Matrix and Tensor Factorization Methods for Natural Language Processing

Guillaume Bouchard[*] **Jason Naradowsky**[#] **Sebastian Riedel**[#]
Tim Rocktäschel[#] and **Andreas Vlachos**[#]
[*] Xerox Research Centre Europe
guillaume.bouchard@xerox.com
[#] Computer Science Department
University College London
{j.narad, s.riedel, t.rocktaschel, a.vlachos}@cs.ucl.ac.uk

1 Tutorial Objectives

Tensor and matrix factorization methods have attracted a lot of attention recently thanks to their successful applications to information extraction, knowledge base population, lexical semantics and dependency parsing. In the first part, we will first cover the basics of matrix and tensor factorization theory and optimization, and then proceed to more advanced topics involving convex surrogates and alternative losses. In the second part we will discuss recent NLP applications of these methods and show the connections with other popular methods such as transductive learning, topic models and neural networks. The aim of this tutorial is to present in detail applied factorization methods, as well as to introduce more recently proposed methods that are likely to be useful to NLP applications.

2 Tutorial Overview

2.1 Matrix/Tensor Factorization Basics

In this part, we first remind essential results on bilinear forms, spectral representations of matrices and low-rank approximation theorems, which are often omitted in undergraduate linear algebra courses. This includes the link between eigenvalue decomposition and singular value decomposition and the trace-norm (a.k.a. nuclear norm) as a convex surrogate of the low-rank constraint on optimization problems. Then, an overview of the most efficient algorithms to solve low-rank constrained problems is made, from the power iteration method, the Lanczos algorithm and the implicitly restarted Arnoldi method that is implemented in the LAPACK library (Anderson et al., 1999). We show how to interpret low-rank models as probabilistic models (Bishop, 1999) and how we can extend SVD algorithms that can factorize non-standard matrices (i.e. with non-Gaussian noise and missing data) using gradient descent, reweighted SVD or Frank-Wolfe algorithms. We then show that combining different convex objectives can be a powerful tool, and we illustrate it by deriving the robust PCA algorithm by adding an L_1 penalty term in the objective function (Candès and Recht, 2009). Furthermore, we introduce Bayesian Personalized Ranking (BPR) for matrix and tensor factorization which deals with implicit feedback in ranking tasks (Rendle et al., 2009). Finally, will introduce the collective matrix factorization model (Singh and Gordon, 2008) and tensor extensions (Nickel et al., 2011) for relational learning.

2.2 Applications in NLP

In this part we will discuss recent work applying matrix/tensor factorization methods in the context of NLP. We will review the Universal Schema paradigm for knowledge base construction (Riedel et al., 2013) which relies on matrix factorization and BPR, as well as recent extensions of the RESCAL tensor factorization (Nickel et al., 2011) approach and methods of injecting logic into the embeddings learned (Rocktäschel et al., 2015). These applications will motivate the connections between matrix factorization and transductive learning (Goldberg et al., 2010), as well as tensor factorization and multi-task learning (Romera-Paredes et al., 2013). Furthermore, we will review work on applying matrix and tensor factorization to sparsity reduction in syntactic dependency parsing (Lei et al., 2014) and word representation learning (Pennington et al., 2014). In addition, we will discuss the connections between matrix factorization, latent semantic analysis and topic modeling (Stevens et al., 2012).

Proceedings of the Tutorials of the 53rd Annual Meeting of the ACL and the 7th IJCNLP, pages 16–18,
Beijing, China, July 26-31, 2015.

3 Structure

Part I: Matrix/Tensor Factorization Basics (90 minutes)

- Matrix factorization basics (40 min): bilinear forms, spectral representations, low rank approximations theorems, optimization with stochastic gradient descent, losses
- Tensor factorization basics (20 minutes): representations,notation decompositions (Tucker etc.)
- Advanced topics (30 minutes): convex surrogates, L1 regularization, alternative losses (ranking loss, logistic loss)

Break (15 minutes)

Part II: Applications in NLP (75 minutes)

- Information extraction, knowledge base population with connections to transductive learning and multitask learning (35 minutes)
- Lexical semantics with connections to neural networks, latent semantic analysis and topic models (30 minutes)
- Structured prediction (10 minutes)

4 About the Speakers

Guillaume Bouchard is a senior researcher in statistics and machine learning at Xerox, focusing on statistical learning using low-rank model for large relational databases. His research includes text understanding, user modeling, and social media analytics. The theoretical part of his work is related to the efficient algorithms to compute high dimensional integrals, essential to deal with uncertainty (missing and noisy data, latent variable models, Bayesian inference). The main application areas of his work includes the design of virtual conversational agents, link prediction (predictive algorithms for relational data), social media monitoring and transportation analytics. His web page is available at `www.xrce.xerox.com/people/bouchard`.

Jason Naradowsky is a postdoc at the Machine Reading group at UCL. Having previously obtained a PhD at UMass Amherst under the supervision of David Smith and Mark Johnson, his current research aims to improve natural language understanding by performing task-specific training of word representations and parsing models. He is also interested in semi-supervised learning, joint inference, and semantic parsing. His web page is available at `http://narad.github.io/`.

Sebastian Riedel is a senior lecturer at University College London and an Allen Distinguished Investigator, leading the Machine Reading Lab. Before, he was a postdoc and research scientist with Andrew McCallum at UMass Amherst, a researcher at Tokyo University and DBCLS with Tsujii Junichi, and a PhD student with Ewan Klein at the University of Edinburgh. He is interested in teaching machines how to read and works at the intersection of Natural Language Processing (NLP) and Machine Learning, investigating various stages of the NLP pipeline, in particular those that require structured prediction, as well as fully probabilistic architectures of end-to-end reading and reasoning systems. Recently he became interested in new ways to represent textual knowledge using low-rank embeddings and how to reason with such representations. His web page is available at `http://www.riedelcastro.org/`.

Tim Rocktäschel is a PhD student in Sebastian Riedel's Machine Reading group at University College London. Before that he worked as research assistant in the Knowledge Management in Bioinformatics group at Humboldt-Universität zu Berlin, where he also obtained his Diploma in Computer Science. He is broadly interested in representation learning (e.g. matrix/tensor factorization, deep learning) for NLP and automated knowledge base completion, and how these methods can take advantage of symbolic background knowledge. His webpage is available at `http://rockt.github.io/`.

Andreas Vlachos is postdoc at the Machine Reading group at UCL working with Sebastian Riedel on automated fact-checking using low-rank factorization methods. Before that he was a postdoc at the Natural Language and Information Processing group at the University of Cambridge and at the University of Wisconsin-Madison. He is broadly interested in natural language understanding (e.g. information extraction, semantic parsing) and in machine learning approaches that would help us towards this goal. He has also worked on active learning, clustering and biomedical text mining. His web page is available at `http://sites.google.com/site/andreasvlachos/`.

References

[Anderson et al.1999] Edward Anderson, Zhaojun Bai, Christian Bischof, Susan Blackford, James Demmel, Jack Dongarra, Jeremy Du Croz, Anne Greenbaum, S Hammerling, Alan McKenney, et al. 1999. *LAPACK Users' guide*, volume 9. SIAM.

[Bishop1999] Christopher M Bishop. 1999. Bayesian PCA. In *Advances in Neural Information Processing Systems*, pages 382–388.

[Candès and Recht2009] Emmanuel J Candès and Benjamin Recht. 2009. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772.

[Goldberg et al.2010] Andrew Goldberg, Ben Recht, Junming Xu, Robert Nowak, and Xiaojin Zhu. 2010. Transduction with matrix completion: Three birds with one stone. In *Advances in Neural Information Processing Systems 23*, pages 757–765.

[Lei et al.2014] Tao Lei, Yu Xin, Yuan Zhang, Regina Barzilay, and Tommi Jaakkola. 2014. Low-rank tensors for scoring dependency structures. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1381–1391.

[Nickel et al.2011] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 809–816.

[Pennington et al.2014] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

[Rendle et al.2009] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 452–461.

[Riedel et al.2013] Sebastian Riedel, Limin Yao, Benjamin M. Marlin, and Andrew McCallum. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, GA.

[Rocktäschel et al.2015] Tim Rocktäschel, Sameer Singh, and Sebastian Riedel. 2015. Injecting Logical Background Knowledge into Embeddings for Relation Extraction . In *Proceedings of the 2015 Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*.

[Romera-Paredes et al.2013] Bernardino Romera-Paredes, Hane Aung, Nadia Bianchi-Berthouze, and Massimiliano Pontil. 2013. Multilinear multitask learning. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1444–1452.

[Singh and Gordon2008] Ajit P. Singh and Geoffrey J. Gordon. 2008. Relational learning via collective matrix factorization. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 650–658.

[Stevens et al.2012] Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 952–961.