2011 WORKSHOP BOOK

Neural Information Processing Systems

NIPS 2011

NEURAL INFORMATION PROCESSING SYSTEMS WORKSHOP

TUTORIALS
December 12, 2011
Granada Congress and
Exhibition Centre, Granada, Spain

CONFERENCE SESSIONS
December 12-15, 2011
Granada Congress and
Exhibition Centre, Granada, Spain

WORKSHOPS
December 16-17, 2011
Melia Sierra Nevada & Melia Sol y
Nieve, Sierra Nevada, Spain

Sponsored by the Neural Information Processing
System Foundation, Inc

The technical program includes six invited talks and 306 accepted papers, selected from a total of 1,400 submissions considered by the program committee. Because the conference stresses interdisciplinary interactions, there are no parallel sessions. Papers presented at the conference will appear in "Advances in Neural Information Processing Systems 23," edited by Rich Zemel, John Shawe-Taylor, Peter Bartlett, Fernando Pereira and Killian Weinberger.

Neural Information
Processing Systems
Foundation

TABLE OF CONTENTS

Contents

Organizing Committee 3

Program Committee 3

NIPS Foundation Offices and Board Members 4

Core Logistics Team 4

Awards 4

Sponsors 5

Program Highlights 6

Maps 7

WS1 Second Workshop on Computational Social Science and the Wisdom of Crowds 10

WS2 Decision Making with Multiple Imperfect Decision Makers 12

WS3 Big Learning: Algorithms, Systems, and Tools for Learning at Scale 16

WS4 Learning Semantics 20

WS5 Integrating Language and Vision 23

WS6 Copulas in Machine Learning 25

WS7 Philosophy and Machine Learning 27

WS8 Relations between machine learning problems: An approach to unify the field 30

WS9 Beyond Mahalanobis: Supervised Large-Scale Learning of Similarity 32

WS10 New Frontiers in Model Order Selection 36

WS11 Bayesian Optimization, Experimental Design and Bandits 38

WS12 Machine Learning for Sustainability 40

WS13 From Statistical Genetics to Predictive Models in Personalized Medicine 41

WS14 Machine Learning meets Computational Photography 43

WS15 Fourth International Workshop on Machine Learning and Music: Learning from Musical Structure 45

WS16 Machine Learning in Computational Biology 48

WS17 Machine Learning and Interpretation in Neuroimaging 50

WS18 Domain Adaptation Workshop: Theory and Application 54

WS19 Challenges in Learning Hierarchical Models: Transfer Learning and Optimization 57

WS20 Cosmology meets Machine Learning 59

WS21 Deep Learning and Unsupervised Feature Learning 61

WS22 Choice Models and Preference Learning 62

WS23 Optimization for Machine Learning 64

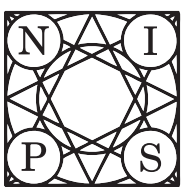
WS24 Computational Trade-offs in Statistical Learning 66

WS25 Bayesian Nonparametric Methods: Hope or Hype? 68

WS26 Sparse Representation and Low-rank Approximation 70

WS27 Discrete Optimization in Machine Learning (DISCML): Uncertainty, Generalization and Feedback 73

Notes 75



™

Neural Information Processing Systems Foundation

ORGANIZING COMMITTEE

General Chairs	**John Shawe-Taylor**, University College London, **Richard Zemel**, University of Toronto
Program Chairs	**Peter Bartlett,** Queensland Univ. of Technology & UC Berkeley**, Fernando Pereira,** Google Research
Spanish Ambassador	**Jesus Cortes**, University of Granada, Spain
Tutorials Chair	**Max Welling,** University of California, Irvine
Workshop Chairs	**Fernando Perez-Cruz,** University Carlos III in Madrid, **Spain; Jeff Bilmes**, University of Washington
Demonstration Chair	**Samy Bengio**, Google Research
Publications Chair & Electronic Proceedings Chair	**Kilian Weinberger**, Washington University in St. Louis
Program Manager	**David Hall**, University of California, Berkeley

PROGRAM COMMITTEE

Cedric Archambeau (Xerox Research Centre Europe)
Andreas Argyriou (Toyota Technological Institute at Chicago)
Peter Auer (Montanuniversität Leoben)
Mikhail Belkin (Ohio State University)
Chiru Bhattarcharyya (Indian Institute of Computer Science)
Charles Cadieu (University of California, Berkeley)
Michael Collins (Columbia University)
Ronan Collobert (IDIAP Research Institute)
Hal Daume III (University of Maryland)
Fei Fei Li (Stanford University)
Rob Fergus (New York University)
Maria Florina Balcan (Georgia Tech)
Kenji Fukumizu (Institute of Statistical Mathematics)
Amir Globerson (The Hebrew University of Jerusalem)
Sally Goldman (Google)
Noah Goodman (Stanford University)
Alexander Gray (Georgia Tech)
Katherine Heller (MIT)
Guy Lebanon (Georgia Tech)
Mate Lengyel (University of Cambridge)
Roger Levy (University of California, San Diego)
Hang Li (Microsoft)
Chih-Jen Lin (National Taiwan University)
Phil Long (Google)
Yi Ma (University of Illinois at Urbana-Champaign)
Remi Munos (INRIA, Lille)
Jan Peters (Max Planck Institute of Intelligent Systems, Tübingen)
Jon Pillow (University of Texas, Austin)
Joelle Pineau (McGill University)
Ali Rahimi (San Francisco, CA)
Sasha Rakhlin (University of Pennsylvania)
Pradeep Ravikumar (University of Texas, Austin)
Ruslan Salakhutdinov (MIT)
Sunita Sarawagi (IIT Bombay)
Thomas Serre (Brown University)
Shai Shalev-Shwartz (The Hebrew University of Jerusalem)
Ingo Steinwart (Universität Stuttgart)
Amar Subramanya (Google)
Masashi Sugiyama (Tokyo Institute of Technology)
Koji Tsuda (National Institute of Advanced Industrial Science and Technology)
Raquel Urtasun (Toyota Technological Institute at Chicago)
Manik Varma (Microsoft)
Nicolas Vayatis (Ecole Normale Supérieure de Cachan)
Jean-Philippe Vert (Mines ParisTech)
Hanna Wallach (University of Massachusetts Amherst)
Frank Wood (Columbia University)
Eric Xing (Carnegie Mellon University)
Yuan Yao (Peking University)
Kai Yu (NEC Labs)
Tong Zhang (Rutgers University)
Jerry Zhu (University of Wisconsin-Madison)

NIPS would like to especially thank Microsoft Research for their donation of Conference Management Toolkit (CMT) software and server space.

NIPS FOUNDATION OFFICERS & BOARD MEMBERS

President	**Terrence Sejnowski**, The Salk Institute
Treasurer	**Marian Stewart Bartlett**, University of California, San Diego
Secretary	**Michael Mozer**, University of Colorado, Boulder
Legal Advisor	**Phil Sotel**, Pasadena, CA
Executive	**John Lafferty**, Carnegie Mellon University; **Chris Williams**, University of Edinburgh; **Dale Schuurmans**, University of Alberta, Canada; **Yoshua Bengio**, University of Montreal, Canada; **Daphne Koller**, Stanford University; **John C. Platt**, Microsoft Research; **Bernhard Schölkopf**, Max Planck Institute for Biological Cybernetics, Tübingen
Advisory Board	**Sue Becker**, McMaster University, Ontario, Canada, **Gary Blasdel**, Harvard Medical School, **Jack Cowan**, University of Chicago, **Thomas G. Dietterich**, Oregon State University, **Stephen Hanson**, Rutgers University, **Michael I. Jordan**, UC Berkeley, **Michael Kearns**, University of Pennsylvania, **Scott Kirkpatrick,** Hebrew University, Jerusalem, **Richard Lippmann**, Massachusetts Institute of Technology, T**odd K. Leen**, Oregon Graduate Institute, **Bartlett Mel**, University of Southern California, **John Moody**, International Computer Science Institute, Berkeley and Portland, **Gerald Tesauro**, IBM Watson Labs, **Dave Touretzky**, Carnegie Mellon University, **Sebastian Thrun,** Stanford University, **Lawrence Saul**, University of California, San Diego, **Sara A. Solla**, Northwestern University Medical School, **Yair Weiss**, Hebrew University of Jerusalem
Emeritus Members	**T. L. Fine**, Cornell University, **Eve Marder**, Brandeis University

CORE LOGISTICS TEAM

The running of NIPS would not be possible without the help of many volunteers, students, researchers and administrators who donate their valuable time and energy to assist the conference in various ways. However, there is a core team at the Salk Institute whose tireless efforts make the conference run smoothly and efficiently every year. This year, NIPS would particularly like to acknowlege the exceptional work of:

Lee Campbell - IT Manager
Chris Hiestand - Webmaster
Mary Ellen Perry - Executive Director
Montse Gamez - Administrator
Ramona Marchand - Administrator

AWARDS

OUTSTANDING STUDENT PAPER AWARDS

Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials
Philipp Krähenbühl * and Vladlen Koltun

Priors over Recurrent Continuous Time Processes
Ardavan Saeedi * and Alexandre Bouchard-Côte

Fast and Accurate k-means For Large Datasets
Michael Shindler * Alex Wong, and Adam Meyerson

*** Winner**

STUDENT PAPER HONORABLE MENTIONS

Learning Sparse Representations of High Dimensional Data on Large Scale Dictionaries
Zhen James Xiang * Hao Xu, and Peter Ramadge

The Manifold Tangent Classifier
Salah Rifai *, Yann Dauphin *, Pascal Vincent, Yoshua Bengio, and Xavier Muller

SPONSORS

NIPS gratefully acknowledges the generosity of those individuals and organizations who have provided financial support for the NIPS 2011 conference. The financial support enabled us to sponsor student travel and participation, the outstanding student paper awards, the demonstration track and the opening buffet.

Microsoft Research

Google™

Artificial Intelligence

D E Shaw & Co

TWO 2σ SIGMA

IBM Research

PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

United Technologies Research Center

TOYOTA

ebaY®
Research Labs

YAHOO! LABS

Telefónica

PROGRAM HIGHLIGHTS

THURSDAY, DECEMBER 15TH

Registration (At Melia Sol y Nieve) 4:30 - 9:30 PM

FRIDAY, DECEMBER 16TH

Registration (At Melia Sol y Nieve) 7:00 - 10:30 AM

Friday Workshops

All workshops run from 7:30 to 10:30AM and from 16:00 to 20:00 PM with breaks from 8:45 to 9:30AM and 5:45 to 6:30PM

WS2 **Decision Making with Multiple Imperfect Decision Makers**
Melia Sol y Nieve: Snow

WS3 **Big Learning: Algorithms, Systems, and Tools for Learning at Scale**
Montebajo: Theater

WS5 **Integrating Language and Vision**
Montebajo: Library

WS6 **Copulas in Machine Learning**
Melia Sierra Nevada: Genil

WS8 **Relations between machine learning problems: An approach to unify the field**
Melia Sierra Nevada: Dilar

WS9 **Beyond Mahalanobis: Supervised Large-Scale Learning of Similarity**
Melia Sierra Nevada: Guejar

WS10 **New Frontiers in Model Order Selection**
Melia Sol y Nieve: Ski

WS11 **Bayesian Optimization, Experimental Design and Bandits**
Melia Sierra Nevada: Hotel Bar

WS13 **From statistical genetics to predictive models in personalized medicine**
Melia Sol y Nieve: Slalom

WS17 **Machine Learning and Interpretation in Neuroimaging**
Melia Sol y Nieve: Aqua

WS20 **Cosmology meets Machine Learning**
Melia Sierra Nevada: Monachil

WS21 **Deep Learning and Unsupervised Feature Learning**
Telecabina: Movie Theater

WS23 **Optimization for Machine Learning**
Melia Sierra Nevada: Dauro

WS24 **Computational Trade-offs in Statistical Learning**
Montebajo: Basketball Court

WS26 **Sparse Representation and Low-rank Approximation**
Montebajo: Room I

SATURDAY, DECEMBER 17TH

Registration (At Melia Sol y Nieve) 7:00 - 11:00 AM

Saturday Workshops

All workshops run from 7:30 to 10:30AM and from 16:00 to 20:00 PM with breaks from 8:45 to 9:30AM and 5:45 to 6:30PM

WS1 **Second Workshop on Computational Social Science and the Wisdom of Crowds**
Telecabina: Movie theater

WS3 **Big Learning: Algorithms, Systems, and Tools for Learning at Scale**
Montebajo: Theater

WS4 **Learning Semantics**
Melia Sol y Nieve: Ski

WS7 **Philosophy and Machine Learning**
Melia Sierra Nevada: Hotel Bar

WS12 **Machine Learning for Sustainability**
Melia Sierra Nevada: Guejar

WS14 **Machine Learning meets Computational Photography**
Melia Sol y Nieve: Snow

WS15 **Fourth International Workshop on Machine Learning and Music: Learning from Musical Structure**
Melia Sierra Nevada: Dilar

WS16 **Machine Learning in Computational Biology**
Melia Sierra Nevada: Genil

WS17 **Machine Learning and Interpretation in Neuroimaging**
Melia Sol y Nieve: Aqua

WS18 **Domain Adaptation Workshop: Theory and Application**
Melia Sierra Nevada: Monachil

WS19 **Challenges in Learning Hierarchical Models: Transfer Learning and Optimization**
Montebajo: Library

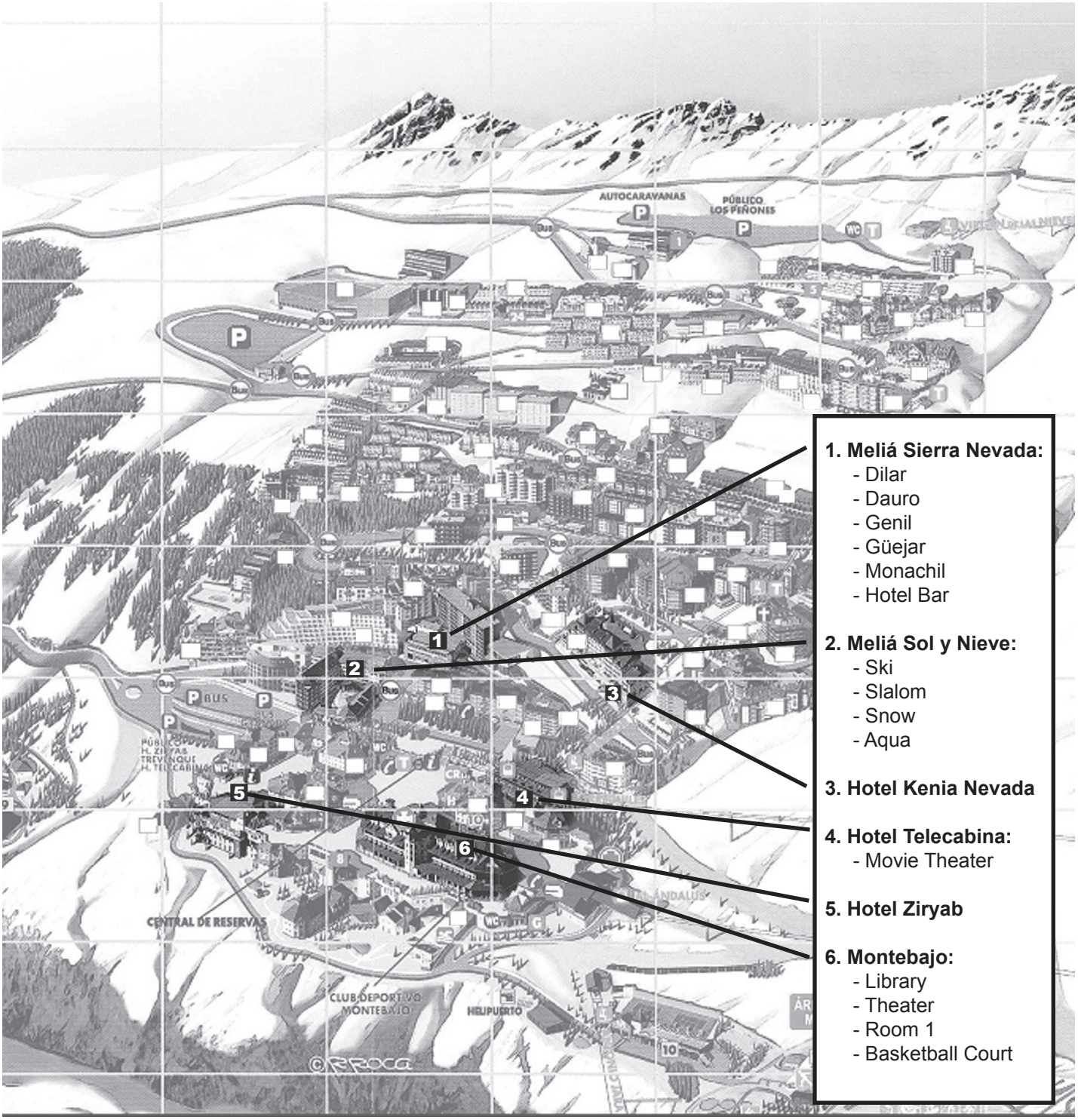
WS22 **Choice Models and Preference Learning**
Montebajo: Room I

WS25 **Bayesian Nonparametric Methods: Hope or Hype?**
Melia Sierra Nevada: Dauro

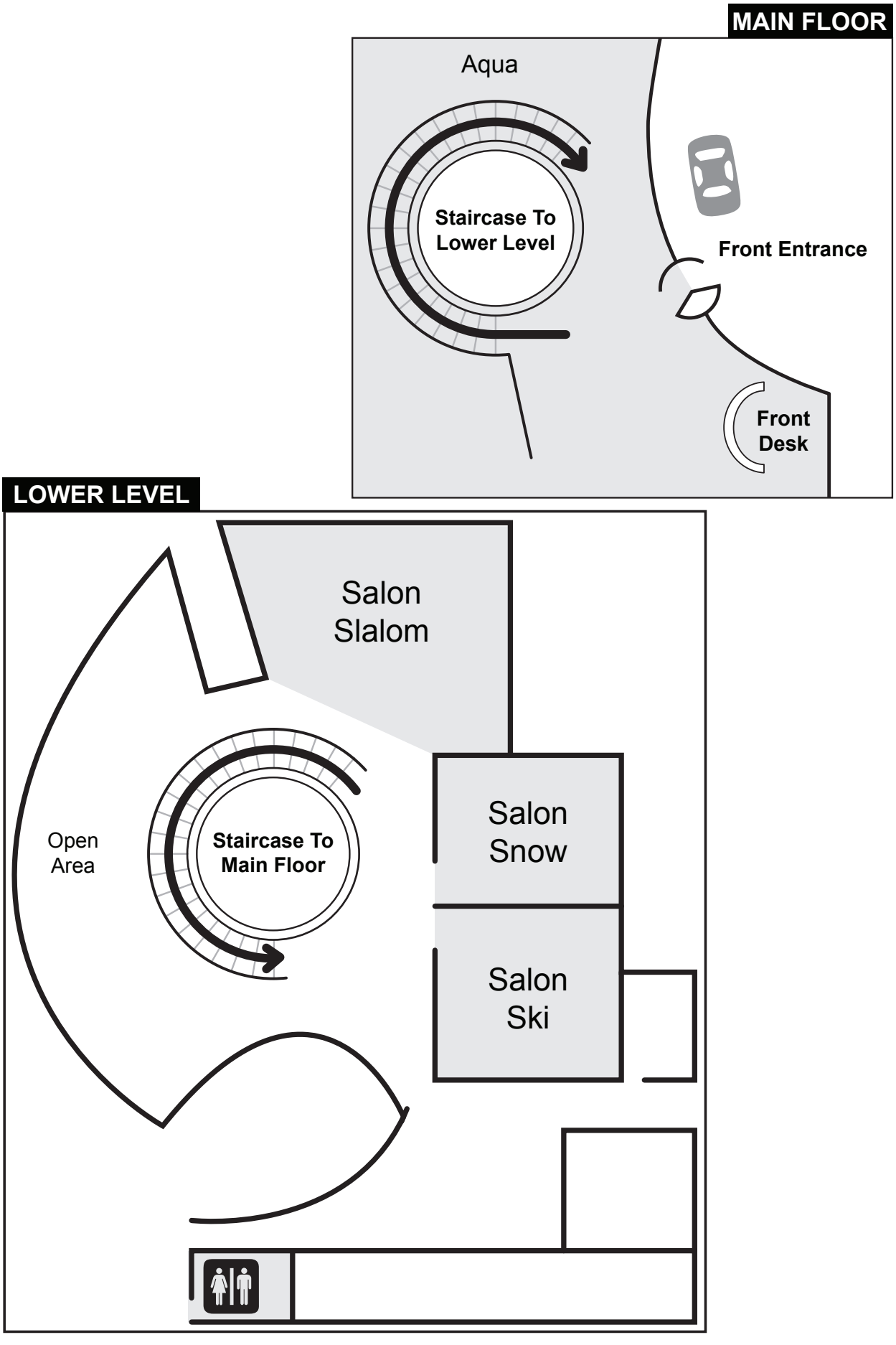
WS27 **Discrete Optimization in Machine Learning (DISCML): Uncertainty, Generalization and Feedback**
Melia Soy y Nieve: Slalom

PLEASE NOTE: Some workshops run on different schedules. Please check timings on the subsequent pages.

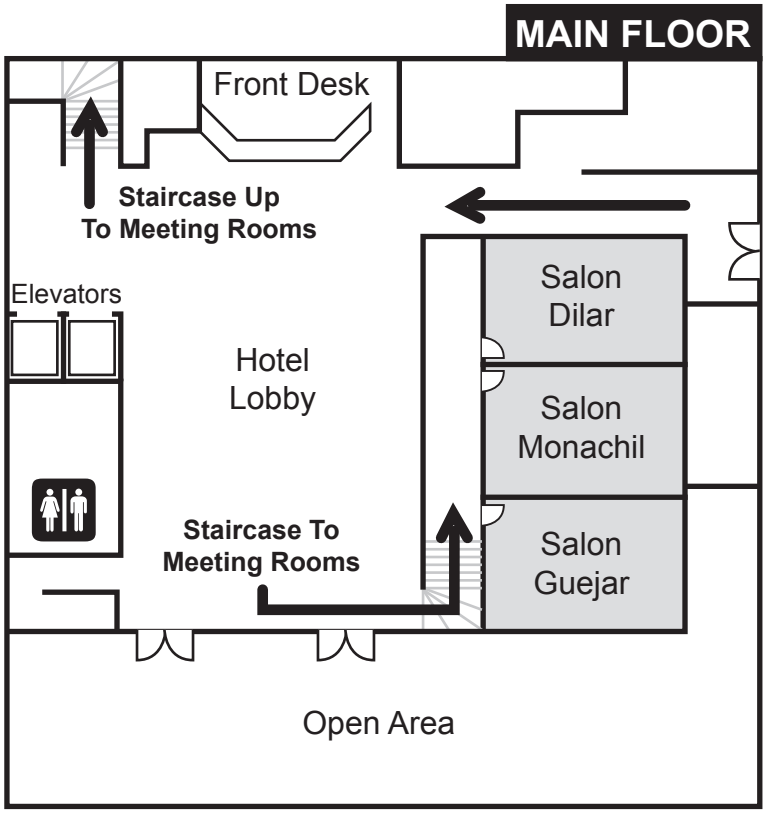
SIERRA NEVADA AREA MAP

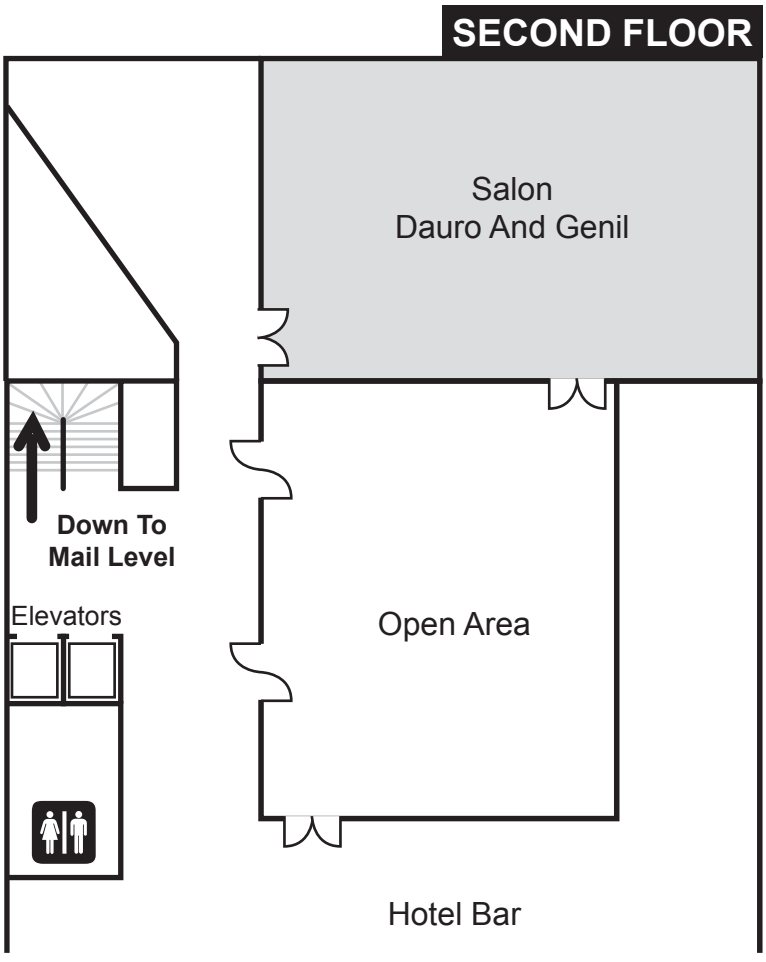


Melia Sol y Nieve - Meeting Rooms



Melia Sierra Nevada - Meeting Rooms





Second Workshop on Computational Social Science and the Wisdom of Crowds

http://www.cs.umass.edu/~wallach/workshops/nips2011css/

LOCATION
Telecabina: Movie theater
Saturday, 7:30 - 10:30 AM & 4:00 - 8:00 PM

Winter Mason m@winteram.com
Stevens Institute of Technology

Jennifer Wortman Vaughan jenn@cs.ucla.edu
UCLA

Hanna Wallach wallach@cs.umass.edu
University of Massachusetts Amherst

Abstract
Computational social science is an emerging academic research area at the intersection of computer science, statistics, and the social sciences, in which quantitative methods and computational tools are used to identify and answer social science questions. The field is driven by new sources of data from the Internet, sensor networks, government databases, crowdsourcing systems, and more, as well as by recent advances in computational modeling, machine learning, statistics, and social network analysis. The related area of social computing deals with the mechanisms through which people interact with computational systems, examining how and why people contribute to crowdsourcing sites, and the Internet more generally. Examples of social computing systems include prediction markets, reputation systems, and collaborative filtering systems, all designed with the intent of capturing the wisdom of crowds. Machine learning plays in important role in both of these research areas, but to make truly ground breaking advances, collaboration is necessary: social scientists and economists are uniquely positioned to identify the most pertinent and vital questions and problems, as well as to provide insight into data generation, while computer scientists are able to contribute significant expertise in developing novel, quantitative methods and tools. The primary goals of this workshop are to provide an opportunity for attendees from diverse fields to meet, interact, share ideas, establish new collaborations, and to inform the wider NIPS community about current research in computational social science and social computing.

SCHEDULE

7.30-7.40	Opening Remarks
7.40-8.25	**Invited Talk**: David Jensen
8.25-8.45	**A Text-based HMM Model of Foreign Affair Sentiment** - Sean Gerrish and David Blei
8.45-9.25	Poster Session 1 and Coffee Break
9.25-10.10	**Invited Talk**: Daniel McFarland
10.10-10.30	**A Wisdom of the Crowd Approach to Forecasting** - Brandon M. Turner and Mark Steyvers
10.30-16.00	Break
16.00-16.45	**Invited Talk**: David Rothschild
16.45-17.05	**Learning Performance of Prediction Markets with Kelly Bettors** - Alina Beygelzimer, John Langford, and David M. Pennock
17.05-17.25	**Approximating the Wisdom of the Crowd** - Seyda Ertekin, Haym Hirsh, Thomas W. Malone, and Cynthia Rudin
17.25-18.05	Poster Session 2 and Coffee Break
18.05-18.50	**Invited Talk**: Aaron Clauset
18.50-19.35	**Invited Talk**: Panagiotis Ipeirotis
19.35-19.45	Closing Remarks and Wrap-up

INVITED SPEAKERS

Invited Talk

David Jensen, University of Massachusetts Amherst

For details on this presentation, please visit the website at the top of this page.

A Text-based HMM Model of Foreign Affair Sentiment

Sean Gerrish
David Blei
Princeton University

We present a time-series model for foreign relations, in which the pairwise sentiment between nations is inferred from news articles. We describe a model of dyadic interaction and illustrate our process of estimating sentiment using Amazon Mechanical Turk labels. Across articles from twenty years of the of the New York Times, we predict with modest error on held out country pairs.

Invited Talk

Daniel McFarland, Stanford University

For details on this presentation, please visit the website at the top of page 8.

A Wisdom of the Crowd Approach to Forecasting

Brandon M. Turner, UC Irvine
Mark Steyvers, UC Irvine

The "wisdom of the crowd" effect refers to the phenomenon that the mean of estimates provided by a group of individuals is more optimal than most of the individual estimates. This effect has mostly been investigated in general knowledge or almanac types of problems that have pre-existing solutions. Can the wisdom of the crowd effect be harnessed to predict the future? We present two probabilistic models for aggregating subjective probabilities for the occurrence of future outcomes. The models allow for individual differences in skill and expertise of participants and correct for systematic distortions in probability judgments. We demonstrate the approach on preliminary results from the Aggregative Contingent Estimation System (ACES), a large-scale project for collecting and combining forecasts of many widely-dispersed individuals.

Invited Talk

David Rothschild, Yahoo! Research
For details on this presentation, please visit the website at the top of page 8.

Learning Performance of Prediction Markets with Kelly Bettors

Alina Beygelzimer, IBM Research
John Langford, Yahoo! Research
David M. Pennock, Yahoo! Research

Kelly betting is an optimal strategy for taking advantage of an information edge in a prediction market, and fractional Kelly is a common variant. We show several consequences that follow by assuming that every participant in a prediction market uses (fractional) Kelly betting. First, the market prediction is a wealth-weighted average of the individual participants' beliefs, where fractional Kelly bettors shift their beliefs toward the market price as if they've seen some fraction of observations. Second, if all fractions are one, the market learns at the optimal rate, the market prediction has low log regret to the best individual participant, and when an underlying true probability exists the market converges to the true objective frequency as if updating a Beta distribution. If fractions are less than one, the market converges to a time-discounted frequency. In the process, we provide a new justification for fractional Kelly betting, a strategy widely used in practice for ad hoc reasons. We propose a method for an agent to learn her own optimal Kelly fraction.

Approximating the Wisdom of the Crowd

Seyda Ertekin, MIT
Haym Hirsh, Rutgers University
Thomas W. Malone, MIT
Cynthia Rudin, MIT

The problem of "approximating the crowd" is that of estimating the crowd's majority opinion by querying only a subset of it. Algorithms that approximate the crowd can intelligently stretch a limited budget for a crowdsourcing task. We present an algorithm, "CrowdSense," that works in an online fashion to dynamically sample subsets of labelers based on an exploration/exploitation criterion. The algorithm produces a weighted combination of the labelers' votes that approximates the crowd's opinion.

Invited Talk

Aaron Clauset, University of Colorado Boulder
For details on this presentation, please visit the website at the top of page 8.

Invited Talk

Panagiotis Ipeirotis, New York University
For details on this presentation, please visit the website at the top of page 8.

Decision Making with Multiple Imperfect Decision Makers

http://www.utia.cz/NIPSHome

LOCATION
Melia Sol y Nieve: Snow
Friday, Dec 16th, 07:30 -- 10:30 AM & 4:00 -- 8:00 PM

Tatiana V. Guy guy@ieee.org
Miroslav Karny school@utia.cas.cz
Institute of Information Theory and Automation, Czech Republic

David Ros Insua david.rios@urjc.es
Royal Academy of Sciences, Spain

Alessandro E.P. Villa Alessandro.Villa@unil.ch
University of Lausanne, Switzerland

David H. Wolpert david.h.wolpert@gmail.com
NASA Ames Research Center, USA

Abstract

Prescriptive Bayesian decision making supported by the efficient theoretically well-founded algorithms is known to be a powerful tool. However, its application within multiple-participants' settings needs an efficient support of an imperfect participant (decision maker, agent), which is characterized by limited cognitive, acting and evaluative resources.

The interacting and multiple-task-solving participants prevail in the natural (societal, biological) systems and become more and more important in the artificial (engineering) systems. Knowledge of conditions and mechanisms in sequencing the participant's individual behavior is a prerequisite to better understanding and rational improving of these systems. The diverse research communities permanently address these topics focusing either on theoretical aspects of the problem or (more often) on practical solution within a particular application. However, different terminology and methodologies used significantly impede further exploitation of any advances occurred. The workshop will bring the experts from different scientific communities to complement and generalize the knowledge gained relying on the multi-disciplinary wisdom. It extends the list of problems of the preceding 2010 NIPS workshop:

How should we formalize rational decision making of a single imperfect decision maker? Does the answer change for interacting imperfect decision makers? How can we create a feasible prescriptive theory for systems of imperfect decision makers?

The workshop especially welcomes contributions addressing the following questions:

What can we learn from natural, engineered, and social systems? How emotions in sequence decision making?

How to present complex prescriptive outcomes to the human? Do common algorithms really support imperfect decision makers? What is the impact of imperfect designers of decision-making systems?

The workshop aims to brainstorm on promising research directions, present relevant case studies and theoretical results, and to encourage collaboration among researchers with complementary ideas and expertise. The workshop will be based on invited talks, contributed talks, posters and demonstrations. Extensive moderated and informal discussions ensure targeted exchange.

SCHEDULE

7.50-8.20	**Emergence of reverse hierarchies in sensing and planning by optimizing predictive information** Naftali Tishby
8.20-8.50	**Modeling Humans as Reinforcement Learners: How to Predict Human Behavior in Multi-Stage Games** Ritchie Lee, David H. Wolpert, Scott Backhaus, Russell Bent, James Bono, Brendan Tracey
8.50-9.20	Coffee Break
9.20-9.50	**Automated Explanations for MDP Policies** Omar Zia Khan, Pascal Poupart, James P. Black
9.50-10.20	**Automated Preference Elicitation** Miroslav Karny, Tatiana V. Guy
10.20-10.40	Poster spotlights
10.40-11.40	Posters & Demonstrations
11.40-4.00	Break
4.00-4.30	**Effect of Emotion on the Imperfectness of Decision Making** Alessandro E. P. Villa, Marina Fiori, Sarah Mesrobian, Alessandra Lintas, Vladyslav Shaposhnyk, Pascal Missonnier
4.30-5.00	**An Adversarial Risk Analysis Model for an Emotional Based Decision Agent** Javier G. Razuri, Pablo G.Esteban, David Rios Insua
5.00-6.00	Posters & Demonstrations (cont.) and Coffee Break
6.00-6.30	**Random Belief Learning** David Leslie
6.30-7.00	**Bayesian Combination of Multiple, Imperfect Classifiers** Edwin Simpson, Stephen Roberts, Ioannis Psorakis, Arfon Smith, Chris Lintott
7.00-8.00	Panel Discussion & Closing Remarks

INVITED SPEAKERS

Emergence of reverse hierarchies in sensing and planning by optimizing predictive information

Naftali Tishby, The Hebrew University of Jerusalem

Efficient planning requires prediction of the future. Valuable predictions are based on information about the future that can only come from observations of past events. Complexity of planning thus depends on the information the past of an environment contains about its future, or on the ”predictive information” of the environment. This quantity, introduced by Bilaek et. al., was shown to be sub-extensive in the past and future time windows, i.e.; to grow sub-linearly with the time intervals, unlike the full complexity (entropy) of events which grow linearly with time in stationary stochastic processes. This striking observation poses interesting bounds on the complexity of future plans, as well as on the required memories of past events. I will discuss some of the implications of this subextesivity of predictive information for decision making and perception in the context of pure information gathering (like gambling) and more general MDP and POMDP settings. Furthermore, I will argue that optimizing future value in stationary stochastic environments must lead to hierarchical structure of both perception and actions and to a possibly new and tractable way of formulating the POMDP problem.

Modeling Humans as Reinforcement Learners: How to Predict Human Behavior in Multi-Stage Games

Ritchie Lee, Carnegie Mellon University
David H. Wolpert, NASA
Scott Backhaus, Los Alamos National Laboratory
Russell Bent, Los Alamos National Laboratory
James Bono, American University, Washington
Brendan Tracey, Stanford University

This paper introduces a novel framework for modeling interacting humans in a multi-stage game environment by combining concepts from game theory and reinforcement learning. The proposed model has the following desirable characteristics: (1) Bounded rational players, (2) strategic (i.e., players account for one another’s reward functions), and (3) is computationally feasible even on moderately large real-world systems. To do this we extend level-K reasoning to policy space to, for the first time, be able to handle multiple time steps. This allows us to decompose the problem into a series of smaller ones where we can apply standard reinforcement learning algorithms. We investigate these ideas in a cyber-battle scenario over a smart power grid and discuss the relationship between the behavior predicted by our model and what one might expect of real human defenders and attackers.

Automated Explanations for MDP Policies

Omar Zia Khan, University of Waterloo
Pascal Poupart, University of Waterloo
James P. Black, University of Waterloo

Explaining policies of Markov Decision Processes (MDPs) is complicated due to their probabilistic and sequential nature. We present a technique to explain policies for factored MDP by populating a set of domain-independent templates. We also present a mechanism to determine a minimal set of templates that, viewed together, completely justify the policy. We demonstrate our technique using the problems of advising undergraduate students in their course selection and evaluate it through a user study.

Automated Preference Elicitation

Miroslav Kárny´, Institute of Information Theory and Automation
Tatiana V. Guy, Institute of Information Theory and Automation

Decision support systems assisting in making decisions became almost inevitable in the modern complex world. Their efficiency depends on the sophisticated interfaces enabling a user take advantage of the support while respecting the increasing on-line information and incomplete, dynamically changing user’s preferences. The best decision making support is useless without the proper preference elicitation. The paper proposes a methodology supporting automatic learning of quantitative description of preferences.

Effect of Emotion on the Imperfectness of Decision Making

Alessandro E. P. Villa, University of Lausanne
Marina Fiori, Lausanne
Sarah Mesrobian, Lausanne
Alessandra Lintas, Lausanne
Vladyslav Shaposhny, Lausanne
Pascal Missonnier, University de Lausanne

Although research has demonstrated the substantial role emotions play in decision-making and behavior traditional economic models emphasize the importance of rational choices rather than their emotional implications. The concept of expected value is the idea that when a rational agent must choose between two options, it will compute the utility of outcome of both actions, estimate their probability of occurrence and finally select the one which offers the highest gain. In the field of neuroeconomics a few studies have analyzed brain and physiological activation during economical monetary exchange revealing that activation of the insula and higher skin conductance were associated to rejecting unfair offers. The aim of the present research is to further extend the understanding of emotions in economic decision-making by investigating the role of basic emotions (happiness, anger, fear, disgust, surprise, and sadness) in the decision-making process. To analyze economic decision-making behavior we used the Ultimatum Game task while recording EEG activity.

In addition, we analyzed the role of individual differences, in particular the personality characteristic of honesty and the tendency to experience positive and negative emotions, as factors potentially affecting the monetary choice.

An Adversarial Risk Analysis Model for an Emotional Based Decision Agent

Javier G. Rázuri, Universidad
Rey Juan Carlos & AISoy Robotics, Madrid
Pablo G. Esteban, Univ. Rey Juan Carlos & AISoy Robotics
David R´ıos Insua, Spanish Royal Academy of Sciences

We introduce a model that describes the decision making process of an autonomous synthetic agent which interacts with another agent and is influenced by affective mechanisms, . This model would reproduce patterns similar to humans and regulate the behavior of agents providing them with some kind of emotional intelligence and improving interaction experience. We sketch the implementation of our model with an edutainment robot.

Random Belief Learning

David Leslie, University of Bristol

When individuals are learning about an environment and other decision-makers in that environment, a statistically sensible thing to do is form posterior distributions over unknown quantities of interest (such as features of the environment and 'opponent' strategy) then select an action by integrating with respect to these posterior distributions. However reasoning with such distributions is very troublesome, even in a machine learning context with extensive computational resources; Savage himself indicated that Bayesian decision theory is only sensibly used in reasonably "small" situations.

Random beliefs is a framework in which individuals instead respond to a single sample from a posterior distribution. There is evidence from the psychological and animal behavior disciplines to suggest that both humans and animals may use such a strategy. In our work we demonstrate such behavior 'solves' the exploration-exploitation dilemma 'better' than other provably convergent strategies. We can also show that such behavior results in convergence to a Nash equilibrium of an unknown game.

Bayesian Combination of Multiple, Imperfect Classifiers

Edwin Simpson, University of Oxford
Stephen Roberts, University of Oxford
Ioannis Psorakis, University of Oxford
Arfon Smith, University of Oxford
Chris Lintott, University of Oxford

In many real-world scenarios we are faced with the need to aggregate information from cohorts of imperfect decision making agents (base classifiers), be they computational or human. Particularly in the case of human agents, we rarely have available to us an indication of how decisions were arrived at or a realistic measure of agent confidence in the various decisions. Fusing multiple sources of information in the presence of uncertainty is optimally achieved using Bayesian inference, which elegantly provides a principled mathematical framework for such knowledge aggregation. In this talk we discuss a Bayesian framework for such imperfect decision combination, where the base classifications we receive are greedy preferences (i.e. labels with no indication of confidence or uncertainty). The classifier combination method we develop aggregates the decisions of multiple agents, improving overall performance. We present a principled framework in which the use of weak decision makers can be mitigated and in which multiple agents, with very different observations, knowledge or training sets, can be combined to provide complementary information. The preliminary application we focus on in this paper is a distributed citizen science project, in which human agents carry out classification tasks, in this case identifying transient objects from images as corresponding to potential supernovae or not. This application, Galaxy Zoo Supernovae, is part of the highly successful Zooniverse family of citizen science projects. In this application the ability of our base classifiers (volunteer citizen scientists) can be very varied and there is no guarantee over any individual's performance, as each user can have radically different levels of domain experience and have different background knowledge. As individual users are not overloaded with decision requests by the system, we often have little performance data for individual users. The methodology we advocate provides a scalable, computationally efficient, Bayesian approach (using variational inference) to learning base classifier performance thus enabling optimal decision combinations. The approach is robust in the presence of uncertainties at all levels and naturally handles missing observations, i.e. in cases where agents do not provide any base classifications. The method far outperforms other established approaches to imperfect decision combination.

Artificial Intelligence Design for Real-time Strategy Games

Firas Safadi, University of Liége
Raphael Fonteneau, University of Liége
Damien Ernst, University of Liége

For now over a decade, real-time strategy (RTS) games have been challenging intelligence, human and artificial (AI) alike, as one of the top genre in terms of overall complexity. RTS is a prime example problem featuring multiple interacting imperfect decision makers. Elaborate dynamics, partial observability, as well as a rapidly diverging action space render rational decision making somehow elusive. Humans deal with the complexity using several abstraction layers, taking decisions on different abstract levels. Current agents, on the other hand, remain largely scripted and exhibit static behavior, leaving them extremely vulnerable to flaw abuse and no match against human players. In this paper, we propose to mimic the abstraction mechanisms used by human players for designing AI for RTS games. A non-learning agent for StarCraft showing promising performance is proposed, and several research directions towards the integration of learning mechanisms are discussed at the end of the paper.

Distributed Decision Making by Categorically-Thinking Agents

Joong Bum Rhim, MIT
Lav R. Varshney, IBM Thomas J. Watson Research Center
Vivek K Goyal, MIT

This paper considers group decision making by imperfect agents that only know quantized prior probabilities for use in Bayesian likelihood ratio tests. Global decisions are made by information fusion of local decisions, but information sharing among agents before local decision making is forbidden. The quantization scheme of the agents is investigated so as to achieve the minimum mean Bayes risk; optimal quantizers are designed by a novel extension to the Lloyd-Max algorithm. Diversity in the individual agents' quantizers leads to optimal performance.

Non-parametric Synthesis of Private Probabilistic Predictions

Phan H. Giang, George Mason University

This paper describes a new non-parametric method to synthesize probabilistic predictions from different experts. In contrast to the popular linear pooling method that combines forecasts with the weights that reflect the average performance of individual experts over the entire forecast space, our method exploits the information that is local to each prediction case. A simulation study shows that our synthesized forecast is calibrated and whose Brier score is close to the theoretically optimal Brier score. Our robust non-parametric algorithm delivers an excellent performance comparable to the best combination method with parametric recalibration - Ranjan-Gneiting's beta-transformed linear pooling.

Decision making and working memory in adolescents with ADHD after cognitive remediation

Michel Bader, Lausanne University
Sarah Leopizzi, Lausanne University
Eleonora Fornari, Biomédicale, Lausanne
Olivier Halfon, Lausanne University
Nouchine Hadjikhani, Harvard Medical School, Lausanne

An increasing number of theoretical frameworks have incorporated an abnormal sensitivity response inhibition as to decision-making and working memory (WM) impairment as key issues in Attention deficit hyperactivity disorder (ADHD). This study reports the effects of 5 weeks cognitive training (RoboMemo, Cogmed) with fMRI paradigm by young adolescents with ADHD at the level of behavioral, neuropsychological and brain activations. After the cognitive remediation, at the level of WM we observed an increase of digit span without significant higher risky choices reflecting decision-making processes. These preliminary results are promising and could provide benefits to the clinical practice. However, models are needed to investigate how executive functions and cognitive training shape high-level cognitive processes as decision-making and WM, contributing to understand the association, or the separability, between distinct cognitive abilities.

Towards Distributed Bayesian Estimation: A Short Note on Selected Aspects

Kamil Dedecius, Institute of Information Theory and Automation
Vladimıra Seckarova, Institute of Information Theory and Automation

The theory of distributed estimation has attained a very considerable focus in the past decade, however, mostly in the classical deterministic realm. We conjecture, that the consistent and versatile Bayesian decision making framework, can significantly contribute to the distributed estimation theory. The paper introduces the problem as a general Bayesian decision making problem and then narrows to the estimation problem. Two mainstream approaches to distributed estimation are presented and the constraints imposed by the environment are studied.

Variational Bayes in Distributed Fully Probabilistic Decision Making

Vaclav Smıdl, Institute of Information Theory and Automation
Ondřej Tichy´, Institute of Information Theory and Automation

We are concerned with design of decentralized control strategy for stochastic systems with global performance measure. It is possible to design optimal centralized control strategy, which often cannot be used in distributed way. The distributed strategy then has to be suboptimal (imperfect) in some sense. In this paper, we propose to optimize the centralized control strategy under the restriction of conditional independence of control inputs of distinct decision makers. Under this optimization, the main theorem for the Fully Probabilistic Design is closely related to that of the well known Variational Bayes estimation method. The resulting algorithm then requires communication between individual decision makers in the form of functions expressing moments of conditional probability densities. This contrasts to the classical Variational Bayes method where the moments are typically numerical. We apply the resulting methodology to distributed control of a linear Gaussian system with quadratic loss function. We show that performance of the proposed solution converges to that obtained using the centralized control.

Ideal and non-ideal predictors in estimation of Bellman function

Jan Zeman, Institute of Information Theory and Automation

The paper considers estimation of Bellman function using revision of the past decisions. The original approach is further extended by employing predictions coming from several imperfect predictors. The resulting algorithm speeds up the convergence of the Bellman function estimation and improves the results quality. The potential of the approach is demonstrated on a futures market data.

Bayesian Combination of Multiple, Imperfect Classifiers

Edwin Simpson, University of Oxford
Stephen Roberts, University of Oxford
Arfon Smith, University of Oxford
Chris Lintott, University of Oxford

Classifier combination methods need to make best use of the outputs of multiple, imperfect classifiers to enable higher accuracy classifications. In many situations, such as when human decisions need to be combined, the base decisions can vary enormously in reliability. A Bayesian approach to such uncertain combination allows us to infer the differences in performance between individuals and to incorporate any available prior knowledge about their abilities when training data is sparse. In this paper we explore Bayesian classifier combination, using the computationally efficient framework of variational Bayesian inference. We apply the approach to real data from a large citizen science project, Galaxy Zoo Supernovae, and show that our method far outperforms other established approaches to imperfect decision combination. We go on to analyze the putative community structure of the decision makers, based on their inferred decision making strategies, and show that natural groupings are formed.

Towards a Supra-Bayesian Approach to Merging of Information

Vladimıra Seckarova, Institute of Information Theory and Automation

Merging of information shared by several decision makers is an important topic in recent years and a lot of solutions has been developed. The main restriction is how to cope with the incompleteness of the information as well as its various forms. The paper introduces merging, which solves the mentioned problems via a Supra-Bayesian approach. The key idea is to unify the forms of the provided information into single one and to treat possible incompleteness. The constructed merging reduces to the Bayesian solution for the particular class of problems.

Demonstration:Interactive Two-Actors game

Ritchie Lee, Carnegie Mellon University

Demonstration:Social Emotional Robot

AISoy Robotics, Madrid

Demonstration: Real-Time Strategy Games

Firas Safadi, University of Liége

Big Learning: Algorithms, Systems, and Tools for Learning at Scale

http://biglearn.org

LOCATION
Montebajo: Theater
Friday & Saturday, December 16th & 17th
07:30 -- 10:30 AM & 4:00 -- 8:00 PM

Joseph Gonzalez jegonzal@cs.cmu.edu
Carlos Guestrin guestrin@cs.cmu.edu
Yucheng Low ylow@cs.cmu.com
Carnegie Mellon University

Sameer Singh sameer@cs.umass.edu
Andrew McCallum mccallum@cs.umass.edu
UMass Amherst

Alice Zheng alicez@microsoft.com
Misha Bilenko mbilenko@microsoft.com
Microsoft Research

Graham Taylor gwtaylor@cs.nyu.edu
New York University

James Bergstra bergstra@rowland.harvard.edu
Harvard

Sugato Basu sugato@google.com
Google Research

Alex Smola alex@smola.org
Yahoo! Research

Michael Franklin franklin@cs.berkeley.edu
Michael Jordan jordan@cs.berkeley.edu
UC Berkeley

Yoshua Bengio yoshua.bengio@umontreal.ca
UMontreal

Abstract

This workshop will address tools, algorithms, systems, hardware, and real-world problem domains related to large-scale machine learning (\Big Learning"). The Big Learning setting has attracted intense interest with active research spanning diverse fields including machine learning, databases, parallel and distributed systems, parallel architectures, and programming languages and abstractions. This workshop will bring together experts across these diverse communities to discuss recent progress, share tools and software, identify pressing new challenges, and to exchange new ideas.

Key topics of interest in this workshop are:

Hardware Accelerated Learning: Practicality and performance of specialized high-performance hardware (e.g. GPUs, FPGAs, ASIC) for machine learning applications.

Applications of Big Learning: Practical application case studies; insights on end-users, typical data work flow patterns, common data characteristics (stream or batch); trade-offs between labeling strategies (e.g., curated or crowd-sourced); challenges of real-world system building.

SCHEDULE

Friday December 16th

7:00-7:30	Poster Setup
7:30-7:40	Introduction
7:40-8:25	Invited talk: **GPU Metaprogramming: A Case Study in Large-Scale Convolutional Neural Networks** Nicolas Pinto
8:25-9:00	Poster Spotlights
9:00-9:25	Poster Session
9:25-9:45	**A Common GPU n-Dimensional Array for Python and C** Arnaud Bergeron
9:45-10:30	Invited talk: **NeuFlow: A Runtime Reconfigurable Data flow Processor for Vision** Yann LeCun and Clement Farabet
4:00-4:45	Invited talk: **Towards Human Behavior Understanding from Pervasive Data: Opportunities and Challenges Ahead** Nuria Oliver
4:45-5:05	**Parallelizing the Training of the Kinect Body Parts Labeling Algorithm** Derek Murray
5:05-5:25	Poster Session
5:25-6:10	Invited talk: **Machine Learning's Role in the Search for Fundamental Particles** Daniel Whiteson
6:10-6:30	**Fast Cross-Validation via Sequential Analysis** Tammo Krueger
6:30-7:00	Poster Session
7:00-7:30	Invited talk: **BigML** Miguel Araujo
7:30-7:50	**Bootstrapping Big Data** Ariel Kleiner

Tools, Software, & Systems: Languages and libraries for large-scale parallel or distributed learning. Preference will be given to approaches and systems that leverage cloud computing (e.g. Hadoop, DryadLINQ, EC2, Azure), scalable storage (e.g. RDBMs, NoSQL, graph databases), and/or specialized hardware (e.g. GPU, Multicore, FPGA, ASIC).

Models & Algorithms: Applicability of different learning techniques in different situations (e.g., simple statistics vs. large structured models); parallel acceleration of computationally intensive learning and inference; evaluation methodology; trade-offs between performance and engineering complexity; principled methods for dealing with large number of features.

INVITED SPEAKERS

GPU Metaprogramming: A Case Study in Large-Scale Convolutional Neural Networks

Nicolas Pinto, Harvard University

Large-scale parallelism is a common feature of many neuro-inspired algorithms. In this short paper, we present a practical tutorial on ways that metaprogramming techniques dynamically generating specialized code at runtime and compiling it just-in-time can be used to greatly accelerate a large data-parallel algorithm. We use filter-bank convolution, a key component of many neural networks for vision, as a case study to illustrate these techniques. We present an overview of several key themes in template metaprogramming, and culminate in a full example of GPU auto-tuning in which an instrumented GPU kernel tem- plate is built and the space of all possible instantiations of this kernel is automatically grid- searched to find the best implementation on various hardware/software platforms. We show that this method can, in concert with traditional hand-tuning techniques, achieve significant speed-ups, particularly when a kernel will be run on a variety of hardware platforms.

A Common GPU n-Dimensional Array for Python and C

Arnaud Bergeron, Universite de Montreal

Currently there are multiple incompatible array/matrix/n-dimensional base object implementations for GPUs. This hinders the sharing of GPU code and causes duplicate development work. This paper proposes and presents a first version of a common GPU n-dimensional array(tensor) named GpuNdArray that works with both CUDA and OpenCL. It will be usable from python, C and possibly other languages.

NeuFlow: A Runtime Reconfigurable Dataflow Processor for Vision

Yann LeCun, New York University
Clément Farabet, New York University

We present a scalable hardware architecture to implement general-purpose systems based on convolutional networks. We will first review some of the latest advances in convolutional networks, their applications and the theory behind them, then present our dataflow processor, a highly-optimized architecture for large vector transforms, which represent 99% of the computations in convolutional networks. It was designed with the goal of providing a high-throughput engine for highly-redundant operations, while consuming little power and remaining completely runtime reprogrammable. We present performance comparisons between software versions of our system executing on CPU and GPU machines, and show that our FPGA implementation can outperform these standard computing platforms.

Towards Human Behavior Understanding from Pervasive Data: Opportunities and Challenges Ahead

Nuria Oliver, Telefonica Research, Barcelona

We live in an increasingly digitized world where our physical and digital interactions leave digital footprints. It is through the analysis of these digital footprints that we can learn and model some of the many facets that characterize people, including their tastes, personalities, social network interactions, and mobility and communication patterns. In my talk, I will present a summary of our research efforts on transforming these massive amounts of user behavioral data into meaningful insights, where machine learning and data mining techniques play a central role. The projects that I will describe cover a broad set of areas, including smart cities and urban computing, psychographics, socioeconomic status prediction and disease propagation. For each of the projects, I will highlight the main results and point at technical challenges still to be solved from a data analysis perspective.

Parallelizing the Training of the Kinect Body Parts Labeling Algorithm

Derek Murray, Microsoft Research

We present the parallelized implementation of decision forest training as used in Kinect to train the body parts classification system. We describe the practical details of dealing with large training sets and deep trees, and describe how to parallelize over multiple dimensions of the problem.

Machine Learning's Role in the Search for Fundamental Particles

Daniel Whiteson, Dept of Physics and Astronomy, UC Irvine

High-energy physicists try to decompose matter into its most fundamental pieces by colliding particles at extreme energies. But to extract clues about the structure of matter from these collisions is not a trivial task, due to the incomplete data we can gather regarding the collisions, the subtlety of the signals we seek and the large rate and dimensionality of the data. These challenges are not unique to high energy physics, and there is the potential for great progress in collaboration between high energy physicists and machine learning experts. I will describe the nature of the physics problem, the challenges we face in analyzing the data, the previous successes and failures of some ML techniques, and the open challenges.

Fast Cross-Validation via Sequential Analysis

Tammo Krueger, Technische Universitaet Berlin

With the increasing size of today's data sets, finding the right parameter configuration via cross-validation can be an extremely time-consuming task. In this paper we propose an improved cross-validation procedure which uses non-parametric testing coupled with sequential analysis to determine the best parameter set on linearly increasing subsets of the data. By eliminating underperforming candidates quickly and keeping promising candidates as long as possible the method speeds up the computation while preserving the capability of the full cross-validation. The experimental evaluation shows that our method reduces the computation time by a factor of up to 70 compared to a full cross-validation with a negligible impact on the accuracy.

Invited talk: BigML

Miguel Araujo

Please visit website at the top of the previous page for details

Bootstrapping Big Data

Ariel Kleiner, UC Berkeley

The bootstrap provides a simple and powerful means of assessing the quality of estimators. However, in settings involving very large datasets, the computation of bootstrap-based quantities can be extremely computationally demanding. As an alternative, we introduce the Bag of Little Bootstraps (BLB), a new procedure which combines features of both the bootstrap and subsampling to obtain a more computationally efficient, though still robust, means of quantifying the quality of estimators. BLB maintains the simplicity of implementation and statistical efficiency of the bootstrap and is furthermore well suited for application to very large datasets using modern distributed computing architectures, as it uses only small subsets of the observed data at any point during its execution. We provide both empirical and theoretical results which demonstrate the efficacy of BLB.

SCHEDULE

Saturday December 17th

7:00-7:30	Poster Setup
7:30-8:15	Invited talk: **Hazy: Making Data-driven Statistical Applications Easier to Build and Maintain** Chris Re
8:15-8:45	Poster Spotlights
8:45-9:05	Poster Session
9:05-9:25	**The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo** Matthew Hoffman
9:25-10:10	Invited talk: **Real Time Data Sketches** Alex Smola
10:10-10:30	**Randomized Smoothing for (Parallel) Stochastic Optimization** John Duchi
4:00-4:20	**Block Splitting for Large-Scale Distributed Learning** Neal Parikh
4:20-5:05	Invited talk: **Spark: In-Memory Cluster Computing for Iterative and Interactive Applications** Matei Zaharia
5:05-5:30	Poster Session
5:30-6:15	Invited talk: **Machine Learning and Hadoop** Jeff Hammerbacher
6:15-6:35	**Large-Scale Matrix Factorization with Distributed Stochastic Gradient Descent** Raimer Gemulla
6:35-7:00	Poster Session
7:00-7:45	Invited talk: **GraphLab 2: The Challenges of Large Scale Computation on Natural Graphs** Carlos Guestrin
7:45-8:00	Closing Remarks

Hazy: Making Data-driven Statistical Applications Easier to Build and Maintain

Chris Re, University of Wisconsin

The main question driving my group's research is: how does one deploy statistical data-analysis tools to enhance data-driven systems? Our goal is to find abstractions that one needs to deploy and maintain such systems. In this talk, I describe my group's attack on this question by building a diverse set of statistical-based data-driven applications: a system whose goal is to read the Web and answer complex questions, a muon detector in collaboration with a neutrino telescope called IceCube, and a social-science applications involving rich content (OCR and speech data). Even in this diverse set, my group has found common abstractions that we are exploiting to build and to maintain systems. Of particular relevance to this workshop is that I have heard of applications in each of these domains referred to as "big data." Nevertheless, in our experience in each of these tasks, after appropriate preprocessing, the relevant data can be stored in a few terabytes -- small enough to fit entirely in RAM or on a handful of disks. As a result, it is unclear to me that scale is the most pressing concern for academics. I argue that dealing with data at TB scale is still challenging, useful, and fun, and I will describe some of our work in this direction. This is joint work with Benjamin Recht, Stephen J. Wright, and the Hazy Team.

The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo

Matthew Hoffman, Columbia University

Hamiltonian Monte Carlo (HMC) is a Markov Chain Monte Carlo (MCMC) algorithm that avoids the random walk behavior and sensitivity to correlations that plague many MCMC methods by taking a series of steps informed by first-order gradient information. These features allow it to converge to high-dimensional target distributions much more quickly than popular methods such as random walk Metropolis or Gibbs sampling. However, HMC's performance is highly sensitive to two user-specified parameters: a step size E and a desired number of steps L. In particular, if L is too small then the algorithm exhibits undesirable random walk behavior, while if L is too large the algorithm wastes computation. We present the No-U-Turn Sampler (NUTS), an extension to HMC that eliminates the need to set a number of steps L. NUTS uses a recursive algorithm to build a set of likely candidate points that spans a wide swath of the target distribution, stopping automatically when it starts to double back and retrace its steps. NUTS is able to achieve similar performance to a well tuned standard HMC method, without requiring user intervention or costly tuning runs. NUTS can thus be used in applications such as BUGS-style automatic inference engines that require efficient "turnkey" sampling algorithms.

Real Time Data Sketches

Alex Smola, Yahoo! Labs

I will describe a set of algorithms for extending streaming and sketching algorithms to real time analytics. These algorithm captures frequency information for streams of arbitrary sequences of symbols. The algorithm uses the Count-Min sketch as its basis and exploits the fact that the sketching operation is linear. It provides real time statistics of arbitrary events, e.g. streams of queries as a function of time. In

particular, we use a factorizing approximation to provide point estimates at arbitrary (time, item) combinations. The service runs in real time, it scales perfectly in terms of throughput and accuracy, using distributed hashing. The latter also provides performance guarantees in the case of machine failure. Queries can be answered in constant time regardless of the amount of data to be processed. The same distribution techniques can also be used for heavy hitter detection in a distributed scalable fashion.

Randomized Smoothing for (Parallel) Stochastic Optimization

John Duchi, UC Berkeley

By combining randomized smoothing techniques with accelerated gradient methods, we obtain convergence rates for stochastic optimization procedures, both in expectation and with high probability, that have optimal dependence on the variance of the gradient estimates. To the best of our knowledge, these are the first variance-based rates for non-smooth optimization. A combination of our techniques with recent work on decentralized optimization yields order-optimal parallel stochastic optimization algorithms. We give applications of our results to statistical machine learning problems, providing experimental results demonstrating the effectiveness of our algorithms.

Block Splitting for Large-Scale Distributed Learning

Neal Parikh, Stanford University

Machine learning and statistics with very large datasets is now a topic of widespread interest, both in academia and industry. Many such tasks can be posed as convex optimization problems, so algorithms for distributed convex optimization serve as a powerful, general-purpose mechanism for training a wide class of models on datasets too large to process on a single machine. In previous work, it has been shown how to solve such problems in such a way that each machine only looks at either a subset of training examples or a subset of features. In this paper, we extend these algorithms by showing how to split problems by both examples and features simultaneously, which is necessary to deal with datasets that are very large in both dimensions. We present some experiments with these algorithms run on Amazon's Elastic Compute Cloud.

Spark: In-Memory Cluster Computing for Iterative and Interactive Applications

Matei Zaharia, AMP Lab, UC Berkeley

MapReduce and its variants have been highly successful in supporting large-scale data-intensive cluster applications. However, these systems are inefficient for applications that share data among multiple computation stages, including many machine learning algorithms, because they are based on an acyclic data flow model. We present Spark, a new cluster computing framework that extends the data flow model with a set of in-memory storage abstractions to efficiently support these applications. Spark outperforms Hadoop by up to 30x in iterative machine learning algorithms while retaining MapReduce's scalability and fault tolerance. In addition, Spark makes programming jobs easy by integrating into the Scala programming language. Finally, Spark's ability to load a dataset into memory and query it repeatedly makes it especially suitable for interactive analysis of big data. We have modified the Scala interpreter to make it possible to use Spark interactively as a highly responsive data analytics tool. At Berkeley, we have used Spark to implement several large-scale machine learning applications, including a Twitter spam classifier and a real-time automobile traffic estimation system based on expectation maximization. We will present lessons learned from these applications and optimizations we added to Spark as a result. Spark is open source and can be downloaded at http://www.spark-project.org.

Machine Learning and Apache Hadoop

Jeff Hammerbacher, Cloudera

We'll review common use cases for machine learning and advanced analytics found in our customer base at Cloudera and ways in which Apache Hadoop supports these use cases. We'll then discuss upcoming developments for Apache Hadoop that will enable new classes of applications to be supported by the system.

Large-Scale Matrix Factorization with Distributed Stochastic Gradient Descent

Rainer Gemulla, MPI

We provide a novel algorithm to approximately factor large matrices with millions of rows, millions of columns, and billions of nonzero elements. Our approach rests on stochastic gradient descent (SGD), an iterative stochastic optimization algorithm. Based on a novel "stratified" variant of SGD, we obtain a new matrix-factorization algorithm, called DSGD, that can be fully distributed and run on web-scale datasets using, e.g., MapReduce. DSGD can handle a wide variety of matrix factorizations; it showed good scalability and convergence properties in our experiments.

GraphLab 2: The Challenges of Large Scale Computation on Natural Graphs

Carlos Guestrin, Carnegie Mellon University

Two years ago we introduced GraphLab to address the critical need for a high-level abstraction for large-scale graph structured computation in machine learning. Since then, we have implemented the abstraction on multicore and cloud systems, evaluated its performance on a wide range of applications, developed new ML algorithms, and fostered a growing community of users. Along the way, we have identified new challenges to the abstraction, our implementation, and the important task of fostering a community around a research project. However, one of the most interesting and important challenges we have encountered is large-scale distributed computation on natural power law graphs. To address the unique challenges posed by natural graphs, we introduce GraphLab 2, a fundamental redesign of the GraphLab abstraction which provides a much richer computational framework. In this talk, we will describe the GraphLab 2 abstraction in the context of recent progress in graph computation frameworks (e.g., Pregel/Giraph). We will review some of the special challenges associated with distributed computation on large natural graphs and demonstrate how GraphLab 2 addresses these challenges. Finally, we will conclude with some preliminary results from GraphLab 2 as well as a live demo. This talk represents joint work with Yucheng Low, Joseph Gonzalez, Aapo Kyrola, Danny Bickson, Alex Smola, and Joseph Hellerstein.

Learning Semantics

http://learningsemanticsnips2011.wordpress.com

LOCATION
Melia Sol y Nieve: Ski
Saturday, 07:30 -- 10:30 AM & 4:00 -- 8:00 PM

Antoine Bordes CNRS UTC	antoine.bordes@hds.utc.fr
Jason Weston Google	jweston@google.com
Ronan Collobert IDIAP	ronan@collobert.com
Leon Bottou Microsoft	leon@bottou.org

Abstract

A key ambition of AI is to render computers able to evolve in and interact with the real world. This can be made possible only if the machine is able to produce a correct interpretation of its available modalities (image, audio, text, etc.), upon which it would then build a reasoning to take appropriate actions. Computational linguists use the term “semantics” to refer to the possible interpretations (concepts) of natural language expressions, and showed some interest in “learning semantics”, that is finding (in an automated way) these interpretations. However, “semantics” are not restricted to natural language modality, and are also pertinent for speech or vision modalities. Hence, knowing visual concepts and common relationships between them would certainly bring a leap forward in scene analysis and in image parsing akin to the improvement that language phrase interpretations would bring to data mining, information extraction or automatic translation, to name a few.

Progress in learning semantics has been slow mainly because this involves sophisticated models which are hard to train, especially since they seem to require large quantities of precisely annotated training data. However, recent advances in learning with weak and limited supervision lead to the emergence of a new body of research in semantics based on multi-task/transfer learning, on learning with semi/ambiguous supervision or even with no supervision at all. The goal of this workshop is to explore these new directions and, in particular, to investigate the following questions:

How should meaning representations be structured to be easily interpretable by a computer and still express rich and complex knowledge?

What is a realistic supervision setting for learning semantics? How can we learn sophisticated representations with limited supervision?

How can we jointly infer semantics from several modalities?

This workshop defines the issue of learning semantics as its main interdisciplinary subject and aims at identifying, establishing and discussing potential, challenges and issues of learning semantics. The workshop is mainly organized around invited speakers to highlight several key current directions, but, it also presents selected contributions and is intended to encourage the exchange of ideas with all the other members of the NIPS community.

SCHEDULE

7.30-7.40	Introduction
7.40-8.20	Invited talk: **Learning Natural Language from its Perceptual Context** Raymond Mooney (UT Austin)
8.20-9.00	Invited talk: **Learning Dependency-Based Compositional Semantics** Percy Liang (Stanford)
9.00-9.10	Coffee
9.10-9.50	Invited talk: **How to Recognize Everything** Derek Hoiem (UIUC)
9.50-10.10	Contributed talk: **Learning What Is Where from Unlabeled Images** A. Chandrashekar and L. Torresani (Dartmouth College)
10.10-10.30	Posters and group discussions
10.30-16.00	Break
16.00-16.40	Invited talk: **From Machine Learning to Machine Reasoning** Leon Bottou (Microsoft)
16.40-17.20	Invited talk: **Towards More Human-like Machine Learning of Word Meanings** Josh Tenenbaum (MIT)
17.20-17.40	Contributed talk: **Learning Semantics of Movement** Timo Honkela et al. (Aalto University)
17.40-17.50	Coffee
17.50-18.30	Invited talk: **Towards Extracting Meaning from Text, and an Autoencoder for Sentences** Chris Burges (Microsoft)
18.30-19.10	Invited talk: **Recursive Deep Learning in Natural Language Processing and Computer Vision** Richard Socher (Stanford)
19.10-20.00	Posters and group discussions

INVITED SPEAKERS

Learning Natural Language from its Perceptual Context

Raymond Mooney, The University of Texas at Austin

Machine learning has become the best approach to building systems that comprehend human language. However, current systems require a great deal of laboriously constructed human-annotated training data. Ideally, a computer would be able to acquire language like a child by being exposed to linguistic input in the context of a relevant but ambiguous perceptual environment. As a step in this direction, we have developed systems that learn to sportscast simulated robot soccer games and to follow navigation instructions in virtual environments by simply observing sample human linguistic behavior. This work builds on our earlier work on supervised learning of semantic parsers that map natural language into a formal meaning representation. In order to apply such methods to learning from observation, we have developed methods that estimate the meaning of sentences from just their ambiguous perceptual context.

Learning Dependency-Based Compositional Semantics

Percy Liang, Stanford University

The semantics of natural language has a highly-structured logical aspect. For example, the meaning of the question ”What is the third tallest mountain in a state not bordering California?” involves superlatives, quantification, and negation. In this talk, we develop a new representation of semantics called Dependency-Based Compositional Semantics (DCS) which can represent these complex phenomena in natural language. At the same time, we show that we can treat the DCS structure as a latent variable and learn it automatically from question/answer pairs. This allows us to build a compositional question-answering system that obtains state-of-the-art accuracies despite using less supervision than previous methods. I will conclude the talk with extensions to handle contextual effects in language.

How to Recognize Everything

Derek Hoiem, UIUC

Our survival depends on recognizing everything around us: how we can act on objects, and how they can act on us. Likewise, intelligent machines must interpret each object within a task context. For example, an automated vehicle needs to correctly respond if suddenly faced with a large boulder, a wandering moose, or a child on a tricycle. Such robust ability requires a broad view of recognition, with many new challenges. Computer vision researchers are accustomed to building algorithms that search through image collections for a target object or category. But how do we make computers that can deal with the world as it comes? How can we build systems that can recognize any animal or vehicle, rather than just a few select basic categories? What can be said about novel objects? How do we approach the problem of learning about many related categories? We have recently begun grappling with these questions, exploring shared representations that facilitate visual learning and prediction for new object categories. In this talk, I will discuss our recent efforts and future challenges to enable broader and more flexible recognition systems.

Learning What Is Where from Unlabeled Images

Ashok Chandrashekar, Dartmouth College
Lorenzo Torresani, Dartmouth College

“What does it mean, to see? The plain man’s answer would be, to know what is where by looking.” This famous quote by David Marr sums up the holy grail of vision: discovering what is present in the world, and where it is, from unlabeled images. To tackle this challenging problem we propose a generative model of object formation and present an efficient algorithm to automatically learn the parameters of the model from a collection of unlabeled images. Our algorithm discovers the objects and their spatial extents by clustering together images containing similar foregrounds. Unlike prior work, our approach does not rely on brittle low-level segmentation methods applied as a first step before the clustering. Instead, it simultaneously solves for the image clusters, the foreground appearance models and the spatial subwindows containing the objects by optimizing a single likelihood function defined over the entire image collection.

From Machine Learning to Machine Reasoning

Léon Bottou, Microsoft

A plausible definition of “reasoning” could be “algebraically manipulating previously acquired knowledge in order to answer a new question”. This definition covers first-order logical inference or probabilistic inference. It also includes much simpler manipulations commonly used to build large learning systems. For instance, we can build an optical character recognition system by first training a character segmenter, an isolated character recognizer, and a language model, using appropriate labeled training sets. Adequately concatenating these modules and fine tuning the resulting system can be viewed as an algebraic operation in a space of models. The resulting model answers a new question, that is, converting the image of a text page into a computer readable text. This observation suggests a conceptual continuity between algebraically rich inference systems, such as logical or probabilistic inference, and simple manipulations, such as the mere concatenation of trainable learning systems. Therefore, instead of trying to bridge the gap between machine learning systems and sophisticated “all-purpose” inference mechanisms, we can instead algebraically enrich the set of manipulations applicable to training systems, and build reasoning capabilities from the ground up.

Towards More Human-like Machine Learning of Word Meanings

Josh Tenenbaum, MIT

How can we build machines that learn the meanings of words more like the way that human children do? I will talk about several challenges and how we are beginning to address them using sophisticated probabilistic models. Children can learn words from minimal data, often just one or a few positive examples (one-shot learning). Children learn to learn: they acquire powerful inductive biases for new word meanings in the course of learning their first words. Children can learn words for abstract concepts or types of concepts that have no little or no direct perceptual correlate. Children’s language can be highly context-sensitive, with parameters of word meaning that must be computed anew for each context rather than simply stored. Children learn function words: words whose meanings are expressed purely in how they

compose with the meanings of other words. Children learn whole systems of words together, in mutually constraining ways, such as color terms, number words, or spatial prepositions. Children learn word meanings that not only describe the world but can be used for reasoning, including causal and counterfactual reasoning. Bayesian learning defined over appropriately structured representations -- hierarchical probabilistic models, generative process models, and compositional probabilistic languages -- provides a basis for beginning to address these challenges.

Learning Semantics of Movement

Timo Honkela, Aalto University
Oskar Kohonen, Aalto University
Jorma Laaksonen, Aalto University
Krista Lagus, Aalto University
Klaus Fórger, Aalto University
Mats Sjóberg, Aalto University
Tapio Takala, Aalto University
Harri Valpola, Aalto University
Paul Wagner, Aalto University

In this presentation, we consider how to computationally model the interrelated processes of understanding natural language and perceiving and producing movement in multimodal real world contexts. Movement is the specific focus of this presentation for several reasons. For instance, it is a fundamental part of human activities that ground our understanding of the world. We are developing methods and technologies to automatically associate human movements detected by motion capture and in video sequences with their linguistic descriptions. When the association between human movement and their linguistic descriptions has been learned using pattern recognition and statistical machine learning methods, the system is also used to produce animations based on written instructions and for labeling motion capture and video sequences. We consider three different aspects: using video and motion tracking data, applying multi-task learning methods, and framing the problem within cognitive linguistics research.

Towards Extracting Meaning from Text, and an Autoencoder for Sentences

Chris J.C. Burges, Microsoft

I will begin with a brief overview of some of the projects underway at Microsoft Research Redmond that are aimed at extracting meaning from text. I will then describe a data set that we are making available and which we hope will be useful to researchers who are interested in semantic modeling. The data is composed of sentences, each of which has several variations: in each variation, one of the words has been replaced by one of several alternatives, in such a way that the low order statistics are preserved, but where a human can determine that the meaning of the new sentence is compromised (the “sentence completion” task). Finally I will describe an autoencoder for sentence data. The autoencoder learns vector representations of the words in the lexicon and maps sentences to fixed length vectors. I’ll describe several possible applications of this work, show some early results on learning Wikipedia sentences, and end with some speculative ideas on how such a system might be leveraged in the quest to model meaning.

Recursive Deep Learning in Natural Language Processing and Computer Vision

Richard Socher, Stanford University

Hierarchical and recursive structure is commonly found in different modalities, including natural language sentences and scene images. I will present some of our recent work on three recursive neural network architectures that learn meaning representations for such hierarchical structure. These models obtain state-of-the-art performance on several language and vision tasks. The meaning of phrases and sentences is determined by the meanings of its words and the rules of compositionality. We introduce a recursive neural network (RNN) for syntactic parsing which can learn vector representations that capture both syntactic and semantic information of phrases and sentences. Our RNN can also be used to find hierarchical structure in complex scene images. It obtains state-of-the-art performance for semantic scene segmentation on the Stanford Background and the MSRC datasets and outperforms Gist descriptors for scene classification by 4%. The ability to identify sentiments about personal experiences, products, movies etc. is crucial to understand user generated content in social networks, blogs or product reviews. The second architecture I will talk about is based on recursive autoencoders (RAE). RAEs learn vector representations for phrases sufficiently well as to outperform other traditional supervised sentiment classification methods on several standard datasets. We also show that without supervision RAEs can learn features which outperform previous approaches for paraphrase detection on the Microsoft Research Paraphrase corpus. This talk presents joint work with Andrew Ng and Chris Manning.

Integrating Language and Vision

https://sites.google.com/site/nips2011languagevisionworkshop/

LOCATION
Montebajo: Library
Friday, December 16th, 07:30 -- 10:30 AM & 4:00 -- 8:00 PM

Trevor Darrell trevor@eecs.berkeley.edu
University of California at Berkeley

Raymond Mooney mooney@cs.utexas.edu
University of Texas at Austin

Kate Saenko saenko@eecs.berkeley.edu

Abstract

A growing number of researchers in computer vision have started to explore how language accompanying images and video can be used to aid interpretation and retrieval, as well as train object and activity recognizers. Simultaneously, an increasing number of computational linguists have begun to investigate how visual information can be used to aid language learning and interpretation, and to ground the meaning of words and sentences in perception. However, there has been very little direct interaction between researchers in these two distinct disciplines. Consequently, researchers in each area have a quite limited understanding of the methods in the other area, and do not optimally exploit the latest ideas and techniques from both disciplines when developing systems that integrate language and vision. Therefore, we believe the time is particularly opportune for a workshop that brings together researchers in both computer vision and natural-language processing (NLP) to discuss issues and ideas in developing systems that combine language and vision.

Traditional machine learning for both computer vision and NLP requires manually annotating images, video, text, or speech with detailed labels, parse-trees, segmentations, etc. Methods that integrate language and vision hold the promise of greatly reducing such manual supervision by using naturally co-occurring text and images/video to mutually supervise each other.

There are also a wide range of important real-world applications that require integrating vision and language, including but not limited to: image and video retrieval, human-robot interaction, medical image processing, human-computer interaction in virtual worlds, and computer graphics generation.

More than any other major conference, NIPS attracts a fair number of researchers in both computer vision and computational linguistics. Therefore, we believe it is the best venue for holding a workshop that brings these two communities together for the very first time to interact, collaborate, and discuss issues and future directions in integrating language and vision.

SCHEDULE

7.30-7:35	**Introductory Remarks** Trevor Darrell, Raymond Mooney, Kate Saenko
7:35-8:00	**Automatic Caption Generation for News Images** Mirella Lapata
8:00-8:25	**Integrating Visible Communicative Behavior with Semantic Interpretation of Language** Stanley Peters
8:25-8:50	**Describing and Searching for Images with Sentences** Julia Hockenmaier
8:50-9:00	Coffee break
9:00-9:25	**Grounding Language in Robot Control Systems** Dieter Fox
9:25-9:50	**Grounding Natural-Language in Computer Vision and Robotics** Jeffery Siskind
9:50-10:30	**Panel on Challenge Problems and Datasets** Tamara Berg, Julia Hockenmaier, Raymond Mooney, Louis-Philippe Morency
16:00-16:25	**Modeling Co-occurring Text and Images** Kate Saenko, Yangqing Jia
16:25-16:50	**Learning from Images and Descriptive Text** Tamara Berg
16:50-17:15	**Harvesting Opinions from the Web: The Challenge of Linguistic, Auditory and Visual Integration** Louis-Philippe Morency
17:15-17:17	Spotlight: **Multimodal Distributional Semantics** Elia Bruni
17:17-17:19	Spotlight: **Joint Inference of Soft Biometric Features** Niyati Chhaya
17:19-17:21	Spotlight: **The Visual Treebank** Desmond Elliott

Continued on Next Page

17:21-17:23 Spotlight: **Text-to-Image Generation based on Crossmodal Association with Hierarchical Hypergraphs**
Jung-Woo Ha

17:23-17:25 Spotlight: **Learning Cross-modality Similarity for Multinomial Data**
Yangqing Jia

17:25-17:27 Spotlight: **From Situated Descriptions of Spatial Scenes to Situated Dialogue**
Staffan Larsson

17:27-17:29 Spotlight: **Learning to Describe Activities in Youtube Clips**
Tanvi Motwani

17:29-17:31 Spotlight: **Semantic Annotation of Soap Videos by Relying on Image and Text Features**
Phi The Pham

17:31-17:33 Spotlight: **Bayesian Mixture Modeling of Joint Vision-Language Concepts from Videos**
Byoung-Tak Zhang

17:35-18:35 Poster Session

18:35-19:00 **Pragmatics, Compositionality, and Spatial Descriptions**
Percy Liang

19:00-19:25 **What Kinds of Representations are Needed to Talk about Visual Scenes?**
Joshua Tenenbaum

19:25-20:00 Closing Discussion

Copulas in Machine Learning

http://pluto.huji.ac.il/~galelidan/CopulaWorkshop

LOCATION
Melia Sierra Nevada: Genil
Friday, December 16th, 07:30 -- 10:30 AM & 4:00 -- 7:30 PM

Gal Elidan galel@huji.ac.il
The Hebrew University of Jerusalem

Zoubin Ghahramani zoubin@eng.cam.ac.uk
Cambridge University and Carnegie Mellon University

John Lafferty lafferty@cs.cmu.edu
University of Chicago and Carnegie Mellon University

Abstract

From high-throughput biology and astronomy to voice analysis and medical diagnosis, a wide variety of complex domains are inherently continuous and high dimensional. The statistical framework of copulas offers a flexible tool for modeling highly non-linear multivariate distributions for continuous data. Copulas are a theoretically and practically important tool from statistics that explicitly allow one to separate the dependency structure between random variables from their marginal distributions. Although bivariate copulas are a widely used tool in finance, and have even been famously accused of “bringing the world financial system to its knees” (Wired Magazine, Feb. 23, 2009), the use of copulas for high dimensional data is in its infancy.

While studied in statistics for many years, copulas have only recently been noticed by a number of machine learning researchers, with this “new” tool appearing in the recent leading machine learning conferences (ICML, UAI and NIPS). The goal of this workshop is to promote the further understanding and development of copulas for the kinds of complex modeling tasks that are the focus of machine learning. Specifically, the goals of the workshop are to:

* Draw the attention of machine learning researchers to the important framework of copulas

* Provide a theoretical and practical introduction to copulas

* Identify promising research problems in machine learning that could exploit copulas

* Bring together researchers from the statistics and machine learning communities working in this area.

The target audience includes leading researchers from academia and industry, with the aim of facilitating cross fertilization between different perspectives.

SCHEDULE

7:30-7:40	Opening remarks
7:40-9:10	Keynote tutorial: **Everything You Always Wanted to Know About Copula Modeling but Were Afraid to Ask** Christian Genest
09:10-09:30	Break
09:30-10:00	**Contributed tutorial: Introduction to Vine Models** Nicole Kraemer and Ulf Schepsmeier
10:00-10:30	**High-dimensional Copula Constructions in Machine Learning: An Overview** Gal Elidan
10:30-16:00	Break
16:00-16:45	Invited Talk: **Exploiting Copula Parameterizations in Graphical Model Construction and Learning** Ricardo Silva
16:45-17:05	**Copula Mixture Model for Dependency-seeking Clustering** Melanie Rey,Volker Roth
17:05-17:20	Poster Spotlights
17:20-18:00	Poster Session
18:00-18:20	Break
18:20-18:40	**Expectation Propagation for the Estimation of Conditional Bivariate Copulas** Jose Miguel Hernandes-Lobato,David Lopez Paz and Zoubin Ghahramani
18:40-19:00	**Robust Nonparametric Copula Based Dependence Estimators** Barnabas Poczos, Sergey Krishner, David Pal, Csaba Szepesvari and Jeff Schneider
19:00-19:30	Group Discussion

INVITED SPEAKERS

Keynote tutorial: Everything You Always Wanted to Know About Copula Modeling but Were Afraid to Ask
Christian Genest, McGill University

Visit the website on the previous page for details

Contributed tutorial: Introduction to Vine Models
Nicole Kraemer, TU Munchen
Ulf Schepsmeier, TU Munchen

In this talk, we introduce the main concepts of vine pair-copula constructions. This framework uses bivariate copulas as building blocks to obtain higher-dimensional distributions. As these bivariate copulas can be selected from a wide range of families, the vine approach leads to a more flexible model compared to traditional approaches. For the estimation of vine pair-copulas, we introduce a mathematically elegant framework that joins (a) graph theory, to determine the dependency structure of the data, and (b) maximum-likelihood estimation, to fit bivariate copulas.

High-dimensional Copula Constructions in Machine Learning: An Overview
Gal Elidan, The Hebrew University of Jerusalem

With the "discovery" of copulas by machine learning researchers, several works have emerged that focus on the high-dimensional scenario. This talk will provide a brief overview of these works and cover tree-averaged distributions (Kirshner), the nonparanormal (Liu, Lafferty and Wasserman), copula processes (Wilson and Ghahramani), kernel-based copula processes (Jaimungal and Ng), and copula networks (Elidan). Special emphasis will be given to the high level similarities and differences between these works.

Invited Talk: Exploiting Copula Parameterizations in Graphical Model Construction and Learning
Ricardo Silva, University College London
Robert Gramacy, University of Cambridge
Charles Blundell, University College London
Yee Whye Teh, University College London

Graphical models and copulas are two sets of tools for multivariate analysis. Both are in some sense pathways to the construction of multivariate distributions using modular representations. The former focuses on languages to express conditional independence constraints, factorizations and efficient inference algorithms. The latter allows for the encoding of some marginal features of the joint distribution (univariate marginals, in particular) directly, without resorting to an inference algorithm. In this talk we exploit copula parameterizations in two graphical modeling tasks: parameterizing decomposable models and building proposal distributions for inference with Markov chain Monte Carlo; parameterizing directed mixed graph models and providing simple estimation algorithms based on composite likelihood methods.

Copula Mixture Model for Dependency-seeking Clustering
Melanie Rey, University of Basel
Volker Roth, University of Basel

We introduce a Dirichlet prior mixture of meta-Gaussian distributions to perform dependency-seeking clustering when co-occurring samples from different data sources are available. The model extends Bayesian mixtures of Canonical Correlation Analysis clustering methods to multivariate data distributed with arbitrary continuous margins. Using meta-Gaussian distributions gives the freedom to specify each margin separately and thereby also enables clustering in the joint space when the data are differently distributed in the different views. The Bayesian mixture formulation retains the advantages of using a Dirichlet prior. We do not need to specify the number of clusters and the model is less prone to overfitting than non-Bayesian alternatives. Inference is carried out using a Markov chain sampling method for Dirichlet process mixture models with non-conjugate prior adapted to the copula mixture model. Results on different simulated data sets show significant improvement compared to a Dirichlet prior Gaussian mixture and a mixture of CCA model.

Expectation Propagation for the Estimation of Conditional Bivariate Copulas
Jose Miguel Hernandes-Lobato, University of Cambridge
David Lopez Paz, MPI for Intelligent Systems
Zoubin Ghahramani, Cambridge University and CMU

We present a semi-parametric method for the estimation of the copula of two random variables X and Y when conditioning to an additional covariate Z. The conditional bivariate copula is described using a parametric model fully specified in terms of Kendall's tau. The dependence of the conditional copula on Z is captured by expressing tau as a function of Z. In particular, tau is obtained by filtering a non-linear latent function, which is evaluated on Z, through a sigmoid-like function. A Gaussian process prior is assumed for the latent function and approximate Bayesian inference is performed using expectation propagation. A series of experiments with simulated and real-world data illustrate the advantages of the proposed approach.

Robust Nonparametric Copula Based Dependence Estimators
Barnabas Poczos, Carnegie Mellon University
Sergey Krishner, Purdue University
David Pal, Google, Inc.
Csaba Szepesvari, University of Alberta
Jeff Schneider, Carnegie Mellon University

A fundamental problem in statistics is the estimation of dependence between random variables. While information theory provides standard measures of dependence (e.g. Shannon-, Renyi-, Tsallis-mutual information), it is still unknown how to estimate these quantities from i.i.d. samples in the most efficient way. In this presentation we review some of our recent results on copula based nonparametric dependence estimators and demonstrate their robustness to outliers both theoretically in terms of finite-sample breakdown points and by numerical experiments in independent subspace analysis and image registration.

Philosophy and Machine Learning

http://www.dsi.unive.it/PhiMaLe2011/

LOCATION
Melia Sierra Nevada: Hotel Bar
Saturday, 07:30 -- 10:30 AM & 4:00 -- 8:00 PM

Marcello Pelillo pelillo@dsi.unive.it
University of Venice

Joachim Buhmann jbuhmann@inf.ethz.ch
ETH Zurich

Tiberio Caetano t iberio.caetano@nicta.com.au
NICTA

Bernhard Schölkopf bs@tuebingen.mpg.de
MPI for Biological Cybernetics

Larry Wasserman larry@stat.cmu.edu
Carnegie Mellon University

Abstract

The fields of machine learning and pattern recognition can arguably be considered as a modern-day incarnation of an endeavor which has challenged mankind since antiquity. In fact, fundamental questions pertaining to categorization, abstraction, generalization, induction, etc., have been on the agenda of mainstream philosophy, under different names and guises, since its inception. Nowadays, with the advent of modern digital computers and the availablity of enormous amount of raw data, these questions have taken a computational flavor.

As it often happens with scientific research, in the early days of machine learning there used to be a genuine interest around philosophical and conceptual issues, but over time the interest shifted almost entirely to technical and algorithmic aspects and became driven mainly by practical applications. In recent years, however, there has been a renewed interest around the foundational and/or philosophical problems of machine learning and pattern recognition, from both the computer scientist's and the philosopher's camps. This suggests that the time is ripe to initiating a long-term dialogue between the philosophy and the machine learning communities with a view to foster cross-fertilization of ideas.

In particular, we do feel the present moment is appropriate for reflection, reassessment and eventually some synthesis, with the aim of providing the machine learning field a self-portrait of where it currently stands and where it is going as a whole, and hopefully suggesting new directions. The aim of this workshop is precisely to consolidate research efforts in this area, and to provide an informal discussion forum for researchers and practitioners interested in this important yet diverse subject.

The workshop is planned to be a one-day meeting. The program will feature invited as well as contributed presentations. We feel that the more informal the better and we would like to solicit open and lively discussions and exchange of ideas from researchers with different backgrounds and perspectives. Plenty of time will be allocated to questions, discussions, and breaks.

SCHEDULE

7.30-7.40	Introduction
7.40-8.40	Invited talk: **Between the Philosophy of Science and Machine Learning** David Corfield
8.40-9.10	**Information, Learning and Falsication** David Balduzzi
9.10-9.30	Coffee Break
9.30-10.00	**On the Computability and Complexity of Bayesian Reasoning** Daniel Roy
10.00-10.15	**A Neural-Symbolic Approach to the Contemporary Theory of Metaphor** Artur dAvila Garcez, Guido Boella, Alan Perotti
10.15-10.30	**Bayesian Causal Induction** Pedro Ortega
10.30-16.00	Ski Break
16.00-17.00	**Invited talk: Foundations of Induction** Marcus Hutter
17.00-17.30	**Beyond calculation: probabilistic computing machines and universal stochastic inference** Vikash Mansinghka
17.30-17.45	Coffee Break
17.45-18.15	**Universal Learning vs. No Free Lunch Results** Shai Ben-David, Nathan Srebro, Ruth Urner
18.15-18.45	**Are Mental Properties Supervenient on Brain Properties?** Joshua T. Vogelstein, R. Jacob Vogelstein, Carey E. Priebe
18.45-19.00	**Toward a New Representation for Causation in Dynamic Systems** Denver Dash, Mark Voortman
19.00-20.00	**Panel and Group Discussion** Joachim Buhmann, Tiberio Caetano, Nello Cristianini, Marcello Pelillo, Bob Williamson

INVITED SPEAKERS

Invited Talk: Between the Philosophy of Science and Machine Learning

David Corfield, University of Kent

In this talk I will recount some of my experiences as a philosopher of science, working alongside a machine learning group for over two years. The philosophy of science in the twentieth century has adopted two radically different approaches, a formal one, as exemplified by inductive logic and Carnap's explications, and a historical one, as exemplified by Lakatos and Kuhn. Although the former approach may seem the natural place for machine learning researchers to look for inspiration, I will suggest that a scientific approach to learning would do well to discover what those who have aimed for a realistic account of natural science have to say about our most powerful form of learning.

Information, Learning and Falsification

David Balduzzi, MPI for Intelligent Systems

Broadly speaking, there are two approaches to quantifying information. The first, Shannon information, takes events as belonging to ensembles and quantifies the information resulting from observing the given event in terms of the number of alternate events that have been ruled out. The second, algorithmic information or Kolmogorov complexity, takes events as strings and, given a universal Turing machine, quantifies the information content of a string as the length of the shortest program producing it. Shannon information provides the mathematical foundation for communication and coding theory. Algorithmic information has been applied by Solomonoff and Hutter to prove remarkable results on universal induction. However, both approaches have shortcomings. Algorithmic information is not computable, severely limiting its practical usefulness. Shannon information refers to ensembles rather than actual events: it makes no sense to compute the Shannon information of a single string or rather, there are many answers to this question depending on how a related ensemble is constructed. Although there are asymptotic results linking algorithmic and Shannon information, it is unsatisfying that there is such a large gap a difference in kind between the two measures. This note describes a new method of quantifying information, effective information, that links algorithmic information to Shannon information, and also links both to capacities arising in statistical learning theory. After introducing the measure, we show that it provides a non-universal analog of algorithmic information. We then apply it to derive basic capacities in statistical learning theory: empirical VC-entropy and empirical Rademacher complexity. A nice byproduct of our approach is an interpretation of the explanatory power of a learning algorithm in terms of the number of hypotheses it falsifies (counted in two different ways for the two different capacities). We also discuss how effective information relates to information gain, Shannon and mutual information. We conclude by discussing some broader implications of our results.

On the Computability and Complexity of Bayesian Reasoning

Daniel Roy, University of Cambridge

If we consider the claim made by some cognitive scientists that the mind performs Bayesian reasoning, and if we simultaneously accept the Physical Church-Turing thesis and thus believe that the computational power of the mind is no more than that of a Turing machine, then what limitations are there to the reasoning abilities of the mind? I purpose to give an overview of joint work with Nathanael Ackerman (Harvard, Mathematics) and Cameron Freer (MIT, CSAIL) that bears on the computability and complexity of Bayesian reasoning. In particular, we prove that conditional probability is in general not computable in the presence of continuous random variables. However, in light of additional structure in the prior distribution, such as the presence of certain types of noise, or of exchangeability, conditioning is possible. At the workshop on Logic and Computational Complexity, we presented results on the computational complexity of conditioning that complement older work. E.g., under cryptographic assumptions, the computational complexity of producing samples and computing probabilities was separated by Ben-David, Chor, Goldreich and Luby. In as yet unpublished work, we also make use of cryptographic assumptions to show that different representations of exchangeable sequences may have vastly different complexity. However, when faced with an adversary that is computational bounded, these different representations have the same complexity, highlighting the fact that knowledge representation and approximation play a fundamental role in the possibility and plausibility of Bayesian reasoning.

A Neural-Symbolic Approach to the Contemporary Theory of Metaphor

Artur dAvila Garcez, City University London
Guido Boella, Universitá di Torino, Italy
Alan Perotti, Universitá di Torino, Italy

Lakoff defined the metaphor as a mapping between knowledge domains. The cognitive role of metaphor is to reuse the knowledge about a source domain we have expertiese on in order to reason about another target domain. We propose in this paper a model of metaphor to implement the idea of reusing existing knowledge about one domain in another one. We propose a model of metaphor using a neural approach, first of all to mimic the neural model of metaphor of our brain and secondly to exploit the learning capability of neural networks to handle the limited knowledge about the target domain. Moreover, since we assume that in some cases partial knowledge about both the target and source domain can be already available in the form of symbolic declarative knowledge, we adopt a neural symbolic approach to compile this knowledge into a neural network. This approach allows also for the inverse process: to extract symbolic declarative knowledge after the learning phase. To build a neural model, first we formalize the definition of metaphor by Lakoff as a monomorphism. To model the invertibility of the mapping in a monomorphism we cannot simply use feedforward models but we resort to RBMs because they display symmetric connections between layers. Our approach can be used for software reuse and flexible commitment in multiagent systems.

Bayesian Causal Induction

Pedro Ortega, MPI for Biological Cybernetics

Discovering causal relationships is a hard task, often hindered by the need for intervention, and often requiring large amounts of data to resolve statistical uncertainty. However, humans quickly arrive at useful causal relationships. One possible reason is that humans use strong prior knowledge; and rather than encoding hard causal relationships, they encode beliefs over causal structures, allowing for sound generalization from the observations they obtain from directly acting in the world. In this work we propose a Bayesian approach to causal induction which allows modeling beliefs over multiple causal hypotheses and predicting the behavior of the world under causal interventions. We then illustrate how this method extracts causal information from data containing interventions and observations.

Invited talk: Foundations of Induction

Marcus Hutter, Australian National University

Humans and many other intelligent systems (have to) learn from experience, build models of the environment from the acquired knowledge, and use these models for prediction. In philosophy this is called inductive inference, in statistics it is called estimation and prediction, and in computer science it is addressed by machine learning. I will first review unsuccessful attempts and unsuitable approaches towards a general theory of induction, including Popper's falsificationism and denial of (the necessity of) confirmation, frequentist statistics and much of statistical learning theory, subjective Bayesianism, Carnap's confirmation theory, the data paradigm, eliminative induction, and deductive approaches. I will also debunk some other misguided views, such as the no-free-lunch myth and pluralism. I will then turn to Solomonoff's formal, general, complete, and essentially unique theory of universal induction and prediction, rooted in algorithmic information theory and based on the philosophical and technical ideas of Ockham, Epicurus, Bayes, Turing, and Kolmogorov. This theory provably addresses most issues that have plagued other inductive approaches, and essentially constitutes a conceptual solution to the induction problem. Some theoretical guarantees, extensions to (re)active learning, practical approximations, applications, and experimental results are mentioned, but they are not the focus of this talk. I will conclude with some general advice to philosophers and scientists interested in the foundations of induction.

Beyond Calculation: Probabilistic Computing Machines and Universal Stochastic Inference

Vikash Mansinghka, MIT

Visit the website on the previous page for details

Universal Learning vs. No Free Lunch Results

Shai Ben-David, University of Waterloo
Nathan Srebro, TTI-Chicago
Ruth Urner, University of Waterloo

The so called No-Free-Lunch principle is a basic insight of machine learning. It may be viewed as stating that in the lack of prior knowledge (or inductive bias), any learning algorithm may fail on some learnable task. In recent years, several paradigms for "universal learning" have been proposed and advocated. These range from paradigms of almost science-fictional nature, like Automation of science, through practically oriented Deep Belief Networks, to theoretical constructs like Universal Kernels, Universal Priors and Universal Coding for MDL-based learning. In this work we investigate this apparent contradiction by examining and analyzing several possible definitions of universal learning, proving a basic no-free-lunch theorem for such notions and discussing how they apply to the above mentioned universal learning paradigms.

Are Mental Properties Supervenient on Brain Properties?

Joshua Vogelstein, Johns Hopkins University
Jacob Vogelstein, Johns Hopkins University
Carey Priebe, Johns Hopkins University

The "mind-brain supervenience" conjecture suggests that all mental properties are derived from the physical properties of the brain. To address the question of whether the mind supervenes on the brain, we frame a supervenience hypothesis in rigorous statistical terms. Specifically, we propose a modified version of supervenience (called ϵ-supervenience) that is amenable to experimental investigation and statistical analysis. To illustrate this approach, we perform a thought experiment that illustrates how the probabilistic theory of pattern recognition can be used to make a one-sided determination of ϵ-supervenience. The physical property of the brain employed in this analysis is the graph describing brain connectivity (i.e., the brain-graph or connectome). E-supervenience allows us to determine whether a particular mental property can be inferred from one's connectome to within any given positive misclassification rate, regardless of the relationship between the two. This may provide motivation for cross-disciplinary research between neuroscientists and statisticians.

Toward a New Representation for Causation in Dynamic Systems

Denver Dash, Carnegie Mellon University
Mark Voortman, University of Pittsburgh

Over the past twenty years, causal modeling has been a growing discipline within the field of machine learning (ML). Inspired by the rallying cry of Pearl and others, the ML community has provided a wealth of methods for approximate and exact causal discovery and causal reasoning. Causation has a unique place in machine learning and AI for several reasons: Causal models are generative probabilistic models which seek to provide the "most parsimonious" representation for describing an entire system, as opposed to discriminitive classification models which are interested in predicting a fixed set of variables. This may make causal models suitable for a "general AI" agent whose scope is intended to exceed any simple classification problem. Another reason is that causal models aid explanation because they mirror the way many humans internally model the world [Sloman, 2005]. Perhaps most importantly, causal models provide a syntax for reasoning about manipulating elements of the system being modeled. In this paper we explore the limits of the state-of-the-art causal representations, we identify some practical obstacles to their use, and suggest a new representation to help remedy them.

Relations between machine learning problems: An approach to unify the field

http://rml.cecs.anu.edu.au/

LOCATION
Melia Sierra Nevada: Dilar
Friday, December 16th, 07:30 -- 10:30 AM & 4:00 -- 8:00 PM

Bob Williamson Bob.Williamson@anu.edu.au
The Australian National University and NICTA

John Langford jl@yahoo-inc.com
Yahoo! Research

Ulrike von Luxburg ulrike.luxburg@tuebingen.mpg.de
University of Hamburg

Mark Reid mark.reid@anu.edu.au
The Australian National University and NICTA

Jennifer Wortman Vaughan jenn@cs.ucla.edu
University of California, Los Angeles

Abstract
The workshop proposes to focus on relations between machine learning problems. We use "relation" quite generally to include (but not limit ourselves to) notions such as: one type of problem being viewed special case of another type (e.g., classification as thresholded probability estimation); reductions between learning problems (e.g., transforming ranking problems into classification problems); the use of surrogate losses (e.g., replacing misclassification loss with some other, convex loss); relations between sets of learning problems, such as those studied in the (old) theory of "comparison of experiments"; connections between machine learning problems and what could be construed as "economic learning problems" such as prediction markets and forecast elicitation.

The point of studying relations between machine learning problems is that it stands a reasonable chance of being a way to be able to understand the field of machine learning as a whole. It could serve to prevent re-invention, and rapidly facilitate the growth of new methods.

SCHEDULE

7.30-8.00	Introductory Overview
8.00-8.55	**Machine Learning Markets: Putting your money where your mouth is** Amos Storkey
8.55-9.05	Coffee break
9.05-10.00	**Efficient Market Making via Convex Optimization, and a Connection to Online Learning** Jenn Wortman Vaughan
10.00-10.30	**Bounded Regret Sequential Learning using Prediction Markets** Sindhu Kutty and Rahul Sami
16.00-16.55	**We need a BIT more GUTS (=Grand Unified Theory of Statistics)** Peter Grűnwald
16.55-17.25	**Anatomy of a Learning Problem** Mark Reid
17.25-17.40	Coffee break
17.40-18.10	**Degrees of Supervision** Daro Garca-Garca
18.10-19.00	Panel Discussion

INVITED SPEAKERS

Machine Learning Markets: Putting Your Money Where Your Mouth Is

Amos Storkey, University of Edinburgh

I will discuss a number of things I think are stumbling blocks to progress in machine learning and issues in statistical modeling. These include the existence of a plethora of algorithms combined with poor knowledge of the performance of most, the difficulty of principled model combination or principled approaches for building on previous/partial results. I will suggest that the idea of Machine Learning Markets, can help in this, and discuss some of the developments that result from this basic idea. Machine Learning Markets involve extending prediction market mechanisms for doing machine learning. Because Machine Learning Markets perform joint inference through decision making, they can bypass some of the arguments between different statistical paradigms, and lead to interesting, continuously improvable, potentially hierarchical probabilistic models for systems. A by-product of this is that each contribution to the model is evaluated and rewarded in terms of its added value.

Efficient Market Making via Convex Optimization, and a Connection to Online Learning

Jenn Wortman Vaughan, University of California, Los Angeles
Jake Abernethy, University of Pennsylvania
Yiling Chen, Harvard University

A prediction market is a financial market designed to aggregate information. To facilitate trades, prediction markets are often operated by automated market makers. The market maker trades a set of securities with payoffs that depend on the outcome of a future event. For example, the market maker might offer a security that will pay off $1 if and only if a Democrat wins the 2012 US presidential election. A risk neutral trader who believes that the probability of a Democrat winning is p should be willing to purchase this security at any price below p, or sell it at any price above p. The current market price can then be viewed as the traders' collective estimate of how likely it is that a Democrat will win the election.

Market-based estimates have proved to be accurate in a variety of domains, including business, entertainment, and politics. However, when the number of outcomes is very large, it is generally infeasible to run a simple prediction market over the full outcome space. We propose a general framework for the design of securities markets over combinatorial or infinite state or outcome spaces. Our framework enables the design of computationally efficient markets tailored to an arbitrary, yet relatively small, space of securities with bounded payoff. We prove that any market satisfying a set of intuitive conditions must price securities via a convex cost function, which is constructed via conjugate duality. Rather than deal with an exponentially large or infinite outcome space directly, our framework only requires optimization over a convex hull. By reducing the problem of automated market making to convex optimization, where many efficient algorithms exist, we arrive at a range of new polynomial-time pricing mechanisms for various problems.

Our framework also provides new insights into the relationship between market design and machine learning. In particular, we show that the tools that have been developed for online linear optimization are strikingly similar to those we have constructed for selecting pricing mechanisms. This is rather surprising, as the problem of learning in an online environment is semantically quite distinct from the problem of pricing securities in a prediction market: a learning algorithm receives losses and selects weights whereas a market maker manages trades and sets prices. We show that although the two frameworks have very different semantics, they have nearly identical syntax in a very strong sense.

Bounded-Regret Sequential Learning using Prediction Markets

Sindhu Kutty, University of Michigan, Ann Arbor
Rahul Sami, University of Michigan, Ann Arbor

We demonstrate a relationship between prediction markets and online learning algorithms by using a prediction market metaphor to develop a new class of algorithms for learning exponential families with expert advice. The specific problem we consider is that of prediction when data is distributed according to a particular member of an exponential family. In such a case, cost function based prediction markets provide a convenient analytical tool for evaluating performance. Prediction markets also provide a natural technique for learning in an environment where expert advice is not available simultaneously but sequentially; and experts are either honest and informative, or dishonest and adversarial. As in traditional models, we combine advice using weights on experts. However, to exploit these particular features, we use a form of Kelly gambling to relate the weight of an expert to her budget. We give a formal description of this new model along with relevant definitions and show an equivalence between learning maximum likelihood estimates of the natural parameters of an exponential family and combining advice in prediction markets. We provide an abstract architecture for learning in this model that uses this equivalence to simulate a prediction market to update budgets of experts based on their individual loss. We apply this technique to construct a concrete algorithm that achieves bounded-regret.

We need a BIT more GUTS (=Grand Unified Theory of Statistics)

Peter Grűnwald, Centrum voor Wiskunde en Informatica

A remarkable variety of problems in machine learning and statistics can be recast as data compression under constraints: (1) sequential prediction with arbitrary loss functions can be transferred to equivalent log loss (data compression) problems. The worst-case optimal regret for the original loss is determined by Vovk's mixability, which in fact measures how many bits we lose if we are not allowed to use mixture codes in the compression formulation. (2) in classification, we can map each set of candidate classifiers C to a corresponding probability model M. Tsybakov's condition (which determines the optimal convergence rate) turns out to measure how much more we can compress data by coding it using the convex hull of M rather than just M. (3) hypothesis testing in the applied sciences is usually based on p-values, a brittle and much-criticized approach. Berger and Vovk independently proposed calibrated p-values, which are much more robust. Again we show these have a data compression interpretation. (4) Bayesian nonparametric approaches usually work well, but fail dramatically in Diaconis and Freedman's pathological cases. We show that in these cases (and only in these) the Bayesian predictive distribution does not compress the data. We speculate that all this points towards a general theory that goes beyond standard MDL and Bayes.

Anatomy of a Learning Problem

Mark Reid, The Australian National University and NICTA
James Montgomery, The Australian National University
Mindika Premachandra, The Australian National University

In order to relate machine learning problems we argue that we need to be able to articulate what is meant by a single machine learning problem. By attempting to name the various aspects of a learning problem we hope to clarify ways in which learning problems might be related to each other. We tentatively put forward a proposal for an anatomy of learning problems that will serve as scaffolding for posing questions about relations. After surveying the way learning problems are discussed in a range of repositories and services. We then argue that the terms used to describe problems to better understand a range of viewpoints within machine learning ranging from the theoretical to the practical.

Degrees of Supervision

Darío García-Garcıa, The Australian National University
Robert C. Williamson, The Australian National University/NICTA

Many machine learning problems can be interpreted as differing just in the level of supervision provided to the learning process. In this work we provide a unifying way of dealing with these different degrees of supervision. We show how the framework developed to accommodate this vision can deal with the continuum between classification and clustering, while also naturally accommodating less standard settings such as learning from label proportions, multiple instance learning,...All this emanates from a simple common principle: when in doubt, assume the simplest possible classification problem on the data.

Beyond Mahalanobis: Supervised Large-Scale Learning of Similarity

http://www.ttic.edu/nips11simworkshop

LOCATION
Melia Sierra Nevada: Guejar
Friday, December 16th, 07:30 -- 10:30 AM & 4:00 -- 8:00 PM

Gregory Shakhnarovich greg@ttic.edy
TTI-Chicago

Dhruv Batra dbatra@ttic.edu
TTI-Chicago

Brian Kulis kulis@eecs.berkeley.edu
UC-Berkeley and Ohio State University

Kilian Weinberger kilian@seas.wustl.edu
Washington University at St. Louis

Abstract

The notion of similarity (or distance) is central in many problems in machine learning: information retrieval, nearest-neighbor based prediction, visualization of high-dimensional data, etc. Historically, similarity was estimated via a fixed distance function (typically Euclidean), sometimes engineered by hand using domain knowledge. Using statistical learning methods instead to learn similarity functions is appealing, and over the last decade this problem has attracted much attention in the community. Much of this work, however, has focused on a specific, restricted approach: learning a Mahalanobis distance, under a variety of objectives and constraints. This effectively limits the setup to learning a linear embedding of the data. In this workshop, we will look beyond this setup, and consider methods that learn non-linear embeddings of the data, either explicitly via non-linear mappings or implicitly via kernels. We will especially encourage discussion of methods that are suitable for large-scale problems increasingly facing practitioner of learning methods: large number of examples, high dimensionality of the original space, and/or massively multi-class problems.

INVITED SPEAKERS

Online Similarity Learning: From Images to Texts

Samy Bengio, Google
Gal Chechik, Google

Learning a model of similarity between pairs of objects is an important topic in machine learning. It has multiple applications like finding similar images, videos, and text documents. In most of these applications the number of objects is usually very large (at least millions, if not billions), so learning the similarity function needs to be very efficient. A few years ago, we presented OASIS, an online algorithm for scalable image similarity, and showed that it scaled well to very large datasets of images, where each image was represented by thousands of sparse features. While this approach is relevant for objects like images, it does not scale well for objects like text documents, where the number of features often scales with the dictionary, which can be in the millions, and the underlying OASIS model would need memory on the order of square of that number. We thus discuss here two recent models that address this problem, Loreta and Wsabie. Both algorithms work in an online manner using a ranking cost. Wsabie learns to embed objects in a low-dimensional space, while Loreta learns a similarity model in the Riemannian manifold of low rank matrices.

SCHEDULE

7.30-7:40	Opening remarks
7:40-8:10	**Online Similarity Learning: From Images to Texts** Samy Bengio
8:10-8:40	**To Be Announced** Prateek Jain
8:40-9:00	Poster spotlights
9:00-9:30	Coffee break
9.30-10:00	**Learning class-sensitive similarity metric from few examples** Ruslan Salakhutdinov
10:00-10:30	**Estimating Similarities Between Tasks for Multi-Task Learning** Maya R. Gupta
10:30-4:00	Break
4:00-4:30	**Learning multi-modal similarity: a novel multiple kernel learning technique** Gert Lanckriet
4:30-5:00	**Actively learning similarity from the crowd** Adam Kalai
5:00-5:20	**Learning Discriminative Metrics via Generative Models and Kernel Learning** Fei Sha
5:20-5:40	**Learning a Degree-Augmented Distance Metric From a Network** Bert Huang
5:40-6:10	Coffee break
6:10-6:30	**Adaptive Image Similarity: The Sharpening Match** Erik Learned-Miller
6:30-6:50	**Gradient Boosting for Large Margin Nearest Neighbors** Dor Kedem
6:50-7:20	**Learning similarity for recognition is best solved by first learning to recognize** Alexander Berg
7:20-7:50	Discussion
7:50	**Wrap-up and closing remarks** Organizers

Online Similarity Learning: From Images to Texts
Samy Bengio, Google

Please visit website on previous page for more information

Title To Be Announced
Prateek Jain, Microsoft Research

Please visit website on previous page for more information

Learning class-sensitive similarity metric from few examples
Ruslan Salakhutdinov, University of Toronto

In this talk I will introduce a new compositional learning architecture that integrates deep learning models, such as Deep Boltzmann machines, with hierarchical Bayesian models. This new class of models learns novel concepts from very few training examples by learning (a) low-level generic features, (b) high-level class-sensitive features that capture correlations among low-level features, and (c) a category hierarchy for sharing priors over the high-level features that are typical of different kinds of concepts. I will show that the model is able to learn class-sensitive similarity metric from few examples on CIFAR-100 object recognition, handwritten character recognition, and human motion capture datasets. I will also demonstrate how a related model can learn a hierarchy for sharing visual appearance across 200 object categories with applications to object localization and detection task on SUN dataset.

Estimating Similarities Between Tasks for Multi-Task Learning
Maya R. Gupta, University of Washington

Multi-task learning often benefits from incorporating information about how similar the tasks to be co-learned are. Sometimes such information about the task similarity is available as side information, but often there is no such extra information. Even when side information is available about the similarity of tasks, it is not always clear how that information should be quantified or used in multi-task learning formulas. For example, what does it mean if someone tells you these two tasks have similarity one, but these tasks have similarity 1/2, and these other tasks have similarity zero? Do we use those numbers or the log of those numbers? How do we ask domain experts for usable information about the similarities between tasks? In this talk we begin to answer these questions. We focus on multi-task averaging, which can be applied to any scenario where related averages are to be computed. Due to its simplicity, multi-task averaging is amenable to analysis, but despite its simplicity we show that many of its properties depend nonlinearly on the task-similarities. We quantify the optimal similarities between tasks, and we show preliminary results in how well these similarities can be estimated.

Learning multi-modal similarity: a novel multiple kernel learning technique
Gert Lanckriet, UC-San Diego

In many applications involving multimedia data, the definition of similarity between items is integral to several key tasks, e.g., nearest-neighbor retrieval, classification, or visualization. Data in such regimes typically exhibits multiple modalities, such as acoustic and visual content of a video, or audio clips, web documents and art work describing musical artists. Integrating such heterogeneous data to form a holistic similarity space is therefore a key challenge to be overcome in many real-world applications. We present a novel multiple kernel learning technique for integrating heterogeneous data into a single, unified similarity space. Instead of finding a weighted linear combination of base kernels, as in the original MKL formulation, we learn a concatenation of linear projections, where each projection extracts the relevant information from a base kernel's feature space. This new formulation results in a more flexible model than previous methods, that can adapt to the case where the discriminative power of a kernel varies over the data set or feature space. It is more general than the original formulation and contains it as a special case.

Actively learning similarity from the crowd
Adam Kalai, Microsoft Research
Omer Tamuz, Weizmann Institute of Sciences
Ce Liu, Microsoft Research
Serge Belongie, UC-San Diego
Ohad Shamir, Microsoft Research

We introduce an algorithm that, given n objects, learns a similarity matrix over all n2 pairs, from crowdsourced data alone. The algorithm samples responses to adaptively chosen triplet-based relative-similarity queries. Each query has the form "is object a more similar to b or to c?" and is chosen to be maximally informative given the preceding responses. The output is an embedding of the objects into Euclidean space (like MDS); we refer to this as the "crowd kernel." SVMs reveal that the crowd kernel captures prominent and subtle features across a number of domains, such as " is striped" among neckties and "vowel vs. consonant" among letters.

Learning Discriminative Metrics via Generative Models and Kernel Learning
Fei Sha, University of Southern California

Metrics specifying distances between data points can be learned in a discriminative manner or from generative models. In this paper, we show how to unify generative and discriminative learning of metrics via a kernel learning framework. Specifically, we learn local metrics optimized from parametric generative models. These are then used as base kernels to construct a global kernel that minimizes a discriminative training criterion. We consider both linear and nonlinear combinations of local metric kernels. Our empirical results show that these combinations significantly improve performance on classification tasks. The proposed learning algorithm is also very efficient, achieving order of magnitude speedup in training time compared to previous discriminative baseline methods.

Learning a Degree-Augmented Distance Metric From a Network

Bert Huang, University of Maryland
Blake Shaw, Foursquare
Tony Jebara, Columbia University

In many naturally occurring networks, connected nodes tend to have empirical similarities, which is a phenomenon commonly referred to as homophily. It is useful to learn how to relate the network homophily to the measurable features from data. However, because of the inherent structural nature of networks, we should not expect the similarity between connected nodes to behave in a purely pairwise independent manner. In an attempt to address the structural nature of networks, we model similarity between nodes with an added structural component: node degree. We present a new algorithm called degree-distributional metric learning (DDML), which learns a similarity metric and a set of degree-based score functions that together provide a structure-aware, distance-based method for link prediction. DDML is a variant of structure-preserving metric learning (SPML), an algorithm we introduce in the main NIPS 2011 conference. Just like SPML, the algorithm learns the metric and degree preference parameters via stochastic sub-gradient descent, which provably converges in time independent of the size of the network, and thus allows learning from large-scale data. Empirically, the algorithm achieves state-of-the-art accuracy results on medium and large-scale data. This talk will detail the algorithm, its derivation, analysis, and related open problems.

Adaptive Image Similarity: The Sharpening Match

Erik Learned-Miller, University of Massachusetts, Amherst

Many image descriptors, including SIFT, HOG, and color histograms, bin features to introduce some degree of invariance to spatial location. The theme of this workshop suggests using large data sets to learn properties of these bins (the number of bins, their width, their weight masks, their degree of overlap) that optimize the utility of the descriptors in computing similarity. We show, however, that even with the optimal parameters for such bins, it is impossible to design a similarity measure that satisfies certain simple requirements. We then show that by adapting the parameters of the representation on the fly with respect to each pair of images to be compared, the required properties of the similarity measure can be achieved. We call our similarity measure the sharpening match and discuss its excellent performance on certain tasks. We argue that no amount of a priori learning can substitute for the adaptation of certain structural parameters at the time of actual comparison of images.

Gradient Boosting for Large Margin Nearest Neighbors

Dor Kedem, Washington University at St. Louis
Zhixiang Eddie Xu, Washington University at St. Louis
Kilian Q. Weinberger, Washington University at St. Louis

Please visit website on page 30 for more information

Learning similarity for recognition is best solved by first learning to recognize

Alexander C. Berg, Stony Brook University

I will present a sampling of our recent work on large scale recognition and similar image retrieval with a hierarchy of more than 10,000,000 labeled images in more than 10,000 classes as part of the ImageNet project. Results from this work indicate some possible conclusions about similarity learning for computer vision applications: (1) feature representations matter, and what works for recognition may not work for similarity learning, (2) when similarity is based on labels, it may be best to first learn to estimate those labels before learning a similarity function! This work is part of an ongoing collaboration with Jia Deng and Fei-Fei Li at Stanford University.

ACCEPTED POSTERS

Mirror Descent for Metric Learning

Gautam Kunapuli, University of Wisconsin-Madison
Jude Shavlik, University of Wisconsin-Madison

We propose a unified approach to Mahalanobis metric learning: an online, regularized, positive semi-definite matrix learning problem, whose update rules can be derived using the composite objective mirror descent (COMID) framework. This approach admits many different types of Bregman and loss functions which allows for the tailoring of several different classes of algorithms. The most novel contribution is the trace norm regularization, which yields a sparse metric in its eigenspectrum, thus simultaneously performing feature selection along with metric learning. The regularized update rules are derived using composite objective mirror descent, which results in a parallelizable update that can be computed efficiently for a large number of features. The proposed approach is also kernelizable, which allows for metric learning in nonlinear domains.

Good Similarity Learning for Structured Data

Aurelien Bellet, University of Jean Monnet
Amaury Habrard, University of Jean Monnet
Marc Sebban, University of Jean Monnet

Similarity functions play an important role in the performance of many learning algorithms, thus a lot of research has gone into training them. In this paper, we focus on learning similarity functions for structured data. We propose a novel edit similarity learning approach (GESL) driven by the idea of (e,g,t)-goodness, a recent theory that bridges the gap between the properties of a similarity function and its performance in classification. We derive generalization guarantees for our method and provide experimental evidence of its practical interest.

Learning Cross-Lingual Similarities

Jan Rupnik, Jožef Stefan Institute
Andrej Muhič, Jožef Stefan Institute
Primož Škraba, Jožef Stefan Institute

This work focuses on learning a cross-lingual similarity function given monolingual similarity functions and an aligned bi-lingual corpus. We consider two popular approaches to infer the similarity function: similarity via the aligned basis and regression-based similarity. As well as analyzing the theoretical and practical aspects, we propose a general approach and relate it to the aforementioned methods. Finally, we present an evaluation of the different approaches for several choices of monolingual similarity function.

A metric learning perspective of SVM: On the relation of LMNN and SVM

Huyen Do, University of Geneva
Alexandros Kalousis, University of Geneva
Jun Wang, University of Geneva
Adam Woznica, University of Geneva

Support Vector Machines (SVMs) and metric learning algorithms are two very popular learning paradigms with quite distinct learning biases. In this paper we focus on SVMs and LMNN--Large Margin Nearest Neighbor, one of the most prominent metric learning algorithms; we bring them into a unified view and show that they have a much stronger relation than what is commonly thought. We analyze SVMs from a metric learning perspective and cast them as a metric learning problem, a view which helps us uncover the relations of the two algorithms. We show that LMNN can be seen as learning a set of local SVM-like models in a quadratic space. Along the way and inspired by the metric-based interpretation of SVMs we derive a novel variant of SVMs, E-SVM, to which LMNN is even more similar. Moreover this strong connection is also valid for other large margin metric learning algorithms. This result has the potential to lead to theoretical error bounds for LMNN and other large-margin-based metric learning algorithms, using the SVM large margin theory. Finally we use the metric-based view to provide a unified view of SVM, E-SVM and LMNN.

Learning sequence neigbourhood metrics

Justin Bayer, Technische Universität Műnchen
Christian Osendorfer, Technische Universität Műnchen
Patrick van der Smagt, Technische Universität Műnchen

In order to process dynamic data such as sequences we propose to map it to static representations. Short descriptors, e.g. $x \in \mathbb{R}^n$, typically require much less memory and thus make processing of large datasets much more efficient. Also, algorithms tailored towards static data can be applied. In the presence of class labels this mapping can be learned by combining recurrent neural networks, a pooling operation and neighbourhood components analysis as an objective function.

Ground Metric Learning

Marco Cuturi, Kyoto University
David Avis, Kyoto University

Transportation distances have been used for more than a decade now in machine learning to compare histograms of features. They have one parameter: the ground metric, which can be any metric between the features. To date, the only option available to practitioners to set this parameter was to rely on a priori knowledge of the features, which limited considerably the scope of application of transportation distances. We propose to lift this limitation and consider instead algorithms that can learn the ground metric using a training set of labeled histograms.

New Frontiers in Model Order Selection

http://people.kyb.tuebingen.mpg.de/seldin/fimos.html

LOCATION
Melia Sol y Nieva: Ski
Friday, December 16th, 07:30 -- 10:30 AM & 4:00 -- 8:00 PM

Yevgeny Seldin seldin@tuebingen.mpg.de
Max Planck Institute for Intelligent Systems

Koby Crammer koby@ee.technion.ac.il
Technion

Nicolo Cesa-Bianchi nicolo.cesa-bianchi@unimi.it
Universita degli Studi di Milano

François Laviolette francois.laviolette@ift.ulaval.ca
Universite Laval, Quebec

John Shawe-Taylor jst@cs.ucl.ac.uk

Abstract

Model order selection, which is a trade-off between model complexity and its empirical data fit, is one of the fundamental questions in machine learning. It was studied in detail in the context of supervised learning with i.i.d. samples, but received relatively little attention beyond this domain. The goal of our workshop is to raise attention to the question of model order selection in other domains, share ideas and approaches between the domains, and identify perspective directions for future research. Our interest covers ways of defining model complexity in different domains, examples of practical problems, where intelligent model order selection yields advantage over simplistic approaches, and new theoretical tools for analysis of model order selection. The domains of interest span over all problems that cannot be directly mapped to supervised learning with i.i.d. samples, including, but not limited to, reinforcement learning, active learning, learning with delayed, partial, or indirect feedback, and learning with submodular functions.

An example of first steps in defining complexity of models in reinforcement learning, applying trade-off between model complexity and empirical performance, and analyzing it can be found in [1-4]. An intriguing research direction coming out of these works is simultaneous analysis of exploration-exploitation and model order selection trade-offs. Such an analysis enables to design and analyze models that adapt their complexity as they continue to explore and observe new data. Potential practical applications of such models include contextual bandits (for example, in personalization of recommendations on the web [5]) and Markov decision processes.

References:
[1] N. Tishby, D. Polani. “Information Theory of Decisions and Actions”, Perception-Reason-Action Cycle: Models, Algorithms and Systems, 2010.

[2] J. Asmuth, L. Li, M. L. Littman, A. Nouri, D. Wingate, “A Bayesian Sampling Approach to Exploration in Reinforcement Learning”, UAI, 2009.

[3] N. Srinivas, A. Krause, S. M. Kakade, M. Seeger, “Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design”, ICML, 2010.

[4] Y. Seldin, N. Cesa-Bianchi, F. Laviolette, P. Auer, J. Shawe-Taylor, J. Peters, “PAC-Bayesian Analysis of the Exploration-Exploitation Trade-off”, ICML-2011 workshop on online trading of exploration and exploitation.

[5] A. Beygelzimer, J. Langford, L. Li, L. Reyzin, R. Schapire, “Contextual Bandit Algorithms with Supervised Learning Guarantees”, AISTATS, 2011.

SCHEDULE

7:50-8:00	Opening Remarks
8:00-8:45	Invited talk: **Model Selection in Markovian Processes** Shie Mannor
8:45-9:05	Contributed talk: **BErMin: A Model Selection Algorithm for Reinforcement Learning Problems** Amir-massoud Farahmand
9:05-9:15	Break
9:15-9:35	Contributed talk: **Selecting the State-Representation in Reinforcement Learning** Odalric-Ambrym Maillard
9:35-10:20	Invited talk: **Autonomous Exploration in Reinforcement Learning** Peter Auer
10:20-16:00	Break
16:00-16:45	Invited talk: **Model Selection when Learning with Exploration** John Langford
16:45-17:45	Posters
17:45-18:30	Invited talk: **Kernel-Information-Bottleneck: Successive Refinement and Model Order Selection** Naftali Tishby
18:30-19:30	Panel Discussion

INVITED SPEAKERS

Model Selection in Markovian Processes
Shie Mannor , Technion

We address the problem of how to use a sample of trajectories to choose from a candidate set of possible state spaces in different types of Markov processes. Standard approaches to solving this problem for static models use penalized maximum likelihood criteria that take the likelihood of the trajectory into account. Surprisingly, these criteria do not work even for simple fully observable finite Markov processes. We propose an alternative criterion and show that it is consistent. We then provide a guarantee on its performance with finite samples and illustrate its accuracy using simulated data and real-world data. We finally address the question of model selection in Markov decision processes, where the decision maker can actively select actions to assist in model selection.

BErMin: A Model Selection Algorithm for Reinforcement Learning Problems
Amir-massoud Farahmand , McGill University
Csaba Szepesvári , University of Alberta

We consider the problem of model selection in the batch (offline, non-interactive) reinforcement learning setting when the goal is to find an action-value function with the smallest Bellman error among a countable set of candidate functions. We propose a complexity regularization-based model selection algorithm, BErMin, and prove that it enjoys an oracle-like property: the estimator's error differs from that of an oracle, who selects the candidate with the minimum Bellman error, by only a constant factor and a small remainder term that vanishes at a parametric rate as the number of samples increases.

Selecting the State-Representation in Reinforcement Learning
Odalric-Ambrym Maillard , Montanuniversität Leoben
Rémi Munos , INRIA Lille
Daniil Ryabko , INRIA Lille

The problem of selecting the right state-representation in a reinforcement learning problem is considered. Several models (functions mapping past observations to a finite set) of the observations are given, and it is known that for at least one of these models the resulting state dynamics are indeed Markovian. Without knowing neither which of the models is the correct one, nor what are the probabilistic characteristics of the resulting MDP, it is required to obtain as much reward as the optimal policy for the correct model (or for the best of the correct models, if there are several). We propose an algorithm that achieves that, with a regret of order T 2/3 where T is the horizon time.

Autonomous Exploration in Reinforcement Learning
Peter Auer , Montanuniversität Leoben

One of the striking differences between current reinforcement learning algorithms and early human learning is that animals and infants appear to explore their environments with autonomous purpose, in a manner appropriate to their current level of skills. For analyzing such autonomous exploration theoretically, an evaluation criterion is required to compare exploration algorithms. Unfortunately, no commonly agreed evaluation criterion has been established yet. As one possible criterion, we consider in this work the navigation skill of a learning agent after a number of exploration steps. In particular, we consider how many exploration steps are required, until the agent has learned reliable policies for reaching all states in a certain distance from a start state. (Related but more general objectives are also of interest.) While this learning problem can be addressed in a straightforward manner for finite MDPs, it becomes much more interesting for potentially infinite (but discrete) MDPs. For infinite MDPs we can analyze how the learning agent increases its navigation skill for reaching more distant states, as the exploration time increases. We show that an optimistic exploration strategy learns reliable policies when the number of exploration steps is linear in the number of reachable states and in the number of actions. The number of reachable states is not known to the algorithm, but the algorithm adapts to this number.
Joint work with Shiau Hong Lim and Chris Watkins.

Model Selection when Learning with Exploration
John Langford , Yahoo! Research

I will discuss model selection in 4 settings: {Selective Sampling, Partial Feedback} x {Agnostic, Realizable}. In selective sampling, you choose on which examples to acquire a label. In partial feedback, you choose on which label (or action) to discover a reward (or loss). In the agnostic setting, your goal is simply competing a set of predictors. In the realizable setting, one of your predictors is perfect, for varying definitions of perfect.

Kernel-Information-Bottleneck: Successive Refinement and Model Order Selection
Naftali Tishby , The Hebrew University of Jerusalem

The Information Bottleneck method (IB) was introduced as a way of computing approximate minimal sufficient statistics from empirical data, through a continuous trade-off between (model) complexity and accuracy, given the joint distribution of data and relevant variables. The original algorithm for solving the problem was a converging alternating projection (EM, Arimoto Blahut like), but was not guaranteed to converge to the global optimum in general. An important exception was the multivariate Gaussian case, for which the IB recovers the classical Canonical Correlation Analysis (CCA), with an important addition of a principled model dimension estimation through the complexity accuracy trade-off. For this case the optimal representation can be found efficiently even for very large datasets. In this paper we present a recent generalization of Gaussian IB, using the Kernel trick, which corresponds to the Kernel-CCA, with the additional principled information theoretic model-order estimation. This new algorithm not only makes the IB practical for large classes of real data, but provides a systematic approach to both dimension reduction and Kernel selection, by optimizing the information accuracy complexity curve. Moving from low to high model complexity requires in general a change in the Kernel structure in ways that can simplify the representation. Exceptions to this are self-similar structures (like fractals, or some natural datasets) where the same Kernel can be optimal for all scales. We provide examples of applications to both images and time-series data.
Joint work with Nori Jacoby

Bayesian Optimization, Experimental Design and Bandits

http://www.cs.ubc.ca/~hutter/nips2011workshop

LOCATION
Melia Sierra Nevada: Hotel Bar
Friday, December 16th, 07:30 -- 10:30 AM & 4:00 -- 8:00 PM

Nando de Freitas nando@cs.ubc.ca
University of British Columbia

Roman Garnett rgarnett@andrew.cmu.edu
Carnegie Mellon University

Frank Hutter hutter@cs.ubc.ca
University of British Columbia

Michael Osborne mosb@robots.ox.ac.uk
University of Oxford

Abstract
Recently, we have witnessed many important advances in learning approaches for sequential decision making. These advances have occurred in different communities, who refer to the problem using different terminology: Bayesian optimization, experimental design, bandits (x-armed bandits, contextual bandits, Gaussian process bandits), active sensing, personalized recommender systems, automatic algorithm configuration, reinforcement learning and so on. These communities tend to use different methodologies too. Some focus more on practical performance while others are more concerned with theoretical aspects of the problem. As a result, they have derived and engineered a diverse range of methods for trading off exploration and exploitation in learning. For these reasons, it is timely and important to bring these communities together to identify differences and commonalities, to propose common benchmarks, to review the many practical applications (interactive user interfaces, automatic tuning of parameters and architectures, robotics, recommender systems, active vision, and more), to narrow the gap between theory and practice and to identify strategies for attacking high dimensionality.

INVITED SPEAKERS

Invited Talk 1: To Be Announced
Remi Munos

See website above for details.

Contributed Talk 1: Dynamic Batch Bayesian Optimization
Javad Azimi
Ali Jalali
Xiaoli Fern

See website above for details.

SCHEDULE

7.30-7:50	Welcome and introduction to Bayesian optimization
7:50-8.20	Invited Talk 1: **To Be Announced** Remi Munos
8.20-8.40	Contributed Talk 1: **Dynamic Batch Bayesian Optimization** Javad Azimi, Ali Jalali, and Xiaoli Fern.
8.40-9.00	Contributed Talk 2: A **penny for your thoughts: the value of information in recommendation systems** Alexandre Passos, Jurgen Van Gael, Ralf Herbrich and Ulrich Paquet.
9.00-9.20	Break
9.20-9.50	Spotlights of poster presentations
9.50-10.30	Poster session
16.00-16.30	**Information-theoretic Regret Bounds for (Contextual) Gaussian Process Bandit Optimization** Andreas Krause
16.30-17.00	Invited Talk 3: **To Be Announced** Csaba Szepesvari
17.00-17.30	Poster Session
17.30-17.50	Break
17.50-18.20	Invited Talk 4: **Gaussian Process Bandits applied to Tree Search** Louis Dorard
18.20-19.00	Panel discussion

Contributed Talk 2: A penny for your thoughts: the value of information in recommendation systems
Alexandre Passos
Jurgen Van Gael
Ralf Herbrich
Ulrich Paquet

See website above for details.

Information-theoretic Regret Bounds for (Contextual) Gaussian Process Bandit Optimization
Andreas Krause, ETH Zurich and Caltech

Many applications require optimizing an unknown, noisy function that is expensive to evaluate. We formalize this task as a multi-armed bandit problem, where the payoff function is either sampled from a Gaussian process (GP) or has low RKHS norm. We resolve the important open problem of deriving regret bounds for this setting, which imply novel convergence rates for GP optimization. We analyze GP-UCB, an intuitive upper-confidence based algorithm, and bound its cumulative regret in terms of maximal information gain, establishing a novel connection between GP optimization and experimental design. Moreover, by bounding the latter in terms of operator spectra, we obtain explicit sublinear regret bounds for many commonly used covariance functions. In some important cases, our bounds have surprisingly weak dependence on the dimensionality. We also present results for the contextual bandit setting, where in each round, additional information is provided to the decision maker and must be taken into account. We empirically evaluate the approach on sensor selection and automated vaccine design problems. This is joint work with Niranjan Srinivas, Sham Kakade, Matthias Seeger and Cheng Soon Ong.

Invited Talk 3: To Be Announced
Csaba Szepesvari

See the website on the previous page for details.

Invited Talk 4: Gaussian Process Bandits applied to Tree Search
Louis Dorard

See the website on the previous page for details.

ACCEPTED POSTERS

Analysis of Thompson Sampling for the multi-armed bandit problem
Shipra Agrawal and Navin Goyal.

Dynamic trees for online analysis of massive data
Christoforos Anagnostopoulos and Robert B. Gramacy.

Dynamic Batch Bayesian Optimization
Javad Azimi, Ali Jalali, and Xiaoli Fern.

Implementations of Algorithms for Hyper-parameter Selection
James Bergstra.

An Optimal Algorithm for Linear Bandits
Nicolò Cesa-Bianchi and Sham Kakade.

Regret Bounds for GP Bandits Without Observation Noise
Nando de Freitas*, Alex Smola, and Masrour Zoghi.

Contextual Bandits for Information Retrieval
Katja Hofmann, Shimon Whiteson, and Maarten de Rijke.

Information-Greedy Global Optimization
Philipp Hennig and Christian J. Schuler.

Bayesian Active Learning for Gaussian Process Classification
Neil M. T. Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel.

Bayesian Optimization With Censored Response Data
Frank Hutter*, Holger Hoos, and Kevin Leyton-Brown.

Sequential Design of Computer Experiments for the Estimation of a Quantile with Application to Numerical Dosimetry
M. Jala, C. Levy-Leduc and E. Moulines, E. Conil and J. Wiart.

On the efficiency of Bayesian bandit algorithms from a frequentist point of view
Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier.

Bayesian Safe Exploration in Markov Decision Processes
Teodor Mihai Moldovan and Pieter Abbeel.

A penny for your thoughts: the value of information in recommendation systems
Alexandre Passos, Jurgen Van Gael, Ralf Herbrich and Ulrich Paquet.

Opportunity Cost in Bayesian Optimization
Jasper Snoek, Hugo Larochelle, and Ryan Prescott Adams.

Adapting Control Policies for Expensive Systems to Changing Environments
Matthew Tesch, Jeff Schneider, and Howie Choset.

Machine Learning for Sustainability

http://people.csail.mit.edu/kolter/mlsust11

LOCATION
Melia Sierra Nevada: Guejar
Friday, December 16th, 07:30 -- 10:30 AM & 4:00 -- 8:00 PM

Thomas G. Dietterich tgd@cs.orst.edu
Oregon State University

J. Zico Kolter kolter@csail.mit.edu
Massachusetts Institute of Technology

Matthew Brown m.brown@bath.ac.uk
University of Bath

Abstract

Sustainability problems pose one of the greatest challenges facing society. Humans consume more than 16TW of power, about 84% of which comes from unsustainable fossil fuels. In addition to simply being a finite resource, the carbon released from fossil fuels is a significant driver of climate change and could have a profound impact on our environment. In addition to carbon releases, humans are modifying the ecosphere in many ways that are leading to large changes in the function and structure of ecosystems. These include huge releases of nitrogen from fertilizers, the collapse and extinction of many species, and the unsustainable harvest of natural resources (e.g., fish, timber). While sustainability problems span many disciplines, several tasks in this space are fundamentally prediction, modeling, and control tasks, areas where machine learning can have a large impact. Many of these problems also require the development of novel machine learning methods, particularly methods that can scale to very large spatio-temporal problem instances.

In recent years there has been growing interest in applying machine to problems of sustainability, spanning applications in energy, environmental management, and climate modeling. The goal of this workshop will be to bring together researchers from both the machine learning and sustainability application fields to continue and build upon this emerging area. The talks and posters will span general discussions of sustainability issues, specific sustainability-related data sets and problem domains, and ongoing work on developing and applying machine learning techniques to these tasks.

SCHEDULE

7.30-8.00	Introduction and Opening Remarks
8.00-8.30	**Machine Learning Challenges in Building Energy Management: Energy Disaggregation as an Example** Mario Berges
8.30-9.00	**Dynamic Resource Allocation in Conservation Planning** Andreas Krause
9.00-9.20	Coffee Break
9.20-10.30	Poster Session
10.30-3.30	Break and Poster Session Continues
4.00-4.30	**Enabling intelligent management of the biosphere** Drew Purves
4.30-5.00	**Climate Informatics** Claire Monteleoni
5.00-5.30	**Machine Learning for Hydrology, Water Monitoring and Environmental Sustainability** Kevin Swersky
5.30-5.50	Coffee Break
5.50-6.20	**Putting the “Smarts” into the Smart Grid: A Grand Challenge for Artificial Intelligence** Alex Rogers
6.20-7.00	Panel and Open Discussion

INVITED SPEAKERS

To view the abstracts for the following talks, please visit the website at the top of this page

Machine Learning Challenges in Building Energy Management: Energy Disaggregation as an Example
Mario Bergés, Carnegie Melon University

Dynamic Resource Allocation in Conservation Planning
Andreas Krause, ETH Zurich

Enabling intelligent management of the biosphere
Drew Purves, Microsoft Research Cambridge

Climate Informatics
Claire Monteleoni, George Washington University

Machine Learning for Hydrology, Water Monitoring and Environmental Sustainability
Kevin Swersky, Aquatic Informatics

Putting the ”Smarts” into the Smart Grid: A Grand Challenge for Artificial Intelligence
Alex Rogers, University of Southampton

From statistical genetics to predictive models in personalized medicine

http://agbs.kyb.tuebingen.mpg.de/wikis/bg/NIPSPM11

LOCATION
Melia Sol y Nieve: Slalom
Friday, December 16th, 07:30 -- 10:30 AM & 4:00 -- 8:00 PM

Karsten Borgwardt karsten.borgwardt@tuebingen.mpg.de
Oliver Stegle oliver.stegle@tuebingen.mpg.de
Max Planck Institutes Tubingen, Germany

Shipeng Yu shipeng.yu@siemens.com
Glenn Fung glenn.fung@siemens.com
Faisal Farooq f.farooq@siemens.com
Balaji Krishnapuram balaji.krishnapuram@siemens.com
Siemens Healthcare, USA

Abstract
Technological advances to profile medical patients have led to a change of paradigm in medical prognoses. Medical diagnostics, traditionally carried out by medical experts, is increasingly complemented by large-scale data collection and quantitative genome-scale molecular measurements. Data that are already available as of today or are to enter medical practice in the near future include personal medical records, genotype information, diagnostic tests, proteomics and other emerging omics data types.

The purpose of this 2nd cross-discipline workshop is to bring together researchers from machine learning, statistical genetics and healthcare who are interested in problems and applications of predictive models in the field of personalized medicine. The goal of the workshop will be to bridge the gap between the theory of predictive models and statistical genetics with respect to medical applications and the pressing needs of the healthcare community. The workshop will promote an exchange of ideas, helping to identify important and challenging applications as well as the discovery of possible synergies. Ideally, we hope that such discussion will lead to interdisciplinary collaborations with resulting collaborative grant submissions. The emphasis will be on the statistical and engineering aspects of predictive models and how it relates to practical medical and biological problems.

SCHEDULE

7.30-7.35	Welcome and Introduction
7.35-8.25	**The personal genome in the clinic: problems and some initiatives in Spain** Joaquin Dopazo
8.25-9.15	**New advances in the genetics of melanoma: how can it contribute to personalized medicine?** Florence Demenais
9.15-10.05	Poster spotlight, poster session and coffee break
10.05-10.30	**Detecting similar high-dimensional responses to experimental factors from human and model organism** Tommi Suvitaival
16.00-16.05	Welcome and Introduction
16.05-16.55	TBD Bertram Muller-Myhsok
16.55-17.20	**Inferring a measure of physiological age from multiple ageing related phenotypes** David Knowles
17.20-18.20	Poster spotlight, poster session and coffee break
18.20-18.45	**Accuracy test for genome wide selection of bio-markers** Adam Kowalczyk
18.45-20.00	Panel discussion

INVITED SPEAKERS

The personal genome in the clinic: problems and some initiatives in Spain

Joaquin Dopazo, Genomics Department, CIPF, Valencia, Spain

The popularization of massive sequencing technologies bring the promises of personalized medicine closer to the reality. However, there are still problems in the way genomic data is managed, analyzed and, in particular, converted to useful information suitable of being integrated in the patient health record. The Medical Genome Project and the FutureClinic projects are public-private partnership that emerges to bridge the gap between the genome sequencing and the clinic application in Andalucia and Valencia communities, respectively.

New advances in the genetics of melanoma: how can it contribute to personalized medicine?

Florence Demenais, Inserm-University Paris Diderot, Fondation Jean-Dausset, Paris, France

Incidence of melanoma continues to rise in fair skinned populations worldwide. When diagnosed at a late stage, melanoma shows poor survival and is the most fatal of all skin cancers. However, recent advances in the development of new targeted therapies show the benefit of improving knowledge in the biological mechanisms underlying this cancer. Melanoma results from genetic and environmental factors. The only established environmental factor is sun exposure, but its relationship with melanoma is complex. Patterns of melanoma familial aggregation indicate a substantial hereditary component. Genetic susceptibility to melanoma shows a wide spectrum of genetic variation from rare mutations conferring high risk to common genetic variants conferring moderate risk. Various strategies can be used to identify these predisposing genes.

We have recently contributed to genome-wide association studies (GWAS) that have been successful in identifying 16 loci harbouring common genetic variants associated with melanoma and melanoma-related phenotypes (pigmentation, nevi). However, the functional role of the genetic variants within these loci is unknown for the most part. By taking MC1R (melanocortin 1 receptor) gene as a paradigm, we show that genotype imputations can greatly help in identifying the functional variants within a disease-associated locus but sequencing is also needed to provide the full picture of the effect of the variants. The loci identified by GWAS explain only a small part of the genetic component of melanoma. Thus, many genes remain to be discovered. To identify new genetic variants, we are proposing a new strategy based on animal models of spontaneous melanoma with characteristics comparable to those of human melanoma. These animal models can not only provide a powerful tool for gene discovery but can greatly facilitate the study of the effect of genetic predisposition on somatic alterations and gene expression in tumours.

Regarding rarer variants, we have recently contributed to the identification of a germline missense substitution in MITF (microphthalmia-associated transcription factor), a candidate gene that plays a role in melanoma and renal cell carcinoma (RCC). The Mi-E318K substitution was found to occur at a significantly higher frequency in genetically enriched patients affected with melanoma, RCC or both cancers, when compared with controls. Overall, Mi-E318K carriers had a higher than fivefold increased risk of developing melanoma, RCC or both cancers. Codon 318 is located in a small-ubiquitin-like modifier (SUMO) consensus site and Mi-E318K severely impairs sumoylation of MITF protein. Mi-E318K was found to modify the gene expression patterns controlled by MITF. This study provides insights into the link between SUMOylation, transcription and cancer.

Altogether, these results show that the study of genetic predisposition to melanoma, and cancer in general, can improve our understanding of the mechanisms underlying tumour development. In a longer run, this can provide novel therapeutic targets and it can allow to better define prevention and surveillance strategies in targeted groups of subjects.

Detecting similar high-dimensional responses to experimental factors from human and model organism

Tommi Suvitaival, Aalto University
Ilkka Huopaniemi, Aalto University
Matej Oresic, VTT Technical Research Centre of Finland
Samuel Kaski, Aalto University and University of Helsinki

We present a Bayesian model for analyzing the effect of multiple experimental factors in two-species studies without the requirement of a priori known matching. From model studies of human diseases, conducted using *omics technologies and various model organisms, the question emerges: is there something similar in the molecular responses of the different organisms under certain conditions, such as healthy vs. diseased? Our approach provides a generative model for the task of analyzing multi-species data, naturally taking into account the additional information about the affecting factors such as gender, age, treatment, or disease status.

To Be Announced

Bertram Műller-Myhsok, Max Planck Institute for Psychiatry

See the website on the previous page for details.

Inferring a measure of physiological age from multiple ageing related phenotypes

David Knowles, Cambridge University, UK

What is ageing? One definition is simultaneous degradation of multiple organ systems. Can an individual be said to be old or young for their (chronological) age in a scientifically meaningful way? We investigate these questions statistically using ageing related phenotypes measured on the 12,000 female twins in the Twins UK study [4]. We propose a simple linear model of ageing, which allows a latent adjustment Δ to be made to an individuals' chronological age to give their physiological age, shared across the observed phenotypes. We note problems with the analysis resulting from the linearity assumption and show how to alleviate these issues using a non-linear extension. We find more gene expression probes are significantly associated with our measurement of physiological age than to chronological age.

Accuracy test for genome wide selection of bio-markers

Adam Kowalczyk, The University of Melbourne
Eder Kikianty, The University of Melbourne
Fan Shi, The University of Melbourne

Biochemistry of the cell involves multitude of very complex nonlinear interactions between proteins coded for by DNA. Thus the in-silico searches for genome search for disease related bio-markers have to consider non-linear interactions between millions of directly measured features such as SNP calls. This poses unprecedented statistical and computational challenges for feature selection and reduction techniques. This paper argues that constructive answers to these challenges are feasible. We focus on presentation of a statistical test for feature selection with sufficient statistical power to overcome principled multiple test correction in an exhaustive evaluation of hundreds of billions of pairwise interactions. For an empirical validation of the methodology we show replication of filtered interactions in multiple independent Genome Wide Association Studies (GWAS) of the same diseases, namely, Celiac and Type 2 Diabetes.

Machine Learning Meets Computational Photography

http://webdav.is.mpg.de/pixel/workshops/mlmcp-nips2011/index.html

LOCATION
Melia Sol y Nieve: Snow
Saturday, 07:30 -- 10:30 AM & 4:00 -- 7:00 PM

Michael Hirsch michael.hirsch@ucl.ac.uk
University College London

Stefan Harmeling stefan.harmeling@tuebingen.mpg.de
Max Planck Institute for Intelligent Systems

Rob Fergus fergus@cs.nyu.edu
New York University

Peyman Milanfar milanfar@ee.ucsc.edu
University of California at Santa Cruz

Abstract

In recent years, computational photography (CP) has emerged as a new field that has put forward a new understanding and thinking of how to image and display our environment. Besides addressing classical imaging problems such as deblurring or denoising by exploiting new insights and methodology in machine learning as well as computer and human vision, CP goes way beyond traditional image processing and photography.

By developing new imaging systems through innovative hardware design, CP not only aims at improving existing imaging techniques but also aims at the development of new ways of perceiving and capturing our surroundings. However, CP is not only about to redefine “everyday” photography but also aims at applications in scientific imaging, such as microscopy, biomedical imaging, and astronomical imaging, and can thus be expected to have a significant impact in many research areas.

After the great success of last year’s workshop on CP at NIPS, this workshop proposal tries to accommodate the strong interest in a follow-up workshop expressed by many workshop participants last year. The objectives of this workshop are: (i) to give an introduction to CP, present current approaches and report about the latest developments in this fast-progressing field, (ii) spot and discuss current limitations and present open problems of CP to the NIPS community, and (iii) to encourage scientific exchange and foster interaction between researchers from machine learning, neuro science and CP to advance the state of the art in CP.

The tight interplay between both hardware and software renders CP an exciting field of research for the whole NIPS community, which could contribute in various ways to its advancement, be it by enabling new imaging devices that are possible due to the latest machine learning methods or by new camera and processing designs that are inspired by our neurological understanding of natural visual systems.

Thus the target group of participants are researchers from the whole NIPS community (machine learning and neuro science) and researchers working on CP and related fields.

SCHEDULE

7.30-7.32	Opening remarks
7.32-8.12	Invited talk: **From Image Restoration to Compressive Sampling in Computational Photography. A Bayesian Perspective.** Rafael Molina
8.12-8.52	Invited talk: **Old and New algorithm for Blind Deconvolution** Yair Weiss
8.52-9.10	Coffee Break
9.10-9.50	Invited talk: **The Light Field Camera: Extended Depth of Field, Aliasing and Superresolution** Paolo Favaro
9.50-10.30	Invited talk: **Efficient Regression for Computational Photography: from Color Management to Omnidirectional Superresolution** Maya Gupta
10.30-16.00	Break
16.00-16.40	Invited talk: **To Be Announced** Hendrik Lensch
16.40-17.20	Invited talk: **Superresolution imaging - from equations to mobile applications** Filip Šroubek
17.20-17:40	Coffee Break
17.40-18.20	Invited talk: **Modeling the Digital Camera Pipeline: From RAW to sRGB and Back** Michael Brown
18.20-19.00	Invited talk: **To Be Announced**
19.00-19.02	Closing remarks

INVITED SPEAKERS

Invited talk: From Image Restoration to Compressive Sampling in Computational Photography. A Bayesian Perspective.
Rafael Molina, Universidad de Granada

In numerous applications where acquired images are degraded and where either improving the quality of the imaging system or reproducing the scene conditions in order to acquire another image is not an option, computational approaches provide a powerful means for the recovery of lost information. Image recovery is the process of estimating the information lost due to the acquisition or processing system and obtaining images with high quality, additional information, and/or resolution from a set of degraded images. Three specific areas of image recovery are today of high interest. The first one is image restoration, blind deconvolution, and super-resolution, with application, for instance, on surveillance, remote sensing, medical and nano-imaging applications, and improving the quality of photographs taken by hand-held cameras. The second area is compressive sensing (CS). CS reformulates the traditional sensing processes as a combination of acquisition and compression, and traditional decoding is replaced by recovery algorithms that exploit the underlying structure of the data. Finally, the emerging area of computational photography has provided effective solutions to a number of photographic problems, and also resulted in novel methods for acquiring and processing images. Image recovery is related to many problems in computational photography and, consequently, its algorithms are efficiently utilized in computational photography tasks. In addition, image recovery research is currently being utilized for designing new imaging hardware. In this talk, we will provide a brief overview of Bayesian modeling and inference methods for image recovery and the very related of compressive sensing, and computational photography.

Invited talk: Old and New algorithm for Blind Deconvolution
Yair Weiss, Hebrew University of Jerusalem

I will discuss blind deconvolution algorithms that have been successfully used in the field of communications for several decades and how they can be adapted to the problem of blind deconvolution of images. This yields algorithms that can be rigorously shown to recover the correct blur kernel under certain conditions. I will also discuss the relationship between these algorithms and some recent heuristic algorithms for blind image deconvolution.

Invited talk: The Light Field Camera: Extended Depth of Field, Aliasing and Superresolution
Paolo Favaro, Heriot-Watt University and University of Edinburgh

Portable light field cameras have demonstrated capabilities beyond conventional cameras. In a single snapshot, they enable digital image refocusing, i.e., the ability to change the camera focus after taking the snapshot, and 3D reconstruction. We show that they also achieve a larger depth of field than conventional cameras while maintaining the ability to reconstruct detail at high resolution. More interestingly, we show that their depth of field is essentially inverted compared to regular cameras. Crucial to the success of the light field camera is the way it samples the light field, trading off spatial vs. angular resolution, and how aliasing affects the light field. We present a novel algorithm that estimates a full resolution sharp image and a full resolution depth map from a single input light field image. The algorithm is formulated in a variational framework and it is based on novel image priors designed for light field images. We demonstrate the algorithm on synthetic and real images captured with our own light field camera, and show that it can outperform other computational camera systems.

Invited talk: Efficient Regression for Computational Photography: from Color Management to Omnidirectional Superresolution
Maya Gupta, University of Washington

Many computational photography applications can be framed as low-dimensional regression problems that require fast evaluation of test samples for rendering. In such cases, storing samples on a grid or lattice that can be quickly interpolated is often a practical approach. We show how to optimally solve for such a lattice given non-lattice data points. The resulting lattice regression is fast and accurate. We demonstrate its usefulness for two applications: color management, and superresolution of omnidirectional images.

Invited talk: TBA
Hendrik Lensch, University of Tűbingen
See the website on the previous page for details.

Invited talk: Superresolution imaging - from equations to mobile applications
Filip Šroubek, Institute of Information Theory and Automation

In the last five years we have witnessed a rapid improvement of methods that perform image restoration, such as, denoising, deconvolution and superresolution. We will provide a brief mathematical background to superresolution as an optimization problem and summarize our contribution. Specifically, we will talk about robustness to misregistration, an extension to space-variant cases and a fast converging method of augmented Lagrangian suitable for constrained optimization problems. We will also give an overview of our past and ongoing commercial applications in which superresolution plays a key role.

Invited talk: Modeling the Digital Camera Pipeline: From RAW to sRGB and Back
Michael Brown, National University of Singapore

This talk presents a study of the in-camera imaging process through an extensive analysis of more than 10,000 images from over 30 cameras. The goal is to investigate if output image values (i.e. sRGB) can be transformed to physically meaningful values, and if so, when and how this can be done. From our analysis, we show that the conventional radiometric model fits well for image pixels with low color saturation but begins to degrade as color saturation level increases. This is due to a color mapping process which includes gamut mapping in the in-camera processing that cannot be modeled with conventional methods. To address this issue we introduce a new imaging model for radiometric calibration together with an effective calibration scheme that allows us to compensate for the nonlinear color correction to convert non-linear sRGB images to CCD RAW responses.

Fourth International Workshop on Machine Learning and Music: Learning from Musical Structure

WS15

https://sites.google.com/site/musicmachinelearning11/

LOCATION
Melia Sierra Nevada: Dilar
Saturday, 07:30 -- 10:30 AM & 4:00 -- 8:00 PM

Rafael Ramirez rafael.ramirez@upf.edu
Universitat Pompeu Fabra

Darrell Conklin darrell conklin@ehu.es
University of the Basque Country

Douglas Eck deck@google.com
Ryan Rifkin rif@google.com
Google

Abstract

With the current explosion and quick expansion of music in digital formats, and the computational power of modern systems, research on machine learning and music is gaining increasing popularity. As complexity of the problems investigated by researchers on machine learning and music increases, there is a need to develop new algorithms and methods to solve these problems. The focus of this workshop is on novel methods which take into account or benefit from musical structure. MML 2011 aims to build on the previous three successful MML editions, MML08, MML09 and MML10. It has been convincingly shown that many useful applications can be built using features derived from short musical snippets (chroma, MFCCs and related timbral features, augmented with tempo and beat representations). Given the great advances in these applications, higher level aspects of musical structure such as melody, harmony, phrasing and rhythm can now be given further attention, and we especially welcome contributions exploring these areas. The MML 2011 workshop intends to concentrate on machine learning algorithms employing higher level features and representations for content-based music processing.

Papers in all applications on music and machine learning are welcome, including but not limited to automatic classification of music (audio and MIDI), style-based interpreter recognition, automatic composition and improvisation, music recommender systems, genre and tag prediction, score alignment, polyphonic pitch detection, chord extraction, pattern discovery, beat tracking, and expressive performance modeling. Audio demonstrations are encouraged when indicated by the content of the paper.

INVITED SPEAKERS

A Topic Model for Melodic Sequences

Athina Spiliopoulou, University of Edinburgh
Amos Storkey, University of Edinburgh

Modeling the real-world complexity of music is an interesting problem for machine learning. In Western music, pieces are typically composed according to a system of musical organization, rendering musical structure as one of the fundamentals of music. Nevertheless, characterizing this structure is particularly difficult, as it depends not only on the realization of several musical elements, such as scale, rhythm and meter, but also on the relation of these elements both within single time frames

SCHEDULE

7.30-8.10	Invited Talk: **To Be Announced**
8.10-8.15	Coffee break
8.15-8.40	**A Topic Model for Melodic Sequences** Athina Spiliopoulou
9.40-9.05	**Hidden Beyond MIDIs Reach: Feature Extraction and Machine Learning with Rich Symbolic Formats in music21** Michael S. Cuthbert
8.05-9.30	**A Unified Probabilistic Model of Note Combinations and Chord Progressions** Kazuyoshi Yoshii
9.30-9.40	Coffee break
9.40-10.05	**Learning melodic analysis rules** David Rizo
10.05-10.30	**Learning musical motives through spectral clustering** Alberto Pinto
16.00-16.40	Invited Talk: **To Be Announced**
16.40-16.50	Coffee break
16.50-17.15	**Modeling the Acoustic Structure of Musical Emotion with Deep Belief Networks** Erik M. Schmidt
17.15-17.40	**Multi-Timescale Principal Mel Spectral Components for Automatic Annotation of Music Audio** Philippe Hamel
17.40-18.05	**Getting Into the Groove with Hierarchical Probabilistic Latent Component Analysis** Qingyuan Kong
18.05-18.30	**Hit Song Science Once Again a Science?** Yizhao Ni
18.30-18.45	Coffee break
18.45-19.10	**Brain-Computer Interfaces for Music Recommendation** Kira Radinsky
19.10-19.35	**Automatic Detection of Ornamentation in Flamenco** Francisco Gomez
19.35-20.00	**This is the Remix: Structural Improvisation using Automated Pattern Discovery** Sean Whalen

and across time. This results in an infinite number of possible variations, even within pieces from the same musical genre, which are typically built according to a single musical form. In this work, we are interested in modeling music melody as a generative process that can be learned directly from a set of musical pieces belonging to the same genre. We want to avoid making assumptions explicit to music and thus only consider models that are able to automatically capture the complex relations that occur in the data.

Hidden Beyond MIDI's Reach: Feature Extraction and Machine Learning with Rich Symbolic Formats in music21

Michael S. Cuthbert, M.I.T.
Chris Ariza, M.I.T.
Jose Cabal-Ugaz, M.I.T.
Beth Hadley, M.I.T.
Neena Parikh, M.I.T.

Symbolic Music Information Retrieval (MIR) and MIDI Feature Extraction are so wedded in the field that they are almost considered synonyms for one another. Yet a wealth of data about notational characteristics and hidden, implied notes are inaccessible in MIDI and require richer data formats not supported by existing musical machine learning tools. Music21's ability to extract features from richer formats and to bring out hidden information improves the accuracy of symbolic MIR and opens up new repertories for study.

A Unified Probabilistic Model of Note Combinations and Chord Progressions

Kazuyoshi Yoshii,
Matthias Mauch,
Masataka Goto,
National Institute of Advanced Industrial Science and Technology

This paper presents a unified simultaneous and sequential model for note combinations and chord progressions. In chord progression analysis, n-gram models have often been used for modeling temporal sequences of chord labels (e.g., C major, D minor, and E# seventh). These models require us to specify the value of n and define a limited vocabulary of chord labels. On the other hand, our model is designed to directly modeling temporal sequences of note combinations without specifying the value of n, because we aim to use our model as a prior distribution on musical notes in polyphonic music transcription. To do this, we extend a nonparametric Bayesian n-gram model that was designed for modeling sequences of words in the field of computational linguistics. More specifically, our model can accept any combinations of notes as chords and allows each chord appearing in a sequence to have an unbounded and variable-length context. All possibilities of n are taken into account when predicting a next chord given precedent chords. Even when an unseen note combination (chord) emerges, we can estimate its n-gram probability by referring to its 0-gram probability, i.e., the combinatorial probability of note components. We tested our model by using the ground-truth chord annotations and automatic chord recognition results of the Beatles songs.

Learning melodic analysis rules

Placido R. Illescas, University of Alicante
David Rizo, University of Alicante
Jose Manuel Inesta, University of Alicante
Rafael Ramirez, Universitat Pompeu Fabra

Musical analysis is a mean to better understand the composer's intentions when creating a piece and can be used as an intermediate description of a musical work for other purposes, e.g. expressive performance or music comparison. A musical analysis can be decomposed in melodic, harmonic, and tonal function analyzes. Melodic analysis studies the stylistic characteristics of a note from a contrapunctal point of view, while tonal and harmonic analyzes investigate chord roles in particular musical pieces. In this work we focus on automatic melodic analysis. One question that arises when building a melodic analysis system using a-priori music theory is whether it is possible to automatically extract analysis rules from examples, and how similar are those learnt rules compared to music theory rules. This work investigates this question, i.e. given a dataset of analyzed melodies our objective is to automatically learn analysis rules and to compare them with music theory rules.

Learning musical motives through spectral clustering

Alberto Pinto, Universita degli Studi di Milano

Spectral clustering methods are getting more and more attention in many fields of investigation for analysis and classification tasks. Here we present a spectral clustering based method for musical motif identification and classification. Scores are represented like a graph of segments which are ranked depending on their centrality within the network itself. Segments with higher centrality are more likely to be relevant for music summarization. An experimental musicological analysis has been performed on J.S.Bach's 2-part Inventions to prove the effectiveness of the method.

Modeling the Acoustic Structure of Musical Emotion with Deep Belief Networks

Erik M. Schmidt, Drexel University
Youngmoo E. Kim, Drexel University

The problem of automated recognition of emotional content (mood) within music has been the subject of increasing attention among the music information retrieval (Music-IR) research community. While there have been many advances in machine learning systems for estimating human emotional response to music, very little progress has been made in terms of compact or intuitive feature representations. Current methods typically focus on combining several feature domains (e.g. loudness, timbre, harmony, rhythm), oftentimes as many as possible, followed by feature selection and dimensionality reduction techniques. While these methods can lead to enhanced classification performance, they leave much to be desired in terms of understanding the complex relationship between acoustic content and emotional associations. In this talk, we will discuss methods to learn representations of music audio that are specifically optimized for the prediction of emotion.

Multi-Timescale Principal Mel Spectral Components for Automatic Annotation of Music Audio

Philipp Hamel, Universite de Montreal
Douglas Eck, Google
Yoshua Bengio, Universite de Montreal

Automatic annotation is the task of applying semantic descriptors, or tags, to music audio. In other words, the goal is to learn how to

describe, in words, the audio content of a given music clip. Feature extraction is a crucial part of any automatic annotation system. Good features should be able to model low-level aspects of music audio such as timbre, loudness and pitch, but also higher-level aspects such as melody, phrasing and rhythm. Low-level aspects can be relatively well modelled by features computed over short-time windows. Higher-level aspects, on the other hand, are salient only at larger timescales and require a better representation of time dynamics. In order to obtain a better representation of time dynamics in music audio, we propose to compute general features at different timescales. We build upon the pooled features classifier (PFC) model and principal mel spectral components (PMSCs) features.

Getting Into the Groove with Hierarchical Probabilistic Latent Component Analysis

Qingyuan Kong, Dartmouth College
Andrew Sarroff, Dartmouth College
Spencer Topel, Dartmouth College
Michael A. Casey, Dartmouth College

We present a novel approach for representing the rhythmic content of music in monophonic audio mixtures. Our approach uses a hierarchical extension to probabilistic latent component analysis (H-PLCA) to extract per-track latent amplitude-time envelopes with a universal model of latent timbre bases. We employ entropic priors at each stage of H-PLCA to automatically reduce the rank from initial high-rank distributions. We describe a ”search by groove” system that uses the latent amplitude-time envelopes to represent the rhythmic content of music in mixed audio and we test the performance of our methods in a rhythm retrieval experiment using a collection of 1138 commercial dance music tracks independently marked-up by professional DJs. For retrieval of relevant tracks against the ground truth markup, our results show significant gains in precision when compared with retrieval using only band-pass filter-bank amplitude-time envelopes

Hit Song Science Once Again a Science?

Yizhao Ni, University of Bristol
Raul Santos-Rodriguez, Universidad Carlos III de Madrid
Matt Mcvicar, University of Bristol
Tijl De Bie, University of Bristol

We are interested in the Music Information Retrieval task that aims at predicting whether a given song will be a commercial success prior to its distribution, based on its audio. This is the topic of what is commonly referred to as ’Hit Song Science’ (HSS). The underlying assumption behind HSS is that popular songs are similar with respect to a set of features that make them appealing to a majority of people. These features could then be exploited by learning machines in order to predict whether a song will rise to a high position in the chart. This abstract overviews part of our investigation of the UK top 40 singles chart from the past 50 years. Here, our aim is to distinguish the most popular (peak position top 5) songs from less popular singles (peak position 30 - 40).

Brain-Computer Interfaces for Music Recommendation

Kira Radinsky, Technion
Ashish Kapoorz, Microsoft Research
Avigad Oronz, Microsoft Research
Keren Master, Microsoft Research

We explore the opportunity to harness electroencephalograph (EEG) signals for the purpose of music recommendation. The core idea lies on the hypothesis that cortical signals captured by off-the-shelf electrodes carry enough information about mental state of a listener and can be used to build preference models over musical taste for each individual user. We present a reinforcement learning algorithm that aims to build such models over a period of time and then use it effectively to provide recommendations. Our experiments on real users indicate that the recommendation policy learnt via the brain-computer interface provides better recommendations than commercial services such as Pandora.

Automatic Detection of Ornamentation in Flamenco

Francisco Gomez, Polytechnic University of Madrid
Aggelos Pikrakis, University of Piraeus
Joaquin Mora, University of Seville
Jose Miguel Diaz-Banez, University of Seville
Emilia Gomez, Universitat Pompeu
Fabra Francisco Escobar, University of Seville

Ornamentation is a characteristic feature in the melody of many musical traditions in the world, including flamenco music. There is no universal definition of ornamentation as it is very associated with musical style. In Western classical music, with its long-standing analysis tradition, ornamentation has been studied extensively; consider the ornamentation tables from Baroque to Romanticism, for example. On the contrary, in spite of the always-heard claim that flamenco is very melismatic, scant research have been carried out to define flamenco ornaments. For example, the structural role of flamenco melismas has not been thoroughly described yet, nor its role in style definition. This paper is concerned with the study of ornamentation in flamenco music. As a starting point, we defined a set of ornaments, mainly taken or adapted from classical music. The next step was to select a flamenco corpus -formed by audio recordings- where the selected set of ornaments would be looked up. In order to do this in an automatic way we had to design an appropriate pattern detection algorithm. The algorithm of Smith-Waterman was adapted to our purposes. The obtained results were promising and they constituted a first step toward a systematic study of ornamentation in flamenco music.

This is the Remix: Structural Improvisation using Automated Pattern Discovery

Sean Whalen, Columbia University
James P. Crutchfield, University of California

The remix is pervasive in modern culture. Popularized in contemporary art by early hip-hop, it traces its roots back to both Dadaism and, later, the cut-ups of William S. Burroughs. The basic approach is well understood: Take pre-existing material, then subtract, add, and adjust until something novel and engaging emerges. The remixer becomes a mash-up unto themselves: part sculptor, part curator, part engineer. In popular music, the remix attempts to inject freshness into a familiar song--perhaps to make a slow song danceable or to blend the vocals of a song with the rhythms of a disparate genre in a way that creates new fans. In contrast, the cut-ups of Burroughs or the randomized piano fragments of Stockhausen’s Klavierstucke XI deny such familiarity. And, as a result, they require more of the reader or listener. In all of these, though, the interplay between structure and randomness is the very foundation of our cognitive appreciation: a perceptual information channel from creator to consumer in which both total order and total chaos become uninteresting. Between these two extremes exists an ideal dynamic trade-off where familiarity and surprise keep our attention when we experience the emergence of structural complexity.

Machine Learning in Computational Biology

http://www.fml.tuebingen.mpg.de/nipscompbio

LOCATION
Melia Sierra Nevada: Genil
Saturday, 07:30 -- 10:30 AM & 4:00 -- 8:00 PM

Anna Goldenberg anna.goldenberg@utoronto.ca
Genetics and Genome Biology, SickKids Research Institute

Tomer Hertz thertz@fhcrc.org
Fred Hutchinson Cancer Research Center

Christina Leslie cleslie@cbio.mskcc.org
Sloan-Kettering Institute

Yanjun Qi yanjun@nec-labs.com
Machine Learning Department, NEC Labs America

Gunnar Ratsch Gunnar.Raetsch@tuebingen.mpg.de
Friedrich Miescher Laboratory, Max Planck Society

Jean-Philippe Vert Jean-Philippe.Vert@mines-paristech.fr

Abstract
The field of computational biology has seen dramatic growth The field of computational biology has seen dramatic growth over the past few years, in terms of newly available data, new scientific questions and new challenges for learning and inference. In particular, biological data is often relationally structured and highly diverse, and thus requires combining multiple weak evidence from heterogeneous sources. These sources include sequenced genomes of a variety of organisms, gene expression data from multiple technologies, protein sequence and 3D structural data, protein interaction data, gene ontology and pathway databases, genetic variation data (such as SNPs), high-content phenotypic screening data, and an enormous amount of text data in the biological and medical literature. These new types of scientific and clinical problems require novel supervised and unsupervised learning approaches that can use these growing resources. The goal of this workshop is to present emerging problems and machine learning techniques in computational biology. During the workshop the audience will hear about the progress on new bioinformatics problems and new methodology for established problems. The targeted audience are people with interest in learning and applications to relevant problems from the life sciences.

INVITED SPEAKERS

Unsupervised unwanted variation removal for gene expression data
Laurent Jacob, Johann Gagnon Bartsch and Terence Speed

Please visit the website on the top of this Page for Details

SCHEDULE

7.30-7.35	Introduction and Welcome
7:35-7:55	**Unsupervised unwanted variation removal for gene expression data** Laurent Jacob, Johann Gagnon Bartsch and Terence Speed
7:55-8:15	**Tumor profile purification using infinite mixture topic models** Amit Deshwar, Gerald Quon and Quaid Morris
8:15-8:35	**Simultaneous RNA-seq-based Transcript Inference and Quantification Using Mixed Integer Programming** Jonas Behr, Regina Bohnert, Andre Kahles and Gunnar Raetsch
8:35-8:55	**Bayesian model of transcript differential expression in RNA-seq data with biological variation** Peter Glaus, Antti Honkela and Magnus Rattray
8:55-9:10	Coffee and Poster Setup
9:10-9:40	Poster Spotlights
9:40-11:00	Poster Session
11:00-4:00	Break
4:00-4:45	Invited talk: **High-throughput 3D BioImage Informatics & Its Applications** Hanchuan Peng
4:45-5:05	**Automatic identication of brain-layer specific genes from ISH images** Lior Kirsch, Noa Liscovitch and Gal Chechik
5:05-5:25	**Structure and Parameter Learning for von Mises Graphical Models** Narges Razavian and Christopher Langmead
5:25-5:45	**Hierarchical Inference on Single-Molecule FRET Time Series** Jan Willem Van De Meent, Jonathan E. Bronson, Jingyi Fei, Ruben L. Gonzalez & Chris H. Wiggins
5:45-6:00	Coffee
6:00-6:20	**Estimating Regulatory Networks from Time Course Gene Expression Data via Adaptive Penalization** Sumanta Basu, Ali Shojaie and George Michailidis
6:20-6:40	**TIGRESS: Trustful Inference of Gene REgulation using Stability Selection** Anne-Claire Haury, Fantine Mordelet, Paola Vera-Licona and Jean- Philippe Vert
6:40-7:00	**Learning Transcription Factor to DNA Binding Interactions Using Affinity Regression** Raphael Pelossof and Christina Leslie
7:00-7:30	Panel: **Where should we focus our ML efforts?**
7:30-7:35	Closing Remarks | Organizers

Tumor profile purification using infinite mixture topic models
Amit Deshwar, Gerald Quon and Quaid Morris

Please visit the website on the previous Page for Details

Simultaneous RNA-seq-based Transcript Inference and Quantification Using Mixed Integer Programming
Jonas Behr, Regina Bohnert, Andre Kahles and Gunnar Raetsch

Please visit the website on the previous Page for Details

Bayesian model of transcript differential expression in RNA-seq data with biological variation
Peter Glaus, Antti Honkela and Magnus Rattray

Please visit the website on the previous Page for Details

Invited Talk: **High-throughput 3D BioImage Informatics & Its Applications**
Hanchuan Peng, HHMI Janelia Farm

Bioinformatics is a powerful way to use computational techniques to gain better understanding of biological processes. In many cases it is desirable to analyze genotypes and phenotypes jointly. There have been a huge amount of existing work on genotype data analysis; in recent years, high-throughput phenotype screening that involves systematic analysis of microscopic images (thus called “Bioimage Informatics”) and other types of data has become more and more promising. Here I will discuss several examples on how to (1) build a single cell resolution gene expression map for C. elegans and detect new cell types, and (2) develop a pipeline of tools to understand the structures of a complicated fruit y’s brain. I will also discuss our high-performance image computing platform, Vaa3D http://vaa3d.org, that has been used in several challenging high-throughput bioimage informatics applications.

Automatic identication of brain-layer specific genes from ISH images
Lior Kirsch, Noa Liscovitch and Gal Chechik

Please visit the website on the previous Page for Details

Structure and Parameter Learning for von Mises Graphical Models
Narges Razavian and Christopher Langmead

Please visit the website on the previous Page for Details

Hierarchical Inference on Single-Molecule FRET Time Series
Jan Willem Van De Meent, Jonathan E. Bronson, Jingyi Fei, Ruben L. Gonzalez & Chris H. Wiggins

Please visit the website on the previous Page for Details

Estimating Regulatory Networks from Time Course Gene Expression Data via Adaptive Penalization
Sumanta Basu, Ali Shojaie and George Michailidis

Please visit the website on the previous Page for Details

TIGRESS: Trustful Inference of Gene REgulation using Stability Selection
Anne-Claire Haury, Fantine Mordelet, Paola Vera-Licona and Jean- Philippe Vert

Please visit the website on the previous Page for Details

Learning Transcription Factor to DNA Binding Interactions Using Affinity Regression
Raphael Pelossof and Christina Leslie

Please visit the website on the previous Page for Details

Machine Learning and Interpretation in Neuroimaging

https://sites.google.com/site/mlini2011/

LOCATION
Melia Sol y Nieve: Aqua
Friday & Saturday, 07:30 -- 10:30 AM & 4:00 -- 8:00 PM

Moritz Grosse-Wentrup moritz.grosse-wentrup@tuebingen.mpg.de
Max Planck Institute for Intelligent Systems, Tubingen

Georg Langs georg.langs@meduniwien.ac.at
Medical University of Vienna

Brian Murphy brianmurphy@cmu.edu
Carnegie Mellon University

Irina Rish rish@us.ibm.com
IBM T.J. Watson Research Center

Abstract

In this two-day workshop we will explore perspectives and novel methodology at the interface of Machine Learning, Inference, Neuroimaging and Neuroscience. We aim to bring researchers from machine learning and neuroscience community together, in order to discuss open questions, identify the core points for a number of the controversial issues, and eventually propose approaches to solving those issues.

The workshop will be structured around 4 main topics:

1. Machine learning and pattern recognition methodology
2. Interpretable decoding of higher cognitive states from neural data
3. Causal inference in neuroimaging
4. Linking machine learning, neuroimaging and neuroscience

INVITED SPEAKERS

Multi-subject models of the resting brain

Gäel Varoquaux, INRIA Neurospin

The brain never rests. In the absence of external stimuli, fluctuations in cerebral activity reveal an intrinsic structure that mirrors brain function during cognitive tasks. Well-conditioned probabilistic models of this so-called on-going activity are needed to support neuroscientific hypothesizes. Estimating such models is a high-dimensional unsupervised learning problem limited, at single subject level, by the scarcity of the data available, and, at the population level, by intersubject variability.

First, I will discuss how to extract spatial patterns from on-going activity. The popular ”Independent Component Analysis” can be extended to multi-subject sparse models. These methods single out spatial atoms of brain activity: functional networks or brain regions. With a probabilistic model of inter-subject variability, they learn population-level atlases and with the corresponding subject-specific functional regions.

SCHEDULE

Friday Dec 16

7:30-8:00	Introduction and Overview
8:00-8:40	Invited talk: **Multi-subject models of the resting brain** Gael Varoquaux
8:40-8:55	**Relating brain functional connectivity to anatomical connections: Model Selection** Fani Deligianni, Gael Varoquaux, Bertrand Thirion, Emma Robinson, David Sharp, David Edwards, Daniel Rueckert
8:55-9:30	Coffee Break, Day 1 Posters
9:30-10:00	Invited talk: **Design and application of a biologically inspired feature model, a Bayesian mixture of experts model, and a topic model to functional brain imaging** David Keator
10:00-10:15	**MKL-based sample enrichment and customized outcomes enable smaller AD clinical trials** Chris Hinrichs, Vikas Singh, Sterling Johnson, Maritza Dowling
10:15-10:30	Discussion
10:30-11:30	Day 1 Posters
16:00-16:40	Invited talk: **Combining functional neuroimaging, machine learning and computational modeling to understand auditory perception and cognition** Elia Formisano
16:40-16:55	**The neural basis of rapid categorization: Linking computational models and electrophysiology** Sebastien Crouzet, Maxime Cauchoix, Denis Fize, Thomas Serre
16:55-17:10	**Deformation-Invariant Sparse Coding for Modeling Spatial Variability of Functional Patterns in the Brain** George Chen, Evelina Fedorenko, Nancy Kanwisher, Polina Golland
17:10-17:30	Coffee Break, Day 1 Posters
17:30-17:45	**Inferring Brain Networks through Graphical Models with Hidden Variables** Justin Dauwels, Hang Yu, Xueou Hang, Francois Vialatte, Charles Latchoumane, Andrzej Cichocki
17:45-18:00	**A new approach to neural decoding: acting on the similarities between activation patterns, rather than on the patterns themselves** Rajeev Raizada
18:00-19:00	Panel Session (all invited speakers)

Subsequently, I will consider graphical models of the interactions between brain regions. Such models can be linked to the correlation structure of evoked activity, or the anatomical wiring of the brain. To mitigate the lack of individual data and learn large scale graphical models, introducing a population prior with sparsity-inducing penalizations. The corresponding graphs display a modular community structure reflecting functional networks.

Relating brain functional connectivity to anatomical connections: Model Selection

Fani Deligianni, Imperial College London
Gael Varoquaux, INRIA, Neurospin, Saclay-Ile-de-France
Bertrand Thirion, INRIA, Neurospin, Saclay-Ile-de-France
Emma Robinson, Imperial College London
David Sharp, Imperial College London
David Edwards, Imperial College London
Daniel Rueckert, Imperial College London

We aim to learn across several subjects a mapping from brain anatomical connectivity to functional connectivity. Following (Deligianni et al, 2011), we formulate this problem as estimating a multivariate autoregressive (MAR) model with sparse linear regression. We introduce a model selection framework based on cross-validation. We select the appropriate sparsity of the connectivity matrices and demonstrate that choosing an ordering for the MAR that lends to sparser models is more appropriate than a random. Finally, we suggest randomized LASSO in order to identify relevant anatomo-functional links with better recovery of ground truth.

Design and application of a biologically inspired feature model, a bayesian mixture of experts model, and a topic model to functional brain imaging

David Keator, University of California, Irvine

Identifying functional brain imaging derived features that can be used to detect neurological abnormalities or patterns across multiple images is of primary importance to the medical community. Often clinics will use PET and/or SPECT scans along with other clinical and behavioral variables to diagnose neurologic disorders. Further, the research community routinely uses PET, SPECT, and fMRI functional imaging to infer differences in cognitive processes associated with disease. In this talk I will briefly discuss the design and application of: (1) A biologically inspired feed-forward hierarchical model which emulates visual processing in the striate cortex and its application to signal detection and classification in PET and SPECT imaging; (2) A Bayesian mixture of experts approach for modeling activation patterns across fMRI scans acquired at different sites; (3) Some preliminary results on applying the LDA topic model to count data in PET scans from mutliple disease groups.

MKL-based sample enrichment and customized outcomes enable smaller AD clinical trials

Chris Hinrichs, University of Wisconsin, Madison
Vikas Singh, University of Wisconsin, Madison
Sterling Johnson, University of Wisconsin, Madison
Maritza Dowling, University of Wisconsin, Madison

Recently, the field of neuroimaging analysis has seen a large number of studies which use machine learning methods to make predictions about the progression of Alzheimer's Disease (AD) in mildly demented subjects. Among these, Multi-Kernel Learning (MKL) has emerged as a powerful tool for systematically aggregating diverse data views, and several groups have shown that MKL is uniquely suited to combining different imaging modalities into a single learned model. The next phase of this research is to employ these predictive abilities to design more efficient clinical trials. Two issues can hamper a trial's effectiveness: the presence of non-pathological subjects in a study, and the sensitivity of the chosen outcome measure to the pathology of interest. We offer two approaches for dealing with these issues: trial enrichment, in which MKL-derived predictions are used to screen out subjects unlikely to benefit from a treatment; and custom outcome measures which use an SVM to select a weighted voxel average for use as an outcome measure. We provide preliminary evidence that these two strategies can lead to significant reductions in sample sizes in hypothetical trials, which directly gives reduced costs and higher efficiency in the drug development cycle.

Combining functional neuroimaging, machine learning and computational modeling to understand auditory perception and cognition

Elia Formisano, Maastricht University

In this talk I will present our research combining functional neuroimaging, machine learning and computational modeling to understand the neural basis of human auditory perception and audition. First, I will illustrate a series of studies in which multivariate classification and regression are used to unravel the neural representation of "auditory objects" and their computational properties (e.g. invariance to acoustic variations and robustness to noise). In contrast to strictly hierarchical models of auditory processing, results suggest that primary and early cortical auditory networks encode abstract representations of complex sounds (including voice, speech and music) - beyond their basic acoustic properties. Second, I will illustrate ongoing studies that combine natural sound stimulation, ultra-high field functional MRI (7 Tesla) and computational modeling to examine the functional architecture of the human auditory system. Results show that the topographic representation of frequency preference (tonotopy) and selectivity is not limited to primary regions but extends to higher-order and category-selective auditory regions. The frequency bias in these category-selective regions matches the typical frequency of the preferred sound category. Furthermore, we find that - for ecologically relevant conspecific vocalizations - neuronal and sound spectral profiles are finely matched, which suggests the existence in primary and non primary areas of 'category-matched' neuronal filters. These neuronal filters may be pivotal to the transformation of a sensory (tonotopic) image of a sound into its representation at the level of a semantic category.

The neural basis of rapid categorization: Linking computational models and electrophysiology

Sebastien Crouzet, Brown University
Maxime Cauchoix, Universite de Toulouse
Denis Fize, Universite de Toulouse
Thomas Serre, Brown University

Studies based on rapid visual presentations have demonstrated the incredible speed and accuracy of our visual system for some of the most challenging visual recognition tasks in natural scenes. Several computational models have been proposed that try to account for the underlying visual processes. However, most of the evidence regarding the plausibility of these computational models is based on their ability to match behavioral data and remains indirect. Here, using multivariate pattern analysis techniques, we tested directly the ability of computational models to explain neural activity measured from electrodes in intermediate visual areas (V2 and PIT) of the primate visual cortex while a monkey was actively engaged in a rapid animal vs. non-animal categorization task.

Deformation-Invariant Sparse Coding for Modeling Spatial Variability of Functional Patterns in the Brain

George Chen, MIT
Evelina Fedorenko, MIT
Nancy Kanwisher, MIT
Polina Golland, MIT

For a given cognitive task such as language processing, the location of corresponding functional regions in the brain may vary across subjects relative to anatomy. We present a probabilistic generative model that accounts for such variability as observed in fMRI data. We relate our approach to sparse coding that estimates a basis consisting of functional regions in the brain. Individual fMRI data is represented as a weighted sum of these functional regions that undergo deformations. We demonstrate the proposed method on a language fMRI study. Our method identified activation regions that agree with known literature on language processing and established correspondences among activation regions across subjects.

Inferring Brain Networks through Graphical Models with Hidden Variables

Justin Dauwels, NTU, Singapore
Hang Yu, NTU, Singapore
Xueou Hang, NTU, Singapore
Francois Vialatte, Riken Institute, Tokyo
Charles Latchoumane, Korea Institute of Science & Technology
Andrzej Cichocki, Riken Institute, Tokyo

Inferring the interactions between different brain areas is an important step towards understanding brain activity. Most often, signals can only be measured from some specific brain areas (e.g., cortex in the case of scalp electroencephalograms). However, those signals may be affected by brain areas from which no measurements are available (e.g., deeper areas such as hippocampus). In this paper, the latter are described as hidden variables in a graphical model; such model quantifies the statistical structure in the neural recordings, conditioned on hidden variables, which are inferred in an automated fashion from the data. As an illustration, electroencephalograms (EEG) of Alzheimer's disease patients are considered. It is shown that the number of hidden variables in AD EEG is not significantly different from healthy EEG. However, there are fewer interactions between the brain areas, conditioned on those hidden variables.

A new approach to neural decoding: acting on the similarities between activation patterns, rather than on the patterns themselves

Rajeev Raizada, Cornell University

It might seem almost too obvious to be worth stating that neural decoding should take as its input neural activation patterns. However, we propose a different approach, in which decoding is applied not the neural activation patterns themselves, but instead to the similarity relations between them. This similarity-based decoding does not use a classifier, but instead permutes condition-labels in order to produce a match between two similarity spaces. We review work demonstrating two advantages of this approach. First, it is able to reveal what makes different people's neural representations alike: people's "neural fingerprint" activation patterns are subject-specific and idiosyncratic, but their neural similarity-spaces turn out to have a shared structure, allowing highly accurate across-subject decoding. Second, a similarity-based approach can match neural similarities directly onto semantic similarities, thereby providing a new and successful method of decoding the meanings of untrained words. Finally, we discuss striking parallels between our proposed decoding approach and work in machine-learning and the philosophy of mind.

SCHEDULE

Saturday Dec 17

7:30-7:35	Opening remarks
7:35-8:20	Invited talk: **To Be Announced** Richard Scheines
8:20-9:05	Invited talk: **Inter-areal synchronization and attention** Pascal Fries
9:05-9:30	Coffee Break, Day 2 Posters
9:30-10:30	Open problems session
10:30-11:30	Day 2 Posters
16:00-16:40	Invited talk: **Alignment-Free Exploratory Analysis of fMRI Data** Polina Golland
16:40-17:20	Invited talk: **Learning and Discovery of Clinically Useful Biomarkers in Neuroimaging** Daniel Rueckert
17:20-17:40	Coffee Break, Day 2 Posters
17:40-18:20	**Aligning feature spaces for multi-subject multivariate pattern analysis of fMRI** James V. Haxby
18:20-19:10	Panel Session (all invited speakers)

SATURDAY WORKSHOP ABSTRACT

TBA
Richard Scheines, CMU

See the website on the top of page 48 for details.

Inter-areal synchronization and attention
Pascal Fries, Ernst Strűngmann Institute

Invited Talk : Attention is most likely implemented by modulations in the effective connectivity among brain areas. We have proposed that effective connectivity depends on rhythmic synchronization. We have therefore assessed neuronal activity with 252 electrodes distributed across one hemisphere between primary visual cortex and prefrontal cortex. We find that attention is subserved by strong and specific enhancements of interareal synchronization. The inter-areal influence is often directed, typically bottom-up in the gamma-band and top-down in the beta-band. These results suggest that inter-areal synchronization subserves effective inter-areal interactions.

Alignment-Free Exploratory Analysis of fMRI Data
Polina Golland, CSAIL, MIT

We present an exploratory method for simultaneous parcellation of multisubject fMRI data into functionally coherent areas. Our motivation comes from visual fMRI studies with increasingly large number of image categories. The method is based on a solely functional representation of the fMRI data and a hierarchical probabilistic model that accounts for both inter-subject and intra-subject forms of variability in fMRI responses. The resulting algorithm finds a functional parcellation of the individual brains along with a set of population-level clusters. The model eliminates the need for spatial normalization while still enabling us to fuse data from multiple subjects. Joint work with Danial Lashkari, Ed Vul and Nancy Kanwisher.

Learning and Discovery of Clinically Useful Biomarkers in Neuroimaging
Daniel Rueckert, Imperial College London

Three-dimensional (3D) and four-dimensional (4D) imaging plays an increasingly important role in neuroimaging. In particular there is now wide range of diverse neuroimaging techniques available that go beyond anatomical imaging. This includes diffusion tensor MR imaging (DT-MRI) as well as functional MR imaging (fMRI). In addition, non-imaging information such as genetics and metabolomics offer complementary information about the patient's health. At the same time, advances in machine learning have transformed many of the classical problems in computer vision into machine learning problems. This talk will focus on machine learning techniques for the discovery and quantification of clinically useful information from medical images. To illustrate this I will show several examples such as the segmentation of neuro-anatomical structures, the discovery of biomarkers for neurodegenerative diseases such as Alzheimers and the quantification of temporal changes such as growth in the developing brain. I will also discuss some of the challenges for machine learning in integrating imaging and non-imaging information in this context.

Aligning feature spaces for multi-subject multivariate pattern analysis of fMRI
James V. Haxby, Dartmouth College

See the website on the top of page 48 for details.

Domain Adaptation Workshop: Theory and Application

http://eecs.berkeley.edu/~arostami/da_workshop

LOCATION
Melia Sierra Nevada: Monachil
Saturday, 07:30 -- 10:30 AM & 4:00 -- 8:00 PM

John Blitzer blitzer@google.com
Corinna Cortes corinna@google.com
Afshin Rostamizadeh rostami@google.com
Google Research

Abstract
The main theme of this workshop is the theoretical, algorithmic, and empirical analysis of such cases where there is a mismatch between the training and test distributions. This includes the crucial scenario of domain adaptation where the training examples are drawn from a source domain distinct from the target domain from which the test examples are extracted, or the more general scenario of multiple source adaptation where training instances may have been collected from multiple source domains, all distinct from the target. The topic of our workshop also covers other important problems such that of sample bias correction and has tight connections with other problems such as active learning where the active distribution corresponding to the learner's labeling request differs from the target distribution. Many other intermediate problems and scenarios appear in practice, which will be all covered by this workshop.

These problems are all critical and appear in almost all real-world applications of machine learning. Ignoring them can lead to dramatically poor results. Some straightforward existing solutions based on importance weighting are not always successful. Which algorithms should be used for domain adaptation? Under what theoretical conditions will they be successful? How do these algorithms scale to large domain adaptation problems? These are some of the questions that the workshop aims to address.

INVITED SPEAKERS

On the utility of unlabeled samples in domain adaptation

Shai Ben-David, University of Waterloo

In many domain adaptation applications, on top of a sample of labeled points generated by the training tasks, the learner can also access unlabeled samples generated by the target distribution. The focus of this talk is to investigate when can such unlabeled samples be (provably) beneficial to the leaner.

We show that depending on the type of prior knowledge available to the leaner, there are setups in which unlabeled target-generated samples can make a big difference in the required size of labeled training samples, while in other scenarios such unlabeled samples do not improve the learning rate.

SCHEDULE

7.30-7.45	Introduction
7.45-8.25	**On the utility of unlabeled samples in domain adaptation** Shai Ben-David
8.25-8.40	**Transportability and the Bias-Variance Tradeoff** Karthika Mohan
8.40-8.50	Coffee Break
8.50-9.05	**Training Structured Prediction Models with Extrinsic Loss Functions** Slav Petrov
9.05-9.45	**Adaptation without Retraining** Dan Roth
9.45-10.30	Poster Session / Discussion
4.00-4.20	Discussion / Review of Morning Session
4.20-5.00	**Discrepancy and Domain Adaptation** Mehryar Mohri
5.00-5.15	**History Dependent Domain Adaptation** Nathan Ratliff
5.15-5.25	Break
5.25-5.40	**Domain Adaptation with Multiple Latent Domains** Kate Saenko
5.40-5.55	**Domain Adaptation: Overfitting and Small Sample Statistics** Ruslan Salakhutdinov
5.55-6.10	**Cool world: domain adaptation of virtual and real worlds for human detection using active learning** David Vázquez
6.10-7.00	Poster Session / Discussion

Transportability and the Bias-Variance Tradeoff

Karthika Mohan, University of California, Los Angeles
Jennifer Wortman Vaughan, University of California, Los Angeles
Judea Pearl, University of California, Los Angeles

Transportability is a recently proposed framework that examines whether or not a particular statistical or causal relation can be transported from a source domain to a related target domain given some knowledge of the differences between the domains, usually by appealing to graphical criteria. Transportability allows us to exploit the structure of the source and target domains, but is rigid in the sense that each relation is said to be either fully transportable or not. We propose a relaxation of transportability, and provide examples illustrating how this relaxation can be used to determine whether or not to conduct a new study or collect new data. Finally, we briefly mention ongoing research formalizing and quantifying the bias-variance tradeoff that arises when determining whether to mix source and target data under various graphical criteria.

Training Structured Prediction Models with Extrinsic Loss Functions

Keith Hall, Google, New York
Ryan McDonald, Google, New York
Slav Petrov, Google, New York

We present an online learning algorithm for training structured prediction models with extrinsic loss functions. This allows us to extend a standard supervised learning objective with additional loss-functions, either based on intrinsic or task-specific extrinsic measures of quality. We present experiments with sequence models on part-of-speech tagging and named entity recognition tasks, and with syntactic parsers on dependency parsing and machine translation reordering tasks.

Adaptation without Retraining

Dan Roth, University of Illinois at Urbana-Champaign

Natural language models trained on labeled data from one domain do not perform well on other domains. Most adaptation algorithms proposed in the literature train a new model for the target domain using a mix of labeled and unlabeled data. We discuss some limitations of existing general purpose adaptation algorithms that are due to the interaction between differences in base feature statistics and task differences and illustrate how this should be taken into account jointly.

With these insights we propose a new approach to adaptation that avoids the need for retraining models. Instead, at evaluation time, we perturb the given instance to be more similar to instances the model can handle well, or perturb the model outcomes to fit our expectation of the target domain better, given some prior knowledge on the task and the target domain. We provide experimental evidence in a range of natural language processing, including semantic role labeling and English as a Second Language (ESL) text correction tasks.

Discrepancy and Domain Adaptation

Mehryar Mohri, Courant Institute of Mathematical Sciences

History Dependent Domain Adaptation

Allen Lavoie, Rensselaer Polytechnic Matthew
Eric Otey, Google, Pittsburgh
Nathan Ratliff, Google, Pittsburgh
D. Sculley, Google, Pittsburgh

We study a novel variant of the domain adaptation problem, in which the loss function on test data changes due to dependencies on prior predictions. One important instance of this problem area occurs in settings where it is more costly to make a new error than to repeat a previous error. We propose several methods for learning effectively in this setting, and test them empirically on the real-world tasks of malicious URL classification and adversarial advertisement detection.

Domain Adaptation with Multiple Latent Domains

Judy Hoffman, UC Berkeley
Kate Saenko, UC Berkeley
Brian Kulis, UC Berkeley
Trevor Darrell, UC Berkeley

Domain adaptation is important for practical applications of supervised learning, as the distribution of inputs can differ significantly between available sources of training data and the test data in a particular target domain. Many domain adaptation methods have been proposed, yet very few of them deal with the case of more than one training domain; methods that do incorporate multiple domains assume that the separation into domains is known a priori, which is not always the case in practice. In this paper, we introduce a method for multi-domain adaptation with unknown domain labels, based on learning nonlinear cross-domain transforms, and apply it to image classification. Our key contribution is a novel version of constrained clustering; unlike many existing constrained clustering algorithms, ours can be shown to provably converge locally while satisfying all constraints. We present experiments on a commonly available image dataset.

Domain Adaptation: Overfitting and Small Sample Statistics

Dean Foster, University of Pennsylvania
Sham Kakade, Microsoft Research
Ruslan Salakhutdinov, University of Toronto

We study the prevalent problem when a test distribution differs from the training distribution. We consider a setting where our training set consists of a small number of sample domains, but where we have many samples in each domain. Our goal is to generalize to a new domain. For example, we may want to learn a similarity function using only certain classes of objects, but we desire that this similarity function be applicable to object classes not present in our training sample (e.g. we might seek to learn that “dogs are similar to dogs” even though images of dogs were absent from our training set). Our theoretical analysis shows that we can select many more features than domains while avoiding overfitting by utilizing data-dependent variance properties. We present a greedy feature selection algorithm based on using T-statistics. Our experiments validate this theory showing that our T-statistic based greedy feature selection is more robust at avoiding overfitting than the classical greedy procedure.

Cool world: domain adaptation of virtual and real worlds for human detection using active learning

David Vázquez, Universitat Autonòma de Barcelona
Antonio M. López, Universitat Autonòma de Barcelona
Daniel Ponsa, Universitat Autonòma de Barcelona
Javier Marin, Universitat Autonòma de Barcelona

Image based human detection is of paramount interest for different applications. The most promising human detectors rely on discriminatively learnt classifiers, i.e., trained with labeled samples. However, labelling is a manual intensive task, especially in cases like human detection where it is necessary to provide at least bounding boxes framing the humans for training. To overcome such problem, in Marin et al. we have proposed the use of a virtual world where the labels of the different objects are obtained automatically. This means that the human models (classifiers) are learnt using the appearance of realistic computer graphics. Later, these models are used for human detection in images of the real world. The results of this technique are surprisingly good. However, these are not always as good as the classical approach of training and testing with data coming from the same camera and the same type of scenario. Accordingly, in Vazquez et al. we cast the problem as one of supervised domain adaptation. In doing so, we assume that a small amount of manually labeled samples from real-world images is required. To collect these labeled samples we use an active learning technique. Thus, ultimately our human model is learnt by the combination of virtual- and real-world labeled samples which, to the best of our knowledge, was not done before. Here, we term such combined space cool world. In this extended abstract we summarize our proposal, and include quantitative results from Vazquez et al. showing its validity.

Challenges in Learning Hierarchical Models: Transfer Learning and Optimization

https://sites.google.com/site/nips2011workshop/schedule

LOCATION
Montebajo: Library
Saturday, 07:30 -- 10:30 AM & 4:00 -- 8:00 PM

Quoc V. Le Stanford University	quocle@cs.stanford.edu
Marc'Aurelio Ranzato Google	ranzato@google.com
Ruslan Salakhutdinov University of Toronto	rsalakhu@utstat.toronto.edu
Andrew Y. Ng Stanford University	ang@cs.stanford.edu
Josh Tenenbaum MIT	jbt@mit.edu

Abstract

The ability to learn abstract representations that support transfer to novel but related tasks lies at the core of solving many AI related tasks, including visual object recognition, information retrieval, speech perception, and language understanding. Hierarchical models that support inferences at multiple levels have been developed and argued as among the most promising candidates for achieving such goal. An important property of these models is that they can extract complex statistical dependencies from high-dimensional sensory input and efficiently learn latent variables by re-using and combining intermediate concepts, allowing these models to generalize well across a wide variety of tasks.

In the past few years, researchers across many different communities, from applied statistics to engineering, computer science and neuroscience, have proposed several hierarchical models that are capable of extracting useful, high-level structured representations. The learned representations have been shown to give promising results for solving a multitude of novel learning tasks. A few notable examples of such models include Deep Belief Networks, Deep Boltzmann Machines, sparse coding-based methods, nonparametric and parametric hierarchical Bayesian models.

Despite recent successes, many existing hierarchical models are still far from being able to represent, identify and learn the wide variety of possible patterns and structure in real-world data. Existing models cannot cope with new tasks for which they have not been specifically trained. Even when applied to related tasks, trained systems often display unstable behavior. Furthermore, massive volumes of training data (e.g., data transferred between tasks) and high-dimensional input spaces pose challenging questions on how to effectively train the deep hierarchical models. The recent availability of large scale datasets (like ImageNet for visual object recognition or Wall Street Journal for large vocabulary speech recognition), the continuous advances in optimization methods, and the availability of cluster computing have drastically changed the working scenario, calling for a re-assessment of the strengths and weaknesses of many existing optimization strategies.

SCHEDULE

7:30-8:00	Introduction & Opening remarks
8:00-8:20	**Hierarchical learning in human and machines** Josh Tenenbaum - MIT
8:20-8:50	**Learning Hierarchical-Deep Models** Ruslan Salakhutdinov - University of Toronto
8:50-9:20	Coffee Break
9:20-9:50	**Scalable nonparametric Bayesian hierarchical Representations** Dilan Gorur - UC Irvine
9:50-10:10	Spotlights
10:10-10:30	Open discussion (followed by poster session)
16:00-16:10	**Transfer Learning and Optimization Challenges** Organizers
16:10-16:35	**Talk by Winner of Transfer Learning Challenge: Spike and Slab Sparse Coding for Unsupervised Feature Discovery** Ian Goodfellow - University of Montreal
16:35-17:00	**Talk by Winner of Optimization Challenge: Averaged Stochastic Gradient Descent for Autoencoders** Sixin Zhang - New York University
17:00-17:40	**Why equivariance is better than premature invariance?** Geoff Hinton - University of Toronto
17:40-18:10	Coffee Break
18:10-18:40	**Learning scalable visual recognition models from images and videos** Kai Yu - NEC
18:40-19:30	Open Discussion
19:10-19:30	Poster Session

The aim of this workshop is to bring together researchers working on such hierarchical models to discuss two important challenges: the ability to perform transfer learning and the best strategies to optimize these systems on large scale problems. These problems are "large" in terms of input dimensionality (in the order of millions), number of training samples (in the order of 100 millions or more) and number of categories (in the order of several tens of thousands).

INVITED SPEAKERS

Hierarchical learning in human and machines
Josh Tenenbaum - MIT
Please visit the website on the previous Page for Details

Learning Hierarchical-Deep Models
Ruslan Salakhutdinov, University of Toronto

The ability to learn abstract representations that support transfer to novel but related tasks, lies at the core of many problems in AI, including computer vision, natural language processing, and machine learning. In this talk I will briefly survey recent developments on learning deep models, including Deep Boltzmann Machines. I will then introduce a new compositional learning architecture that integrates deep learning models with nonparametric hierarchical Bayesian models. Specifically, I will show how we can learn a hierarchical Dirichlet process (HDP) prior over the activities of the top-level features in a Deep Boltzmann Machine (DBM), coming to represent both a layered hierarchy of increasingly abstract features, and a tree-structured hierarchy of classes. I will demonstrate that this model is able to learn new concepts from very few examples. Time permits, I will also introduce a new approach to learning these hierarchical models based on adaptive MCMC.

Scalable nonparametric Bayesian hierarchical Representations
Dilan Gorur, UC Irvine

Real phenomena is of essentially unbounded complexity and the immense amount of data available today calls for models with increased capacity. The nonparametric Bayesian framework is a principled way to address this challenge by providing models of unbounded capacity that can flexibly adjust to available data. I will talk about a class of nonparametric Bayesian models for hierarchical representations, called Kingman's coalescent models. The coalescent is a powerful model widely used in population genetics, and recent developments have begun to employ the approach as a model for hierarchical representations in machine learning. I will describe scalable greedy and sequential Monte Carlo algorithms for the coalescent that will make possible deployment of this model on a scale of data that was previously out of reach.

Why equivariance is better than premature invariance?
Geoff Hinton Hinton, University of Toronto

Most people who use deep neural networks for object recognition aim to learn a hierarchy of features that progressively lose information about pose. Unfortunately, this prevents the networks from modeling the precise spatial relationships between high-level parts and it means that precise geometric relationships cannot be used for recognition. A better approach is to learn features that explicitly encode the pose (and other properties such as brightness, contrast and deformation) in addition to the type of the feature. This can be done using "capsules" that contain quite a few neurons which perform a lot of internal computation and then encapsulate the results into a compact vector of activities that is easy to communicate. The pose information in this vector can be expressed in a way that allows a very economical and totally invariant representation of the spatial relationships between parts. This allows the recognizer to generalize to very different poses, so by keeping an explicit, equivariant representation of pose we get much better invariance to pose. This is exactly how computer graphics deals with pose so effectively in the inverse direction.

Learning scalable visual recognition models from images and videos
Kai Yu, NEC

Today we are facing the challenges of dealing with huge amount of every-day visual sensory data, in order to learn models for both extracting meaningful features and making accurate recognitions. The models have to be scalable in the sense that it can automatically explore the spatial/temporal structure of data without requiring too much human labeling, and meanwhile be computationally efficient. In this talk I will introduce two pieces of our recent work related to the goal: one is about learning hierarchical models from natural videos, which explores temporal consistency to obtain invariant features for improving image classification; in the second part, we will share some of our experiences in learning large-scale models for image classification and metric learning using average stochastic gradient descent (ASGD), which has demonstrated superior performances in tasks like ImageNet challenge.

ACCEPTED POSTERS

On Imbalanced Data, Hierarchical Models and Transfer Learning
Srinath Ravindran and Dennis Bahler
North Carolina State University

Multi-Label Boosting via Hypothesis Reuse
Sheng-Jun Huang, Yang Yu and Zhi-Hua Zhou, Nanjing University

Structured Latent Factor Analysis
Yunlong He, Koray Kavukcuoglu, Yanjun Qi and Haesun Park, Georgia Tech and NEC

Correlations and Anticorrelations in LDA Inference
Alexandre Passos, Hanna Wallach and Andrew McCallum, University of Massachusetts Amherst

Multi-category and Taxonomy Learning : A Regularization Approach
Youssef Mroueh, Tomaso Poggio and Lorenzo Rosasco, MIT

Spike-and-Slab Sparse Coding for Unsupervised Feature Discovery
Ian J. Goodfellow, Aaron Courville and Yoshua Bengio, University of Montreal

Leveraging Different Transfer Learning Assumptions: Shared Features, Hierarchical and Semi-Supervised
Bernardino Romera-Paredes, Massimiliano Pontil & Nadia Berthouze, University College London

Cosmology meets Machine Learning

http://cmml-nips2011.wikispaces.com

LOCATION
Melia Sierra Neveda: Monachil
Friday, December 16th, 07:30 -- 10:30 AM & 4:00 -- 8:00 PM

Sarah Bridle sarah@sarahbridle.net
Mark Girolami girolami@stats.ucl.ac.uk
Michael Hirsch michael.hirsch@ucl.ac.uk
University College London

Stefan Harmeling stefan.harmeling@tuebingen.mpg.de
Berhnard Schölkopf bs@tuebingen.mpg.de
MPI for Intelligent Systems Tubingen

Phil Marshall dr.phil.marshall@gmail.com
University of Oxford

Abstract

Cosmology aims at the understanding of the universe and its evolution through scientific observation and experiment and hence addresses one of the most profound questions of human mankind. With the establishment of robotic telescopes and wide sky surveys cosmology already now faces the challenge of evaluating vast amount of data. Multiple projects will image large fractions of the sky in the next decade, for example the Dark Energy Survey will culminate in a catalogue of 300 million objects extracted from petabytes of observational data. The importance of automatic data evaluation and analysis tools for the success of these surveys is undisputed.

Many problems in modern cosmological data analysis are tightly related to fundamental problems in machine learning, such as classifying stars and galaxies and cluster finding of dense galaxy populations. Other typical problems include data reduction, probability density estimation, how to deal with missing data and how to combine data from different surveys. An increasing part of modern cosmology aims at the development of new statistical data analysis tools and the study of their behavior and systematics often not aware of recent developments in machine learning and computational statistics.

Therefore, the objectives of this workshop are two-fold:

1. The workshop aims to bring together experts from the Machine Learning and Computational Statistics community with experts in the field of cosmology to promote, discuss and explore the use of machine learning techniques in data analysis problems in cosmology and to advance the state of the art.

2. By presenting current approaches, their possible limitations, and open data analysis problems in cosmology to the NIPS community, this workshop aims to encourage scientific exchange and to foster collaborations among the workshop participants.

The workshop is proposed as a one-day workshop organized jointly by experts in the field of empirical inference and cosmology. The target group of participants are researchers working in the field of cosmological data analysis as well as researchers from the whole NIPS community sharing the interest in real-world applications in a fascinating, fast-progressing field of fundamental research. Due to the mixed participation of computer scientists and cosmologists the invited speakers will be asked to give talks with tutorial character and make the covered material accessible for both computer scientists and cosmologists.

SCHEDULE

07:30-07:40 Welcome: organizers

07:40-08:20 Invited Talk: **Data Analysis Problems in Cosmology**
Robert Lupton

08:20-08:50 Spotlight
Very short talks by poster contributors

08:50-09:20 Coffee Break

09:20-10:00 Invited Talk: **Theories of Everything**
David Hogg

10:00-10:30 Spotlight
Very short talks by poster contributors

10:30-16:00 Break

16:00-16:40 Invited Talk: **Challenges in Cosmic Shear**
Alex Refregier

16:40-17:20 Invited Talk: **Astronomical Image Analysis**
Jean-Luc Starck

17:20-18:00 Coffee Break

18:00-19:00 Panel Discussion: **Opportunities for cosmology to meet machine learning**

19:00-19:15 Closing Remarks: organizers

19:15-20:00 General Discussion: **Opportunities for cosmologists to meet machine learners**

07:30-20:00 Posters will be displayed, in coffee area.

INVITED SPEAKERS

Data Analysis Problems in Cosmology
Robert Lupton, Princeton University

See the website on the top of the previous page for details.

Theories of Everything
David Hogg, New York University

See the website on the top of the previous page for details.

Challenges in Cosmic Shear
Alexandre Refregier, ETH Zurich

Recent observations have shown that the Universe is dominated by two mysterious components, Dark Matter and Dark Energy. Their nature pose some of the most pressing questions in fundamental physics today. Weak gravitational lensing, or 'cosmic' shear', is a powerful technique to probe these dark components. We will first review the principles of cosmic shear and its current observational status. We will describe the future surveys which will be available for cosmic shear studies. We will then highlight key challenges in data analysis which need to be met for the potential of these future surveys to be fully realized.

Astromical Image Analysis
Jean-Luc Starck, CEA Saclay, Paris

See the website on the top of the previous page for details.

Deep Learning and Unsupervised Feature Learning

http://deeplearningworkshopnips2011.wordpress.com

LOCATION
Telecabina: Movie Theater
Friday, December 16th, 07:30 -- 10:30 AM & 4:00 -- 8:00 PM

Adam Coates — acoates@cs.stanford.edu
Stanford University

Nicolas Le Roux — nicolas@le-roux.name
INRIA

Yoshua Bengio — bengioy@iro.umontreal.ca
University of Montreal

Yann LeCun — yann@cs.nyu.edu
New York University

Andrew Ng — ang@cs.stanford.edu
Stanford University

Abstract

In recent years, there has been a lot of interest in algorithms that learn feature representations from unlabeled data. Deep learning algorithms such as deep belief networks, sparse coding-based methods, convolutional networks, ICA methods, and deep Boltzmann machines have shown promise and have already been successfully applied to a variety of tasks in computer vision, audio processing, natural language processing, information retrieval, and robotics. In this workshop, we will bring together researchers who are interested in deep learning and unsupervised feature learning, review the recent technical progress, discuss the challenges, and identify promising future research directions. Through invited talks, panel discussions and presentations by the participants, this workshop attempts to address some of the more controversial topics in deep learning today, such as whether hierarchical systems are more powerful, and what principles should guide the design of objective functions used to train these models. Panel discussions will be led by the members of the organizing committee as well as by prominent representatives of the community. The goal of this workshop is two-fold. First, we want to identify the next big challenges and to propose research directions for the deep learning community. Second, we want to bridge the gap between researchers working on different (but related) fields, to leverage their expertise, and to encourage the exchange of ideas with all the other members of the NIPS community.

SCHEDULE

7.30-8.30	**Tutorial on deep learning and unsupervised feature learning** Workshop organizers
8.30-9.10	Invited Talk: **Classification with Stable Invariants** Stephane Mallat
9.10-9.30	Break
9.30-9.50	Poster Presentation Spotlights
9.50-10.30	Poster Session #1 and group discussions
4.00-4.40	Invited Talk: **Structured sparsity and convex optimization** Francis Bach
4.40-5.05	Panel Discussion #1 Francis Bach, Samy Bengio, Yann LeCun, Andrew Ng
5.05-5.25	Break
5.25-5.43	Contributed Talk: **Online Incremental Feature Learning with Denoising Autoencoders** Guanyu Zhou, Kihyuk Sohn, Honglak Lee
5.43-6.00	Contributed Talk: **Improved Preconditioner in Hessian Free Optimization** Olivier Chapelle, Dumitru Erhan
6.00-6.25	Panel Discussion #2 Yoshua Bengio, Nando de Freitas, Stephane Mallat
6.25-7.00	Poster Session #2 and group discussions

INVITED SPEAKERS

Classification with Stable Invariants

Stéphane Mallat, IHES, Ecole Polytechnique, Paris
Joan Bruna, IHES, Ecole Polytechnique, Paris

Classification often requires to reduce variability with invariant representations, which are stable to deformations, and retain enough information for discrimination. Deep convolution networks provide architectures to construct such representations. With adapted wavelet filters and a modulus pooling non-linearity, a deep convolution network is shown to compute stable invariants relatively to a chosen group of transformations. It may correspond to translations, rotations, or a more complex group learned from data. Renormalizing this scattering transform leads to a representation similar to a Fourier transform, but stable to deformations as opposed to Fourier. Enough information is preserved to recover signal approximations from their scattering representation. Image and audio classification examples are shown with linear classifiers.

Structured sparsity and convex optimization

Francis Bach, INRIA

The concept of parsimony is central in many scientific domains. In the context of statistics, signal processing or machine learning, it takes the form of variable or feature selection problems, and is commonly used in two situations: First, to make the model or the prediction more interpretable or cheaper to use, i.e., even if the underlying problem does not admit sparse solutions, one looks for the best sparse approximation. Second, sparsity can also be used given prior knowledge that the model should be sparse. In these two situations, reducing parsimony to finding models with low cardinality turns out to be limiting, and structured parsimony has emerged as a fruitful practical extension, with applications to image processing, text processing or bioinformatics. In this talk, I will review recent results on structured sparsity, as it applies to machine learning and signal processing. (Joint work with R. Jenatton, J. Mairal and G. Obozinski)

Choice Models and Preference Learning

https://sites.google.com/site/cmplnips11/

LOCATION
Montebajo: Room I
Saturday, 07:30 -- 10:30 AM & 4:00 -- 8:00 PM

Jean-Marc Andreoli jean-marc.andreoli@xrce.xerox.com
Cedric Archambeau cedric.archambeau@xrce.xerox.com
Guillaume Bouchard guillaume.bouchard@xrce.xerox.com
Shengbo Guo shengbo. guo@xrce.xerox.com
Onno Zoeter onno. zoeter@xrce.xerox.com
Xerox Research Centre Europe

Kristian Kersting kristian.kersting@iais.fraunhofer.de
Fraunhofer IAIS - University of Bonn

Scott Sanner scott.sanner@nicta.com.au
NICTA

Martin Szummer szummer@microsoft.com
Microsoft Research Cambridge

Paolo Viappiani paolo.viappiani@gmail.com
Aalborg University

Abstract
Preference learning has been studied for several decades and has drawn increasing attention in recent years due to its importance in diverse applications such as web search, ad serving, information retrieval, recommender systems, electronic commerce, and many others. In all of these applications, we observe (often discrete) choices that reflect preferences among several entities, such as documents, webpages, products, songs etc. Since the observation then is partial, or censored, the goal is to learn the complete preference model, e.g. to reconstruct a general ordering function from observed preferences in pairs.

Traditionally, preference learning has been studied independently in several research areas, such as machine learning, data and web mining, artificial intelligence, recommendation systems, and psychology among others, with a high diversity of application domains such as social networks, information retrieval, web search, medicine, biology, etc. However, contributions developed in one application domain can, and should, impact other domains. One goal of this workshop is to foster this type of interdisciplinary exchange, by encouraging abstraction of the underlying problem (and solution) characteristics during presentation and discussion. In particular, the workshop is motivated by the two following lines of research:

SCHEDULE

7.30-7:45	Opening
7:45-8:30	Invited talk: **Online Learning with Implicit User Preferences** Thorsten Joachims
8:30-9:00	Contributed talk: **Exact Bayesian Pairwise Preference Learning and Inference on the Uniform Convex Polytope** Scott Sanner and Ehsan Abbasnejad
9:00-9:15	Coffee break
9:15-10:00	Invited talk: **Towards Preference-Based Reinforcement Learning** Johannes Fuernkranz
10:00-10:30	3-minute pitch for posters Authors with poster papers
10:30-15:30	Coffee break, poster session, lunch, skiing break
15:30-16:15	Invited talk: **Collaborative Learning of Preferences for Recommending Games and Media** Thore Graepel
16:15-16:45	Contributed talk: **Label Ranking with Abstention: Predicting Partial Orders by Thresholding Probability Distributions** Weiwei Cheng and Eyke Huellermeier
16:45-17:00	Coffee break
17:00-17:45	Invited talk by Zoubin Ghahramani
17:45-18:15	Contributed talk: **Approximate Sorting of Preference Data** Ludwig M. Busse, Morteza Haghir Chehreghani and Joachim M. Buhmann
18:15-18:20	Break
18:20-19:05	Invited talk by Craig Boutilier
19:05-19:30	Discussion & Open research problems

1. Large scale preference learning with sparse data: There has been a great interest and take-up of machine learning techniques for preference learning in learning to rank, information retrieval and recommender systems, as supported by the large proportion of preference learning based literature in the widely regarded conferences such as SIGIR, WSDM, WWW, CIKM. Different paradigms of machine learning have been further developed and applied to these challenging problems, particularly when there is a large number of users and items but only a small set of user preferences are provided.

2. Personalization in social networks: recent wide acceptance of social networks has brought great opportunities for services in different domains, thanks to Facebook, Linkin, Douban, Twitter, etc. It is important for these service providers to offer personalized service (e.g., personalization of Twitter recommendations). Social information can improve the inference for user preferences. However, it is still challenging to infer user preferences based on social relationship.

INVITED SPEAKERS

Invited talk: Online Learning with Implicit User Preferences

Thorsten Joachims, Cornell University

Many systems, ranging from search engines to smart homes, aim to continually improve the utility they are providing to their users. While clearly a machine learning problem, it is less clear what the interface between user and learning algorithm should look like. Focusing on learning problems that arise in recommendation and search, this talk explores how the interactions between the user and the system can be modeled as an online learning process. In particular, the talk investigates several techniques for eliciting implicit feedback, evaluates their reliability through user studies, and then proposes online learning models and methods that can make use of such feedback. A key finding is that implicit user feedback comes in the form of preferences, and that our online learning methods provide bounded regret for (approximately) rational users.

Exact Bayesian Pairwise Preference Learning and Inference on the Uniform Convex Polytope

Scott Sanner, NICTA
Ehsan Abbasnejad, NICTA{ANU

In Bayesian approaches to utility learning from preferences, the objective is to infer a posterior belief distribution over an agent's utility function based on previously observed agent preferences. From this, one can then estimate quantities such as the expected utility of a decision or the probability of an unobserved preference, which can then be used to make or suggest future decisions on behalf of the agent. However, there remains an open question as to how one can represent beliefs over agent utilities, perform Bayesian updating based on observed agent pairwise preferences, and make inferences with this posterior distribution in an exact, closed-form. In this paper, we build on Bayesian pairwise preference learning models under the assumptions of linearly additive multi-attribute utility functions and a bounded uniform utility prior. These assumptions lead to a posterior form that is a uniform distribution over a convex polytope for which we then demonstrate how to perform exact, closed-form inference w.r.t. this posterior, i.e., without resorting to sampling or other approximation methods.

Invited talk: Towards Preference-Based Reinforcement Learning

Johannes Fuernkranz, TU Darmstadt

Preference Learning is a recent learning setting, which may be viewed as a generalization of several conventional problem settings, such as classification, multi-label classification, ordinal classification, or label ranking. In the first part of this talk, I will give a brief introduction into this area, and brie y recapitulate some of our work on learning by pairwise comparisons. In the second part of the talk, I will present a framework for preference-based reinforcement learning, where the goal is to replace the quantitative reinforcement signal in a conventional RL setting with a qualitative reward signal in the form of preferences over trajectories. I will motivate this approach and show first results in simple domains.

Invited talk: Collaborative Learning of Preferences for Recommending Games and Media

Thore Graepel, Microsoft Research Cambridge

The talk is motivated by our recent work on a recommender system for games, videos, and music on Microsoft's Xbox Live Marketplace with over 35M users. I will discuss the challenges associated with such a task including the type of data available, the nature of the user feedback data, implicit versus explicit, and the scale of the problem. I will then describe a probabilistic graphical model that combines the prediction of pairwise and list-wise preferences with ideas from matrix factorisation and content-based recommender systems to meet some of these challenges. The new model combines ideas from two other models, TrueSkill and Matchbox, which will be reviewed. TrueSkill is a model for estimating players' skills based on outcome rankings in online games on Xbox Live, and Matchbox is a Bayesian recommender system based on mapping user/item features into a common trait space. This is joint work with Tim Salimans and Ulrich Paquet. Contributors to TrueSkill include Ralf Herbrich and Tom Minka, contributors to Matchbox include Ralf Herbrich and David Stern.

Contributed talk: Label Ranking with Abstention: Predicting Partial Orders by Thresholding Probability Distributions

Weiwei Cheng, University of Marburg
Eyke Huellermeier, University of Marburg

We consider an extension of the setting of label ranking, in which the learner is allowed to make predictions in the form of partial instead of total orders. Predictions of that kind are interpreted as a partial abstention: If the learner is not sufficiently certain regarding the relative order of two alternatives, it may abstain from this decision and instead declare these alternatives as being incomparable. We propose a new method for learning to predict partial orders that improves on an existing approach, both theoretically and empirically. Our method is based on the idea of thresholding the probabilities of pairwise preferences between labels as induced by a predicted (parameterized) probability distribution on the set of all rankings.

Contributed talk: Approximate Sorting of Preference Data

Ludwig M. Busse, Morteza Haghir Chehreghani and Joachim M. Buhmann Ludwig Busse, Swiss Federal Institute of Technology
Morteza Chehreghani, Swiss Federal Institute of Technology
Joachim Buhmann, Swiss Federal Institute of Technology

We consider sorting data in noisy conditions. Whereas sorting itself is a well studied topic, ordering items when the comparisons between objects can suffer from noise is a rarely addressed question. However, the capability of handling noisy sorting can be of a prominent importance, in particular in applications such as preference analysis. Here, orderings represent consumer preferences ("rankings") that should be reliably computed despite the fact that individual, simple pairwise comparisons may fail. This paper derives an information theoretic method for approximate sorting. It is optimal in the sense that it extracts as much information as possible from the given observed comparison data conditioned on the noise present in the data. The method is founded on the maximum approximation capacity principle. All formulas are provided together with experimental evidence demonstrating the validity of the new method and its superior rank prediction capability.

Optimization for Machine Learning

http://opt.kyb.tuebingen.mpg.de/index.html

LOCATION
Melia Sierra Nevada: Dauro
Friday, December 16th, 07:30 -- 10:30 AM & 4:00 -- 8:00 PM

Suvrit Sra suvrit@tuebingen.mpg.de
Max Planck Institute for Intelligent Systems

Sebastian Nowozin senowozi@microsoft.com
Microsoft Research

Stephen Wrights wright@cs.wisc.edu
University of Wisconsin

Abstract
Optimization is a well-established, mature discipline. But the way we use this discipline is undergoing a rapid transformation: the advent of modern data intensive applications in statistics, scientific computing, or data mining and machine learning, is forcing us to drop theoretically powerful methods in favor of simpler but more scalable ones. This changeover exhibits itself most starkly in machine learning, where we have to often process massive datasets; this necessitates not only reliance on large-scale optimization techniques, but also the need to develop methods "tuned" to the specific needs of machine learning problems.

INVITED SPEAKERS

Stochastic optimization with non-i.i.d. noise
Alekh Agarwal, University California, Berkeley
John Duchi, University California, Berkeley

We study the convergence of a class of stable online algorithms for stochastic convex optimization in settings where we do not receive independent samples from the distribution over which we optimize, but instead receive samples that are coupled over time. We show the optimization error of the averaged predictor output by any stable online learning algorithm is upper bounded with high probability by the average regret of the algorithm, so long as the underlying stochastic process is β- or Φ-mixing. We additionally show sharper convergence rates when the expected loss is strongly convex, which includes as special cases linear prediction problems including linear and logistic regression, least-squares SVM, and boosting.

Steepest Descent Analysis for Unregularized Linear Prediction with Strictly Convex Penalties
Matus Telgarsky, University of California, San Diego

This manuscript presents a convergence analysis, generalized from a study of boosting, of unregularized linear prediction. Here the empirical risk incorporating strictly convex penalties composed with a linear term may fail to be strongly convex, or even attain a minimizer. This analysis is demonstrated on linear regression, decomposable objectives, and boosting.

SCHEDULE

7:30-7:40	Opening remarks
7:40-8.00	**Stochastic Optimization With Non-i.i.d. Noise** Alekh Agarwal
8:00-8:20	**Steepest Descent Analysis for Unregularized Linear Prediction with Strictly Convex Penalties** Matus Telgarsky
8:20-9:00	Poster Spotlights
9:00-9:30	Coffee Break (POSTERS)
9:30-10:30	Invited Talk: **Convex Optimization: from Real-Time Embedded to Large-Scale Distributed** Stephen Boyd
10:30-16:00	Break (POSTERS)
16:00-17:00	**Invited Talk: To Be Announced** Ben Recht
17:00-17:30	Coffee Break
17:30-17:50	**Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization** Ohad Shamir
17:50-18:10	**Fast First-Order Methods for Composite Convex Optimization with Large Steps** Katya Scheinberg
18:10-18:30	**Coding Penalties for Directed Acyclic Graphs** Julien Mairal
18:30-20:00	Posters continue

Convex Optimization: from Real-Time Embedded to Large-Scale Distributed
Stephen Boyd, Stanford University

Please visit the website at the top of this page for details

Invited Talk: To Be Announced
Ben Recht, University of Wisconsin

Please visit the website at the top of this page for details

Making Gradient Descent Optimal for Strongly Convex Stochastic Optimization

Ohad Shamir, Microsoft Research

Stochastic gradient descent (SGD) is a simple and popular method to solve stochastic optimization problems which arise in machine learning. For strongly convex problems, its convergence rate was known to be $O(log(T)/T)$, by running SGD for T iterations and returning the average point. However, recent results showed that using a different algorithm, one can get an optimal O(1/T) rate. This might lead one to believe that standard SGD is suboptimal, and maybe should even be replaced as a method of choice. In this paper, we investigate the optimality of SGD in a stochastic setting. We show that for smooth problems, the algorithm attains the optimal O(1/T) rate. However, for non-smooth problems, the convergence rate with averaging might really be $\Omega(log(T)/T)$, and this is not just an artifact of the analysis. On the IP side, we show that a simple modification of the averaging step success to recover the O(1/T) rate, and no other change of the algorithm is necessary. We also present experimental results which support our findings, and point out open problems.

Fast First-Order Methods for Composite Convex Optimization with Large Steps

Katya Scheinberg, Lehigh University
Donald Goldfarb, Columbia University

We propose accelerated first-order methods with non-monotonic choice of the prox parameter, which essentially controls the step size. This is in contrast with most accelerated schemes where the prox parameter is either assumed to be constant or non-increasing. In particular we show that a backtracking strategy can be used within FISTA and FALM algorithms starting from an arbitrary parameter value preserving their worst-case iteration complexities of $O(\sqrt{L(f)/\epsilon})$. We also derive complexity estimates that depend on the \average" step size rather than the global Lipschitz constant for the function gradient, which provide better theoretical justification for these methods, hence the main contribution of this paper is theoretical.

Coding Penalties for Directed Acyclic Graphs

Julien Mairal, University California, Berkeley
Bin Yu, University California, Berkeley

We consider supervised learning problems where the features are embedded in a graph, such as gene expressions in a gene network. In this context, it is of much interest to automatically select a subgraph which has a small number of connected components, either to improve the prediction performance, or to obtain better interpretable results. Existing regularization or penalty functions for this purpose typically require solving among all connected subgraphs a selection problem which is combinatorially hard. In this paper, we address this issue for directed acyclic graphs (DAGs) and propose structured sparsity penalties over paths on a DAG (called \path coding" penalties). We design minimum cost flow formulations to compute the penalties and their proximal operator in polynomial time, allowing us in practice to efficiently select a subgraph with a small number of connected components. We present experiments on image and genomic data to illustrate the sparsity and connectivity benefits of path coding penalties over some existing ones as well as the scalability of our approach for prediction tasks.

ACCEPTED POSTERS

Krylov Subspace Descent for Deep Learning

Oriol Vinyals, University California, Berkeley
Daniel Povey, Microsoft Research

Relaxation Schemes for Min Max Generalization in Deterministic Batch Mode Reinforcement Learning

Raphael Fonteneau, University of Liéege
Damien Ernst, University of Liéege
Bernard Boigelot, University of Liéege
Quentin Louveaux, University of Liéege

Non positive SVM

Gaelle Loosli, Clermont Universite, LIMOS
Stephane Canu, LITIS, Insa de Rouen

Accelerating ISTA with an active set strategy

Matthieu Kowalski, Univ Paris-Sud
Pierre Weiss, INSA Toulouse
Alexandre Gramfort, Harvard Medical School
Sandrine Anthoine, CNRS

Learning with matrix gauge regularizers

Miroslav Dudik, Yahoo!
Zaid Harchaoui, INRIA
Jerome Malick, CNRS, Lab. J. Kuntzmann

Online solution of the average cost Kullback-Leibler optimization problem

Joris Bierkens, SNN, Radboud University
Bert Kappen, SNN, Radboud University

An Accelerated Gradient Method for Distributed Multi-Agent Planning with Factored MDPs

Sue Ann Hong, Carnegie Mellon University
Geoffrey Gordon, Carnegie Mellon University

Group Norm for Learning Latent Structural SVMs

Daozheng Chen, University of Maryland, College Park
Dhruv Batra, Toyota Technological Institute at Chicago
Bill Freeman, MIT
Micah Kimo Johnson, GelSight, Inc.

Computational Trade-offs in Statistical Learning

https://sites.google.com/site/costnips/

LOCATION
Montebajo: Basketball Court
Friday, 07:30 -- 10:30 AM & 4:00 -- 8:00 PM

Alekh Agarwal alekh@cs.berkeley.edu
UC Berkeley

Alexander Rakhlin rakhlin@wharton.upenn.edu
U Penn

Abstract
Since its early days, the field of Machine Learning has focused on developing computationally tractable algorithms with good learning guarantees. The vast literature on statistical learning theory has led to a good understanding of how the predictive performance of different algorithms improves as a function of the number of training samples. By the same token, the well-developed theories of optimization and sampling methods have yielded efficient computational techniques at the core of most modern learning methods. The separate developments in these fields mean that given an algorithm we have a sound understanding of its statistical and computational behavior. However, there hasn't been much joint study of the computational and statistical complexities of learning, as a consequence of which, little is known about the interaction and trade-offs between statistical accuracy and computational complexity. Indeed a systematic joint treatment can answer some very interesting questions: what is the best attainable statistical error given a finite computational budget? What is the best learning method to use given different computational constraints and desired statistical yardsticks? Is it the case that simple methods outperform complex ones in computationally impoverished scenarios?

INVITED SPEAKERS

Stochastic Algorithms for One-Pass Learning
Leon Bottou Microsoft Research,

The goal of the presentation is to describe practical stochastic gradient algorithms that process each training example only once, yet asymptotically match the performance of the true optimum. This statement needs, of course, to be made more precise. To achieve this, we'll review the works of Nevel'son and Has'minskij (1972), Fabian (1973, 1978), Murata & Amari (1998), Bottou & LeCun (2004), Polyak & Juditsky (1992), Wei Xu (2010), and Bach & Moulines (2011). We will then show how these ideas lead to practical algorithms and new challenges.

SCHEDULE

7.30-7:40	Opening Remarks
7:40-8:40	Keynote: **Stochastic Algorithms for One-Pass Learning** Leon Bottou
8:40-9:00	Coffee Break and Poster Session
9:00-9:30	**Early stopping for non-parametric regression: An optimal data-dependent stopping rule** Garvesh Raskutti
9:30-10:00	**Statistical and computational tradeoffs in biclustering** Sivaraman Balakrishnan
10-10:30	Contributed short talks
10:30-16:00	Ski break
16:00-17:00	Keynote: **Using More Data to Speed-up Training Time** Shai Shalev-Shwartz
17:00-17:30	Coffee break and Poster Session
17:30-18:00	**Theoretical Basis for "More Data Less Work"?** Nati Srebro
18:00-18:15	Discussion
18:15-18:45	**Anticoncentration regularizers for stochastic combinatorial problems** Shiva Kaul
18:45-19:05	Contributed short talks
19:05-20:00	Last chance to look at posters

Early stopping for non-parametric regression: An optimal data-dependent stopping rule

Garvesh Raskutti University of California Berkeley,
Martin Wainwright University of California Berkeley,
Bin Yu University of California Berkeley,

The goal of non-parametric regression is to estimate an unknown function f based on numobs i.i.d. observations of the form $y_i = f^*(x_i) + w_i$, where $\{w_i\}_{i=1}^{numjobs}$ are additive noise variables. Simply choosing a function to minimize the least-squares loss $\frac{1}{2n}\sum_{i=1}^{n}(y_i - f(x_i))^2$ will lead to "overfitting", so that various estimators are based on different types of regularization. The early stopping strategy is to run an iterative algorithm such as gradient descent for a fixed but finite number of iterations. Early stopping is known to yield estimates with better prediction accuracy than those obtained by running the algorithm for an infinite number of iterations. Although bounds on this prediction error are known for certain function classes and step size choices, the bias-variance tradeoffs for arbitrary reproducing kernel Hilbert spaces (RKHSs) and arbitrary choices of step-sizes have not been well-understood to date. In this paper, we derive upper bounds on both the LTPn and LTP error for arbitrary RKHSs, and provide an explicit and easily computable data-dependent stopping rule. In particular, it depends only on the sum of step-sizes and the eigenvalues of the empirical kernel matrix for the RKHS. For Sobolev spaces and finite-rank kernel classes, we show that our stopping rule yields estimates that achieve the statistically optimal rates in a minimax sense.

Statistical and computational tradeoffs in biclustering

Sivaraman Balakrishnan Carnegie Mellon University,
Mladen Kolar Carnegie Mellon University,
Alessandro Rinaldo Carnegie Mellon University,
Aarti Singh Carnegie Mellon University,
Larry Wasserman Carnegie Mellon University,

We consider the problem of identifying a small sub-matrix of activation in a large noisy matrix. We establish the minimax rate for the problem by showing tight (up to constants) upper and lower bounds on the signal strength needed to identify the sub-matrix. We consider several natural computationally tractable procedures and show that under most parameter scalings they are unable to identify the sub-matrix at the minimax signal strength. While we are unable to directly establish the computational hardness of the problem at the minimax signal strength we discuss connections to some known NP-hard problems and their approximation algorithms.

Using More Data to Speed-up Training Time

Shai-Shalev Shwartz Hebrew University,

Recently, there has been a growing interest in understanding how more data can be leveraged to reduce the required training runtime. I will describe a systematic study of the runtime of learning as a function of the number of available training examples, and underscore the main high-level techniques. In particular, a formal positive result will be presented, showing that even in the unrealizable case, the runtime can decrease exponentially while only requiring a polynomial growth of the number of examples. The construction corresponds to a synthetic learning problem and an interesting open question is if and how the tradeoff can be shown for more natural learning problems. I will spell out several interesting candidates of natural learning problems for which we conjecture that there is a tradeoff between computational and sample complexity.
Based on joint work with Ohad Shamir and Eran Tromer.

Theoretical Basis for "More Data Less Work"?

Nati Srebro TTI Chicago,
Karthik Sridharan TTI Chicago,

We argue that current theory cannot be used to analyze how more data leads to less work, that in-fact for a broad generic class of convex learning problems more data does not lead to less work in the worst case, but in practice, actually more data does lead to less work.

Anticoncentration regularizers for stochastic combinatorial problems

Shiva Kaul Carnegie Mellon University,
Geoffrey Gordon Carnegie Mellon University,

Statistically optimal estimators often seem difficult to compute. When they are the solution to a combinatorial optimization problem, NP-hardness motivates the use of suboptimal alternatives. For example, the non-convex ℓ_0 norm is ideal for enforcing sparsity, but is typically overlooked in favor of the convex ℓ_1 norm. We introduce a new regularizer which is small enough to preserve statistical optimality but large enough to circumvent worst-case computational intractability. This regularizer rounds the objective to a fractional precision and smooths it with a random perturbation. Using this technique, we obtain a combinatorial algorithm for noisy sparsity recovery which runs in polynomial time and requires a minimal amount of data.

Bayesian Nonparametric Methods: Hope or Hype?

http://people.seas.harvard.edu/~rpa/nips2011npbayes.html

LOCATION
Melia Sierra Nevada: Dauro
Saturday, 07:30 -- 10:30 AM & 4:00 -- 8:00 PM

Ryan P. Adams rpa@seas.harvard.edu
Harvard University

Emily B. Fox ebfox@wharton.upenn.edu
University of Pennsylvania

Abstract
Bayesian nonparametric methods are an expanding part of the machine learning landscape. Proponents of Bayesian nonparametrics claim that these methods enable one to construct models that can scale their complexity with data, while representing uncertainty in both the parameters and the structure. Detractors point out that the characteristics of the models are often not well understood and that inference can be unwieldy. Relative to the statistics community, machine learning practitioners of Bayesian nonparametrics frequently do not leverage the representation of uncertainty that is inherent in the Bayesian framework. Neither do they perform the kind of analysis | both empirical and theoretical | to set skeptics at ease. In this workshop we hope to bring a wide group together to constructively discuss and address these goals and shortcomings.

Mini Talks

Transformation Process Priors
Nicholas Andrews, Johns Hopkins University
Jason Eisner, Johns Hopkins University

Latent IBP Compound Dirichlet Allocation
Balaji Lakshminarayanan, Gatsby Computational Neuroscience Unit

Bayesian Nonparametrics for Motif Estimation of Transcription Factor Binding Sites
Philipp Benner, Max Planck Institute
Pierre-Yves Bourguignon, Max Planck Institute
Stephan Poppe, Max Planck Institute

Nonparametric Priors for Finite Unknown Cardinalities of Sampling Spaces
Philipp Benner, Max Planck Institute
Pierre-Yves Bourguignon, Max Planck Institute
Stephan Poppe, Max Planck Institute

SCHEDULE

7:30-7:45	Welcome
7:45-8:45	Plenary Talk: **To Be Announced** Zoubin Ghahramani
8:45-9:15	Coffee Break
9:15-9:45	Poster Session
9:45-10:15	Invited Talk: **To Be Announced** Erik Sudderth
10:15-10:30	Discussant: **To Be Announced** Yann LeCun
16:00-16:30	Invited Talk: **To Be Announced** Igor Pruenster
16:30-16:45	Invited Talk: **To Be Announced** Peter Orbanz
16:45-17:15	Invited Talk: **Designing Scalable Models for the Internet** Alex Smola
17:15-17:30	Discussant: **To Be Announced** Yee Whye Teh
17:30-18:00	Coffee Break
18:00-18:30	Invited Talk: **To Be Announced** Christopher Holmes
18:30-18:45	Discussant: **To Be Announced** To Be Determined
18:45-19:00	Closing Remarks

A Discriminative Nonparametric Bayesian Model: Infinite Hidden Conditional Random Fields
Konstantinos Bousmalis, Imperial College London
Louis-Philippe Morency, University of Southern California
Stefanos Zafeiriou, Imperial College Londong
Maja Pantic, Imperial College London

Infinite Exponential Family Harmoniums
Ning Chen, Tsinghua University Jun Zhu, Tsinghua University
Fuchun Sun, Tsinghua University

Learning in Robotics Using Bayesian Nonparametrics
Marc Peter Deisenroth, TU Darmstadt
Dieter Fox, University of Washington
Carl Edward Rasmussen, University of Cambridge

An Analysis of Activity Changes in MS Patients: A Case Study in the Use of Bayesian Nonparametrics
Finale Doshi--Velez, Massachusetts Institute of Technology
Nicholas Roy, Massachusetts Institute of Technology

GNSS Urban Localization Enhancement Using Dirichlet Process Mixture Modeling
Emmanuel Duflos,

Infinite Multiway Mixture with Factorized Latent Parameters
Isık Barıs Fidaner, Boğaziçi University
A. Taylan Cemgil, Boğaziçi University

A Semiparametric Bayesian Latent Variable Model for Mixed Outcome Data
Jonathan Gruhl,

Nonparametric Bayesian State Estimation in Nonlinear Dynamic Systems with Alpha-Stable Measurement Noise
Nouha Jaoua, Emmanuel Duflos, Philippe Vanheeghe,

Bayesian Multi-Task Learning for Function Estimation with Dirichlet Process Priors
Marcel Hermkes, University of Potsdam Nicolas Kuehn, University of Potsdam Carsten Riggelsen, University of Potstdam

A Bayesian Nonparametric Clustering Application on Network Traffic Data
Barıs Kurt, Boğaziçi University
A. Taylan Cemgil, Boğaziçi University

Video Streams Semantic Segmentation Utilizing Multiple Channels with Different Time Granularity
Bado Lee, Seoul National University
Ho-Sik Seok, Seoul National University
Byoung-Tak Zhang, Seoul National University

Efficient Inference in the Infinite Multiple Membership Relational Model
Morten Mørup, Technical University of Denmark
Mikkel N. Schmidt, Technical University of Denmark

Gaussian Process Dynamical Models for Phoneme Classification
Hyunsin Park,
Chang D. Yoo,

PBART: Parallel Bayesian Additive Regression Trees
Matthew T. Pratola, Los Alamos National Laboratory Robert E. McCulloch, University of Texas at Austin James Gattiker, Los Alamos National Laboratory Hugh A. Chipman, Acadia University, David M. Higdon, Los Alamos National Laboratory

Bayesian Nonparametric Methods Are Naturally Well Suited to Functional Data Analysis
Asma Rabaoui, LAPS-IMS/CNRS Hachem Kadri, Sequel-INRIA Lille Manuel Davy, LAGIS/CNRS/Vekia SAS

Hierarchical Models of Complex Networks
Mikkel N. Schmidt, Technical University of Denmark
Morten Mørup, Technical University of Denmark
Tue Herlau, Technical University of Denmark

Pathological Properties of Deep Bayesian Hierarchies
Jacob Steinhardt, Massachusetts Institute of Technology
Zoubin Ghahramani, University of Cambridge

Modeling Streaming Data in the Absence of Sufficiency
Frank Wood, Columbia University

Bayesian Nonparametric Imputation of Missing Design Information Under Informative Survey Samples
Sahar Zangeneh, University of Michigan

Fast Variational Inference for Dirichlet Process Mixture Models
Matteo Zanotto, Istituto Italiano di Tecnologia
Vittorio Murino, Istituto Italiano di Tecnologia

Sparse Representation and Low-rank Approximation

http://www.cs.berkeley.edu/~ameet/sparse-low-rank-nips11

LOCATION
Montebajo: Room I
Friday, December 16th, 07:30 -- 10:30 AM & 4:00 -- 8:00 PM

Francis Bach
INRIA - Ecole Normale Supérieure

Michael Davies
University of Edinburgh

Rémi Gribonval
INRIA

Lester Mackey
University of California at Berkeley

Michael Mahoney
Stanford University

Mehryar Mohri
Courant Institute (NYU) and Google Research

Guillaume Obozinski
INRIA - Ecole Normale Supérieure

Ameet Talwalkar
University of California at Berkeley

Abstract

Sparse representation and low-rank approximation are fundamental tools in fields as diverse as computer vision, computational biology, signal processing, natural language processing, and machine learning. Recent advances in sparse and low-rank modeling have led to increasingly concise descriptions of high dimensional data, together with algorithms of provable performance and bounded complexity. Our workshop aims to survey recent work on sparsity and low-rank approximation and to provide a forum for open discussion of the key questions concerning these dimensionality reduction techniques. The workshop will be divided into two segments, a "sparsity segment" emphasizing sparse dictionary learning and a "low-rank segment" emphasizing scalability and large data.

The sparsity segment will be dedicated to learning sparse latent representations and dictionaries: decomposing a signal or a vector of observations as sparse linear combinations of basis vectors, atoms or covariates is ubiquitous in machine learning and signal processing. Algorithms and theoretical analyzes for obtaining these decompositions are now numerous. Learning the atoms or basis vectors directly from data has proven useful in several domains and is often seen from different viewpoints: (a) as a matrix factorization problem with potentially some constraints such as pointwise non-negativity, (b) as a latent variable model which can be treated in a probabilistic and potentially Bayesian way, leading in particular to topic models, and (c) as dictionary learning with often a goal of signal representation or restoration. The goal of this part of the workshop is to confront these various points of view and foster exchanges of ideas among the signal processing, statistics, machine learning and applied mathematics communities.

The low-rank segment will explore the impact of low-rank methods for large-scale machine learning. Large datasets often take the form of matrices representing either a set of real-valued features for each data point or pairwise similarities between data points. Hence, modern learning problems face the daunting task of storing and operating on matrices with millions to billions of entries. An attractive solution to this problem involves working with low-rank approximations of the original matrix. Low-rank approximation is at the core of widely used algorithms such as Principal Component Analysis and Latent Semantic Indexing, and low-rank matrices appear in a variety of applications including lossy data compression, collaborative filtering, image processing, text analysis, matrix completion, robust matrix factorization and metric learning. In this segment we aim to study new algorithms, recent theoretical advances and large-scale empirical results, and more broadly we hope to identify additional interesting scenarios for use of low-rank approximations for learning tasks.

SCHEDULE

7.30-7.40	Opening remarks
7.40-8.10	Invited Talk: **Local Analysis of Sparse Coding in the Presence of Noise** Rodolphe Jenatton
8.10-8.30	**Recovery of a Sparse Integer Solution to an Underdetermined System of Linear Equations** T.S. Jayram, Soumitra Pal, Vijay Arya
8.30-8.40	Coffee Break
8.40-9.10	Invited Talk: **Robust Sparse Analysis Regularization** Gabriel Peyre
9.10-9.40	Poster Session
9.40-10.10	Invited Talk: **Dictionary-Dependent Penalties for Sparse Estimation and Rank Minimization** David Wipf
10.10-10.30	**Group Sparse Hidden Markov Models** Jen-Tzung Chien, Cheng-Chun Chiang
10.30-16.00	Break
16.00-16.35	Invited Talk: **To Be Announced** Martin Wainwright
16.35-17.20	Contributed Mini-Talks
17.20-17.50	Poster Session/Coffee Break
17.50-18.25	Invited Talk: **To Be Announced** Yi Ma
18.25-19.00	Invited Talk: **To Be Announced** Inderjit Dhillon

INVITED SPEAKERS

Local Analysis of Sparse Coding in the Presence of Noise

Rodolphe Jenatton, INRIA / Ecole Normale Superieure

A popular approach within the signal processing and machine learning communities consists in modeling signals as sparse linear combinations of atoms selected from a learned dictionary. While this paradigm has led to numerous empirical successes in various fields ranging from image to audio processing, there have only been a few theoretical arguments supporting these evidences. In particular, sparse coding, or sparse dictionary learning, relies on a non-convex procedure whose local minima have not been fully analyzed yet. In this paper, we consider a probabilistic model of sparse signals, and show that, with high probability, sparse coding admits a local minimum around the reference dictionary generating the signals. Our study takes into account the case of over complete dictionaries and noisy signals, thus extending previous work limited to noiseless settings and/or under-complete dictionaries. The analysis we conduct is non-asymptotic and makes it possible to understand how the key quantities of the problem, such as the coherence or the level of noise, are allowed to scale with respect to the dimension of the signals, the number of atoms, the sparsity and the number of observations.

Recovery of a Sparse Integer Solution to an Underdetermined System of Linear Equations

T.S. Jayram, IBM Research - Almaden
Soumitra Pal, CSE, IIT - Bombay
Vijay Arya, IBM Research - India

We consider a system of m linear equations in n variables Ax = b where A is a given m x n matrix and b is a given m-vector known to be equal to A$\bar{x}$ for some unknown solution $\bar{x}$ that is integer and k-sparse: $\bar{x}$ ϵ {0; 1}^{n} and exactly k entries of x are 1. We give necessary and sufficient conditions for recovering the solution $\bar{x}$ exactly using an LP relaxation that minimizes the ℓ_1 norm of x. When A is drawn from a distribution that has exchangeable columns, we show an interesting connection between the recovery probability and a well known problem in geometry, namely the k-set problem. To the best of our knowledge, this connection appears to be new in the compressive sensing literature. We empirically show that for large n if the elements of A are drawn i.i.d. from the normal distribution then the performance of the recovery LP exhibits a phase transition, i.e., for each k there exists a value $\bar{m}$ of m such that the recovery always succeeds if $m > \bar{m}$ and always fails if $m < \bar{m}$. Using the empirical data we conjecture that $\bar{m} = n$H(k/n)/2 where H(x) = -$x\ log_2\ x$ - (1 - x) log_2(1 - x) is the binary entropy function.

Robust Sparse Analysis Regularization

Gabriel Peyré, CNRS, CEREMADE, Université Paris-Dauphine

In this talk I will detail several key properties of ℓ_1-analysis regularization for the resolution of linear inverse problems. Most previous theoretical works consider sparse synthesis priors where the sparsity is measured as the norm of the coefficients that synthesize the signal in a given dictionary. In contrast, the more general analysis regularization minimizes the ℓ_1 norm of the correlations between the signal and the atoms in the dictionary. The corresponding variational problem includes several well-known regularizations such as the discrete total variation, the fused lasso and sparse correlation with translation invariant wavelets. I will first study the variations of the solution with respect to the observations and the regularization parameter, which enables the computation of the degrees of freedom estimator. I will then give a sufficient condition to ensure that a signal is the unique solution of the analysis regularization when there is no noise in the observations. The same criterion ensures the robustness of the sparse analysis solution to a small noise in the observations. Lastly I will define a stronger condition that ensures robustness to an arbitrary bounded noise. In the special case of synthesis regularization, our contributions recover already known results, that are hence generalized to the analysis setting. I will illustrate these theoretical results on practical examples to study the robustness of the total variation, fused lasso and translation invariant wavelets regularizations.
This is joint work with S. Vaiter, C. Dossal, J. Fadili

Dictionary-Dependent Penalties for Sparse Estimation and Rank Minimization

David Wipf, University of California at San Diego

In the majority of recent work on sparse estimation algorithms, performance has been evaluated using ideal or quasi-ideal dictionaries (e.g., random Gaussian or Fourier) characterized by unit `2 norm, incoherent columns or features. But these types of dictionaries represent only a subset of the dictionaries that are actually used in practice (largely restricted to idealized compressive sensing applications). In contrast, herein sparse estimation is considered in the context of structured dictionaries possibly exhibiting high coherence between arbitrary groups of columns and/or rows. Sparse penalized regression models are analyzed with the purpose of finding, to the extent possible, regimes of dictionary invariant performance. In particular, a class of non-convex, Bayesian-inspired estimators with dictionary-dependent sparsity penalties is shown to have a number of desirable invariance properties leading to provable advantages over more conventional penalties such as the `1 norm, especially in areas where existing theoretical recovery guarantees no longer hold. This can translate into improved performance in applications such model selection with correlated features, source localization, and compressive sensing with constrained measurement directions. Moreover, the underlying methodology naturally extends to related rank minimization problems.

Group Sparse Hidden Markov Models
Jen-Tzung Chien, National Cheng Kung University, Taiwan
Cheng-Chun Chiang, National Cheng Kung University, Taiwan

This paper presents the group sparse hidden Markov models (GS-HMMs) for speech recognition where a sequence of acoustic features is driven by a Markov chain and each feature vector is represented by two groups of basis vectors. The group of common bases is used to represent the features corresponding to different states within an HMM. The group of individual bases is used to compensate intra-state residual information. Importantly, the sparse prior for sensing weights is specified by the Laplacian scale mixture distribution which is obtained by multiplying Laplacian distribution with an inverse scale mixture parameter. This parameter makes the distribution even sparser and serves as an automatic relevance determination to control the degree of sparsity through selecting the relevant bases in two groups. The parameters of GS-HMMs, including weights and two sets of bases, are estimated via Bayesian learning. We apply this framework for acoustic modeling and show the effectiveness of GS-HMMs for speech recognition in presence of different noises types and SNRs.

Invited Talk: To Be Announced
Martin Wainwright, University of California at Berkeley

Please see the website on page 68 for details

Invited Talk: To Be Announced
Yi Ma, University of Illinois at Urbana-Champaign

Please see the website on page 68 for details

Invited Talk: To Be Announced
Inderjit Dhillon, University of Texas at Austin

Please see the website on page 68 for details

Mini Talks

Automatic Relevance Determination in Nonnegative Matrix Factorization with the fi-Divergence (mini-talk)
Vincent Y. F. Tan, University of Wisconsin-Madison
Cédric Févotte, CNRS LTCI, TELECOM ParisTech

Coordinate Descent for Learning with Sparse Matrix Regularization (mini-talk)
Miroslav Dudik, Yahoo! Research
Zaid Harchaoui, LEAR, INRIA and LJK
Jerome Malick, CNRS and LJK

Divide-and-Conquer Matrix Factorization (mini-talk)
Lester Mackey, University of California, Berkeley
Ameet Talwalkar, University of California, Berkeley
Michael I. Jordan, University of California, Berkeley

Learning with Latent Factors in Time Series (mini-talk)
Ali Jalali, University of Texas at Austin
Sujay Sanghavi, University of Texas at Austin

Low-rank Approximations and Randomized Sampling (mini-talk)
Ming Gu, University of California, Berkeley

Discrete Optimization in Machine Learning (DISCML): Uncertainty, Generalization and Feedback

WS27

http://discml.cc

LOCATION
Melia Sol y Nieve: Slalom
Saturday, 07:30 -- 10:30 AM & 4:00 -- 8:00 PM

Andreas Krause
ETH Zurich

Pradeep Ravikumar,
University of Texas, Austin

Stefanie Jegelka,
Max Planck Institute for Biological Cybernetics

Jeff Bilmes
University of Washington

Abstract

Solving optimization problems with ultimately discrete solutions is becoming increasingly important in machine learning: At the core of statistical machine learning is to infer conclusions from data, and when the variables underlying the data are discrete, both the tasks of inferring the model from data, as well as performing predictions using the estimated model are discrete optimization problems. Many of the resulting optimization problems are NP-hard, and typically, as the problem size increases, standard off-the-shelf optimization procedures become intractable.

Fortunately, most discrete optimization problems that arise in machine learning have specific structure, which can be leveraged in order to develop tractable exact or approximate optimization procedures. For example, consider the case of a discrete graphical model over a set of random variables. For the task of prediction, a key structural object is the “marginal polytope,” a convex bounded set characterized by the underlying graph of the graphical model. Properties of this polytope, as well as its approximations, have been successfully used to develop efficient algorithms for inference. For the task of model selection, a key structural object is the discrete graph itself. Another problem structure is sparsity: While estimating a high-dimensional model for regression from a limited amount of data is typically an ill-posed problem, it becomes solvable if it is known that many of the coefficients are zero. Another problem structure, submodularity, a discrete analog of convexity, has been shown to arise in many machine learning problems, including structure learning of probabilistic models, variable selection and clustering. One of the primary goals of this workshop is to investigate how to leverage such structures.

The focus of this year’s workshop is on the interplay between discrete optimization and machine learning: How can we solve inference problems arising in machine learning using discrete optimization? How can one solve discrete optimization problems that themselves are learned from training data? How can we solve challenging sequential and adaptive discrete optimization problems where we have the opportunity to incorporate feedback (online and active learning with combinatorial decision spaces)? We will also explore applications of such approaches in computer vision, NLP, information retrieval, etc.

SCHEDULE

7:30-7:50	Introduction
7:50-8:40	Invited talk: **Exploiting Problem Structure for Efficient Discrete Optimization** Pushmeet Kohli
8.40-9:00	Poster Spotlights
9:00-9:15	Coffee Break
9:15-10:05	Invited talk: **Learning with Submodular Functions: A Convex Optimization Perspective** Francis Bach
10.05-10.30	Poster Spotlights
10.30-4.00	Break
4.00-4.30	Poster Spotlights
4.30-5.50	Keynote talk: **Polymatroids and Submodularity** Jack Edmonds
5.50-6.20	Coffee & Posters
6.20-7.10	Invited Talk: **Combinatorial prediction games** Nicolo Cesa-Bianchi

INVITED SPEAKERS

Exploiting Problem Structure for Efficient Discrete Optimization

Pushmeet Kohli, Microsoft Research

Many problems in computer vision and machine learning require inferring the most probable states of certain hidden or unobserved variables. This inference problem can be formulated in terms of minimizing a function of discrete variables. The scale and form of computer vision problems raise many challenges in this optimization task. For instance, functions encountered in vision may involve millions or sometimes even billions of variables. Furthermore, the functions may contain terms that encode very high-order interaction between variables. These properties ensure that the minimization of such functions using conventional algorithms is extremely computationally expensive. In this talk, I will discuss how many of these challenges can be overcome by exploiting the sparse and heterogeneous nature of discrete optimization problems encountered in real world computer vision problems. Such problem-aware approaches to optimization can lead to substantial improvements in running time and allow us to produce good solutions to many important problems.

Learning with Submodular Functions: A Convex Optimization Perspective

Francis Bach, INRIA

Submodular functions are relevant to machine learning for mainly two reasons: (1) some problems may be expressed directly as the optimization of submodular functions and (2) the Lovasz extension of submodular functions provides a useful set of regularization functions for supervised and unsupervised learning. In this talk, I will present the theory of submodular functions from a convex analysis perspective, presenting tight links between certain polyhedra, combinatorial optimization and convex optimization problems. In particular, I will show how submodular function minimization is equivalent to solving a wide variety of convex optimization problems. This allows the derivation of new efficient algorithms for approximate submodular function minimization with theoretical guarantees and good practical performance. By listing examples of submodular functions, I will also review various applications to machine learning, such as clustering or subset selection, as well as a family of structured sparsity-inducing norms that can be derived and used from submodular functions.

Polymatroids and Submodularity

Jack Edmonds, University of Waterloo (Retired)
John von Neumann, Theory Prize Recipient

Many polytime algorithms have now been presented for minimizing a submodular function f (S) over the subsets S of a finite set E. We provide a tutorial in (somewhat hidden) theoretical foundations of them all. In particular, f can be easily massaged to a set function g(S) which is submodular, non-decreasing, and zero on the empty set, so that minimizing f (S) is equivalent to repeatedly determining whether a point x is in the polymatroid, P (g) = {x : x ≥ 0 and, for every S, sum of $x(j)$ over j in S is at most g(S)}. A fundamental theorem says that, assuming g(S) is integer, the 0,1 vectors x in P (g) are the incidence vectors of the independent sets of a matroid M (P (g)). Another gives an easy description of the vertices of P (g). We will show how these ideas provide beautiful, but complicated, polytime algorithms for the possibly useful optimum branching system problem.

Combinatorial prediction games

Nicoló Cesa-Nianchi, Universitá degli Studi di Milano

Combinatorial prediction games are problems of online linear optimization in which the action space is a combinatorial space. These games can be studied under different feedback models: full, semi-bandit, and bandit. In first part of the talk we will describe the main known facts about these models and mention some of the open problems. In the second part we will focus on the bandit feedback and describe some recent results which strengthen the link between bandit optimization and convex geometry.

NOTES

NOTES