CS 84010 Big Data Analytics

Fall 2013 - Hanghang Tong

Course Information

We are in the age of big data - data that is both large and complex. Big data analytics is the science of analyzing the data, generating insights, and making decisions by addressing all three dimensions of the big data challenges including variety, velocity and volume. It is essential behind many high impact real applications, such as social networks analysis, finance and business intelligence, climate modeling, health care, political science and so on.

This class aims to provide a comprehensive overview of recent advance in machine learning and data mining to analyze big data. Selected topics include big data clustering and classification, anomaly and fraud detection, time-series analysis, big graph mining, and massive-scale data analytics; as well as case studies in social networks analysis, healthcare, business intelligence, etc.

- Instructor: Hanghang Tong (tong at ccny dot cuny dot edu)
  - Gc 4420, x8196
  - Office hour: 12:45-1:45pm Monday, 4420
- Class meets: Mon 2-4pm, 3212
- Late policy:
  - each person has 2 slip days in total for the whole semester. After that, 20% deduction per day of delay
  - no penalty if medical emergence (need doctor's notes)
- Big Data Seminar (10/1st, bi-weekly afterwards): Science Center, room 4102

Schedule

Class meeting			
Lecture	Date	Topics	Remarks
1	*Sep 2*	*Labor day, no class*	
2	Sep 9	Introduction + Basic Concepts *Recommended reading* (1) "Challenges and Opportunities with Big Data"	
3	Sep 16	Classification 1 (Bayes Clasifier, kNN) + Candidate projects *Recommended reading* (1) Theoretic Properties of knn (2) K-d-tree tutorial	
4	Sep 23	Classification 2 – Indexing, Logistic Regression	

		Recommended reading *(1) Logistic Regression with SGD*	
5	Sep 30	Classification 3 - SVM *Recommended reading* *(1) SVM_Perf* *(2) Pegasos*	Proj. proposal seminar reading 1 due
6	Oct 7	Clustering 1 – Kmeans and PCA	seminar reading 2 due
7	*Oct 14*	*Columbus day, no class*	
8	Oct 15	Oct 14 schedule: Web Fraud Detection (Tim Pan)	
9	Oct 21	Clustering -2 – LSH, GMM and spectral clustering	seminar reading 3 due
10	**Oct 28**	**Midterm exam**	
11	Nov 4	Time Series *Recommended reading* *(1) DnS algorithm* *(2) Spirit*	Proj midterm
12	Nov 11	No class, Prof. Terzi's seminar moved to Sep. 20th	reading 1 due
13	Nov 18	Big Graph Mining 1 – Co-clustering *Recommended reading* (1) Co-clustering (2) Cross-association	
14	Nov 25	Big Graph Mining 2 –Low-rank approximation (1) Colibri (2) NMF	reading 2 due
15	Dec 2	Big Graph Mining 3 – Pattern, Dissemination and Proximity Mining Rare Events from Big Data (Jingrui He)	reading 3 due
16	Dec 11	project presentation at noon, 4421	
17	Dec 12	Project final report due	

Big Data Seminar

Seminar	Date	Speaker	Topics
1	Sep. 20th	Evimaria Terzi (Boston)	Entity Selection and Ranking in Data Mining Applications Related paper
2	Oct. 1	Fei Wang (IBM)	Feature Engineering for Predictive Modeling with Large Scale Electronic Medical Records: Augmentation, Densification and Selection Related Papers (a); (b)
3	Oct. 15	Tim Pan (Google)	Click Fraud - Challenges and Remedies Related Papers: (a) - TSum: Fast, Principled Table Summarization (b) - The Goals and Challenges of Click Fraud Penetration Testing Systems
4	Nov. 5	Han Liu (Princeton)	From High Dimensional Data to Big Data

			Related Paper (a) Challenges of big data analysis (b) Huge Package
5	Nov. 12	Ruoming Jin (Kent)	Finally, Simple, Fast and Scalable Reachability Oracle!
6	Dec 10	Tao Li (FIU)	Learning to Understand Documents