Literature Review: Spectral Learning for Natural Language Processing

Zihan Zhang

Abstract

Spectral learning method is used to estimate parameters for natural language models with latent variables. Compared with the EM algorithm, spectral learning guarantees consistent parameter estimates and is much more efficient. This article reviews the applications of spectral learning method, especially on Hidden Markov Models and latent-variable probabilistic context-free grammars. At the end, we discuss the possible future research directions in this area.

1 Introduction

Statistical models with latent variables play an important role in natural language processing (NLP), where the latent variables are not observable from the data. Latent-variable models are involved in a wide range of NLP tasks, including Hidden Markov Models (HMMs) (Baum and Eagon, 1967; Rabiner, 1989), latent-variable probabilistic context-free grammars (L-PCFGs) (Matsuzaki et al., 2005; Petrov et al., 2006), IBM alignment models (Brown et al., 1993), Brown clustering (Brown et al., 1992), etc.

The latent variables increase the capacity of the models and improve their expressiveness. However, learning the latent variables is usually an intractable problem, because they cannot be directly learned from the observable data. Some heuristic methods are introduced to solve this problem. Expectation Maximization (EM) (Dempster et al., 1977) is a typically method to learn parameters of latent-variable models, which involves two steps after all parameters are randomly initialized at the beginning. The expectation step computes the distributions of the latent variables given the observable variables and the fixed parameters. In the maximization step, the parameters are re-estimated by fixing the distributions just obtained. These two steps are iterated until convergence. i.e. the change of parameters for each iteration is small enough. Estimating parameters for models involving latent variables is a non-convex optimization problem, so these black-box heuristic methods are not theoretically guaranteed to converge to the global optima.

Spectral learning is introduced to solve the optimization problem based on singular value decomposition (SVD). The general first step is to perform SVD on observable variables learned from training examples through the method of moments, i.e. estimating population moments by sample moments. Then singular matrices and values are used to compute observable representations, which are transformations of original parameter matrices and sufficient for model inference. Under certain conditions on parameter matrices and sample complexity, estimation accuracy can be guaranteed in PAC framework (Valiant, 1984). Intuitively, the latent variables often reflect the correlations of the observable variables or structures, which can also be reflected by the singular matrices and values for the observable samples. Compared with the EM algorithm, spectral learning guarantees consistent parameter estimates, which are exempted from local optima, and is much more efficient because of no iterations.

2 Applications

Spectral learning algorithms are widely used in natural language processing, including Hidden Makov Models (HMMs) (Hsu et al., 2008), finite state transducers (Balle et al., 2011), split-head automaton grammers (Luque et al., 2012), reduced rank HMMs (Siddiqi et al., 2010), kernel-based methods for HMMs (Song et al., 2010), refinement HMMs (Stratos et al., 2013), tree graphical models (Parikh et al., 2011; Song et al., 2011), weighted tree automata (Bailly et al., 2010), latent-variable probabilistic context-free grammars (L-PCFGs) (Cohen et al., 2012, 2013), latent Dirichlet allocation (LDA) (Anandkumar et al., 2012), etc. This summary is based on Cohen et al. (2014).

This section reviews the applications of spectral learning algorithms for HMMs and L-PCFGs.

2.1 Hidden Markov Models

Hidden Markov Models, statistical models for discrete time series, have discrete latent variables that cannot be directly learned from data. The EM algorithm, as a traditional method, is used to learn the parameters in HMMs. As discussed before, EM cannot guarantee consistent estimates and easily falls into local optima.

Hsu et al. (2008) proposed a spectral algorithm to learn the parameters in Hidden Markov Models. In his paper, SVD is performed on the matrix of the joint probabilities of the nearest word pairs. The singular matrices and values, involving the correlation information of the nearest words, are used along with the joint probabilities of the previous and later words for each word to compute the proxy parameters for the HMMs. The probabilities are estimated from the observations. These proxy parameters are actually the transformation of the original ones. The original parameters do not have to be recovered because the proxy parameters are sufficient to inference the HMMs. Conditioned on the singular values, the estimation error is guaranteed to be small enough.

The spectral algorithm is only applicable for discrete observations, while in reality many observable variables are continuous or structured. Song et al. (2010) extended discrete observation spaces for HMMs to Hilbert spaces by Hilbert space embedding (Smola et al., 2007; Sriperumbudur et al., 2008), and proposed a kernelized spectral algorithm to learn the parameters. Stratos et al. (2013) defined the refinement HMMs that successful solve supervised sequence labelling tasks. In addition to the observations and their labels, latent variables are involved in the refinement HMMs. The parameter estimation task for the refinement HMMs can also be solved by a specific spectral algorithm.

2.2 Latent-Variable PCFGs

Latent-variable PCFGs (L-PCFGs) (Matsuzaki et al., 2005; Petrov et al., 2006) are designed for natural language parsing, where each node of skeletal trees is decorated with a latent state. EM is previously used to learn the parameters in L-PCFGS (Matsuzaki et al., 2005). Alternatively, Cohen et al. (2012) proposed a spectral algorithm for the learning task, which is depend on the inside-outside algorithm. The correlations of the inside and outside trees for every node reveals the associated latent state. Thus, the specific mapping functions are selected to map the inside and outside trees for every node to feature vectors, making up a co-occurrence matrix. Then this matrix is approximately decomposed to two low-rank singular matrices and the associated singular values that transform the feature vectors to low dimensions. The observable representations, which is the transformation of original parameters, are computed by these low-dimension feature vectors along with the observations. The models with these observable representations can be used to parse sentences. Cohen et al. (2013) carried out experiments on spectral learning for L-PCFGs, who considered the choices of mapping functions, smooth estimation methods and handling negative values. Spectral learning can achieve similar accuracy with the EM algorithm within a much short time, while may lag behind other state-of-the-art algorithms. Narayan and Cohen (2015) used a clustering algorithm to perform spectral learning, where the latent state of each node is clustered to a constant through k-means clustering. Decorated with the latent state, the probabilities can be estimated directly from the data. They added noises to the underlying features and trained a set of models which are decoded to improve the accuracy. Narayan and Cohen (2016) globally optimized the number of latent states by the coarse-to-fine techniques (Petrov et al., 2006) and improved the accuracy of L-PCFGs.

3 Discussion

Spectral learning guarantees consistent parameters and makes the learning process more efficient than the EM algorithm. However, spectral learning cannot recover the original parameters or probabilities (spectral clustering Narayan and Cohen (2015) can do this but break the consistent guarantee due to k-means clustering). There are several directions that worth exploring in the future.

1. Apply spectral learning on other latent-variable models where the EM algorithms are effective, e.g. IBM alignment models (Brown et al., 1993). However, the specific spectral algorithms may have to be further optimized. As discussed before, Cohen et al. (2013) considered many practical problems in applying spectral learning on L-PCFGs. The things to be considered may include feature selection and scaling, the number of latent variables, model regularization, etc.

2. Optimize the existing spectral learning algorithms to improve the accuracy. For example, to regularize L-PCFGs, Cohen et al. (2013) used a smoothing method, and Narayan and Cohen (2015) added noises to features and combined diverse models. More regularization methods might be borrowed from the deep learning area, such as batch normalization.

3. Extend spectral learning algorithms to semi-supervised learning or active learning (Settles, 2009) tasks. Labelling data is usually a troublesome and expensive work. Given an acceptable error, the PAC-style framework of spectral learning could give a sample size that guarantees this error. We only have to label samples up to this size, and use some strategies to add the unlabelled data into the labelled data set.

4. Generalize spectral learning algorithms to general non-convex optimization problems. Anandkumar et al. (2014) and Chaganty and Liang (2014) used the method of moments and tensor factorization on parameter estimates for a broader class of latent-variable models. Janzamin et al. (2015) combined tensor decomposition with neural networks.

References

Anandkumar, A., Foster, D. P., Hsu, D. J., Kakade, S. M., and kai Liu, Y. (2012). A spectral algorithm for latent dirichlet allocation. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 917–925. Curran Associates, Inc.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832.

Bailly, R., Habrard, A., and Denis, F. (2010). A spectral approach for probabilistic grammatical inference on trees. In Hutter, M., Stephan, F., Vovk, V., and Zeugmann, T., editors, *Algorithmic Learning Theory*, pages 74–88, Berlin, Heidelberg. Springer Berlin Heidelberg.

Balle, B., Quattoni, A., and Carreras, X. (2011). A spectral learning algorithm for finite state transducers. In Gunopulos, D., Hofmann, T., Malerba, D., and Vazirgiannis, M., editors, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011. Proceedings, Part I*, volume 6911 of *Lecture Notes in Computer Science*, pages 156–171. Springer.

Baum, L. E. and Eagon, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. *Bull. Amer. Math. Soc.*, 73(3):360–363.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18:467–479.

Chaganty, A. T. and Liang, P. (2014). Estimating latent-variable graphical models using moments and likelihoods. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1872–1880, Bejing, China. PMLR.

Cohen, S. B., Stratos, K., Collins, M., Foster, D. P., and Ungar, L. (2012). Spectral learning of latent-variable PCFGs. In *Proceedings of ACL*.

Cohen, S. B., Stratos, K., Collins, M., Foster, D. P., and Ungar, L. (2013). Experiments with spectral learning of latent-variable PCFGs. In *Proceedings of NAACL*.

Cohen, S. B., Stratos, K., Collins, M., Foster, D. P., and Ungar, L. (2014). Spectral learning of latent-variable PCFGs: Algorithms and sample complexity. *Journal of Machine Learning Research*.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38.

Hsu, D. J., Kakade, S. M., and Zhang, T. (2008). A spectral algorithm for learning hidden markov models. *CoRR*, abs/0811.4413.

Janzamin, M., Sedghi, H., and Anandkumar, A. (2015). Generalization bounds for neural networks through tensor factorization. *CoRR*, abs/1506.08473.

Luque, F. M., Quattoni, A., Balle, B., and Carreras, X. (2012). Spectral learning for non-deterministic dependency parsing. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 409–419, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matsuzaki, T., Miyao, Y., and Tsujii, J. (2005). Probabilistic cfg with latent annotations. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 75–82. Association for Computational Linguistics.

Narayan, S. and Cohen, S. B. (2015). Diversity in spectral learning for natural language parsing. In *Proceedings of EMNLP*.

Narayan, S. and Cohen, S. B. (2016). Optimizing spectral learning for parsing. In *Proceedings of ACL*.

Parikh, A. P., Song, L., and Xing, E. P. (2011). A spectral algorithm for latent tree graphical models. In *ICML*, pages 1065–1072. Omnipress.

Petrov, S., Barrett, L., Thibaux, R., and Klein, D. (2006). Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 433–440, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Settles, B. (2009). Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Siddiqi, S., Boots, B., and Gordon, G. (2010). Reduced-rank hidden markov models. In Teh, Y. W. and Titterington, M., editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 741–748, Chia Laguna Resort, Sardinia, Italy. PMLR.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007). A hilbert space embedding for distributions. In Hutter, M., Servedio, R. A., and Takimoto, E., editors, *Algorithmic Learning Theory*, pages 13–31, Berlin, Heidelberg. Springer Berlin Heidelberg.

Song, L., Boots, B., Siddiqi, S. M., Gordon, G., and Smola, A. (2010). Hilbert space embeddings of hidden markov models. In *Proceedings of ICML*.

Song, L., Xing, E. P., and Parikh, A. P. (2011). Kernel embeddings of latent tree graphical models. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 24*, pages 2708–2716. Curran Associates, Inc.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Lanckriet, G. R. G., and Schölkopf, B. (2008). Injective hilbert space embeddings of probability measures. In *COLT*.

Stratos, K., Rush, A. M., Cohen, S. B., and Collins, M. (2013). Spectral learning of refinement HMMs. In *Proceedings of CoNLL*.

Valiant, L. G. (1984). A theory of the learnable. *Commun. ACM*, 27(11):1134–1142.