Bayesian Nonparametric Structured Models

Percy Liang
Computer Science Division
University of California at Berkeley
Berkeley, CA 94720
pliang@cs.berkeley.edu

Dan Klein
Computer Science Division
University of California at Berkeley
Berkeley, CA 94720
klein@cs.berkeley.edu

Abstract

Probabilistic modeling is a dominant approach for both supervised and unsupervised learning in NLP. One constant challenge for models with latent variables is determining the appropriate model complexity, i.e. the question of "how many clusters." While cross-validation can be used to select between a limited number of options, it cannot be feasibly applied in the context of larger hierarchical models where we must balance complexity in many parts of the model at the same time. Nonparametric "infinite" priors such as the *Dirichlet process* are tools from the Bayesian statistics literature which present an elegant solution to this problem and have seen increasing use in recent NLP work. Models based on the Dirichlet process have an infinite number of clusters and rely on the prior to penalize their use, thus allowing the complexity of the model to adapt to the data.

In explaining how to do inference in these new models, we try to dispel two myths: first, that Bayesian methods are too slow and cumbersome, and, second, that Bayesian techniques require a whole new set of algorithmic ideas. We depart from the traditional sampling methodology which has dominated past expositions and focus on *variational inference*, an efficient technique which is a natural extension of EM. This approach allows us to tackle structured models such as HMMs and PCFGs with the benefits of Bayesian nonparametrics while maintaining much of the existing EM machinery so familiar to this community. We provide intuition about the Dirichlet process, variational inference, and their interaction at several different levels. In addition to our foundational presentation, we discuss concrete implementation issues, and demonstrate the empirical properties of these methods on several different models.

Presenter Biography

Percy Liang is a Ph.D. student in computer science at UC Berkeley. He has a BS in math and a BS/MS in computer science from MIT. His research interests include probabilistic modeling for semi-supervised learning in NLP, especially using Bayesian nonparametrics, and approximate inference algorithms for such models. He holds an NSF Graduate Fellowship and a National Defense Science and Engineering Graduate Fellowship.

Dan Klein is an assistant professor of computer science at the University of California, Berkeley (PhD Stanford, MS Oxford, BA Cornell). Professor Klein's research focuses on statistical natural language processing, including unsupervised methods, syntactic parsing, and machine translation. His academic honors include a British Marshall Fellowship, an inaugural Microsoft New Faculty Fellowship, and best paper awards at the ACL, NAACL, and EMNLP conferences.