Spectral Learning Algorithms for Natural Language Processing

Shay Cohen*, Michael Collins*, Dean P. Foster§, Karl Stratos*, Lyle Ungar§
*Columbia University
§University of Pennsylvania
scohen,mcollins,stratos@cs.columbia.edu
dean@foster.net
ungar@cis.upenn.edu

1 Introduction

Recent work in machine learning and NLP has developed spectral algorithms for many learning tasks involving latent variables. Spectral algorithms rely on singular value decomposition as a basic operation, usually followed by some simple estimation method based on the method of moments. From a theoretical point of view, these methods are appealing in that they offer consistent estimators (and PAC-style guarantees of sample complexity) for several important latent-variable models. This is in contrast to the EM algorithm, which is an extremely successful approach, but which only has guarantees of reaching a local maximum of the likelihood function.

From a practical point of view, the methods (unlike EM) have no need for careful initialization, and have recently been shown to be highly efficient (as one example, in work under submission by the authors on learning of latent-variable PCFGs, a spectral algorithm performs at identical accuracy to EM, but is around 20 times faster).

2 Outline

In this tutorial we will aim to give a broad overview of spectral methods, describing theoretical guarantees, as well as practical issues. We will start by covering the basics of singular value decomposition and describe efficient methods for doing singular value decomposition. The SVD operation is at the core of most spectral algorithms that have been developed.

Tutorials, NAACL-HLT 2013, pages 13–15,
Atlanta, Georgia, June 9 2013.

We will then continue to cover canonical correlation analysis (CCA). CCA is an early method from statistics for dimensionality reduction. With CCA, two or more views of the data are created, and they are all projected into a lower dimensional space which maximizes the correlation between the views. We will review the basic algorithms underlying CCA, give some formal results giving guarantees for latent-variable models and also describe how they have been applied recently to learning lexical representations from large quantities of unlabeled data. This idea of learning lexical representations can be extended further, where unlabeled data is used to learn underlying representations which are subsequently used as additional information for supervised training.

We will also cover how spectral algorithms can be used for structured prediction problems with sequences and parse trees. A striking recent result by Hsu, Kakade and Zhang (2009) shows that HMMs can be learned efficiently using a spectral algorithm. HMMs are widely used in NLP and speech, and previous algorithms (typically based on EM) were guaranteed to only reach a local maximum of the likelihood function, so this is a crucial result. We will review the basic mechanics of the HMM learning algorithm, describe its formal guarantees, and also cover practical issues.

Last, we will cover work about spectral algorithms in the context of natural language parsing. We will show how spectral algorithms can be used to estimate the parameter models of latent-variable PCFGs, a model which serves as the base for state-of-the-art parsing models such as the one of Petrov et al. (2007). We will show what are the practical steps that are needed to be taken in order to make spectral algorithms for L-PCFGs (or other models in general) practical and comparable to state of the art.

3 Speaker Bios

Shay Cohen[1] is a postdoctoral research scientist in the Department of Computer Science at Columbia University. He is a computing innovation fellow. His research interests span a range of topics in natural language processing and machine learning. He is especially interested in developing efficient and scalable parsing algorithms as well as learning algorithms for probabilistic grammars.

Michael Collins[2] is the Vikram S. Pandit Professor of computer science at Columbia University. His research is focused on topics including statistical parsing, structured prediction problems in machine learning, and applications including machine translation, dialog systems, and speech recognition. His awards include a

[1] `http://www.cs.columbia.edu/~scohen/`

[2] `http://www.cs.columbia.edu/~mcollins/`

Sloan fellowship, an NSF career award, and best paper awards at EMNLP (2002, 2004, and 2010), UAI (2004 and 2005), and CoNLL 2008.

Dean P. Foster[3] is currently the Marie and Joseph Melone Professor of Statistics at the Wharton School of the University of Pennsylvania. His current research interests are machine learning, stepwise regression and computational linguistics. He has been searching for new methods of finding useful features in big data sets. His current set of hammers revolve around fast matrix methods (which decompose 2nd moments) and tensor methods for decomposing 3rd moments.

Karl Stratos[4] is a Ph.D. student in the Department of Computer Science at Columbia. His research is focused on machine learning and natural language processing. His current research efforts are focused on spectral learning of latent-variable models, or more generally, uncovering latent structure from data.

Lyle Ungar[5] is a professor at the Computer and Information Science Department at the University of Pennsylvania. His research group develops scalable machine learning and text mining methods, including clustering, feature selection, and semi-supervised and multi-task learning for natural language, psychology, and medical research. Example projects include spectral learning of language models, multi-view learning for gene expression and MRI data, and mining social media to better understand personality and well-being.

[3] http://gosset.wharton.upenn.edu/~foster/index.pl
[4] http://www.cs.columbia.edu/~stratos/
[5] http://www.cis.upenn.edu/~ungar/