2. Beyond EM: Bayesian Techniques for Human Language Technology Researchers

Hal Daume III, USC-ISI

The Expectation-Maximization (EM) algorithm has proved to be a great and useful technique for unsupervised learning problems in natural language, but, unfortunately, its range of applications is largely limited by intractable E- or M-steps, and its reliance on the maximum likelihood estimator. The natural language processing community typically resorts to ad-hoc approximation methods to get (some reduced form of) EM to apply to our tasks. However, many of the problems that plague EM can be solved with Bayesian methods, which are theoretically more well justified. This tutorial will cover Bayesian methods as they can be used in natural language processing. The two primary foci of this tutorial are specifying prior distributions and performing the necessary computations to perform inference in Bayesian models. The focus of the tutorial will be primarily on unsupervised techniques (for which EM is the obvious choice). Supervised and discriminative techniques will also be mentioned at the conclusion of the tutorial, and pointers to relevant literature will be provided.

2.1 Tutorial Outline

1. Introduction to the Bayesian Paradigm
2. Background Material
   - Graphical Models (naive Bayes, maximum entropy, HMMs)
   - Expectation Maximization
   - Non-Bayesian Inference Techniques
3. Common Statistical Distributions
   - Uniform
   - Binomial and Multinomial
   - Beta and Dirichlet
   - Poisson, Gaussian and Gamma
4. Simple Bayesian Inference Techniques
   - Inference = Integration
   - Integration by Summing
   - Monte Carlo Integration
5. Advanced Bayesian Inference Techniques
   - Markov Chain Monte Carlo Integration
   - Laplace Approximation
   - Variational Approximation
   - Others (Message Passing Algorithms)
6. Survey of Popular Models
   - Latent Dirichlet Allocation
   - Integrating Topics and Syntax
   - Matching Words and Pictures
7. Pointers to Literature on Other Topics
8. Conclusions

2.2 Target Audience

This tutorial should be accessible to anyone with a basic understanding of statistics (familiarity with EM would help, but is not necessary). I use a query-focused summarization task as a motivating running example for the tutorial, which should be of interest to researchers in natural language processing and in information retrieval.

Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume, pages 301–302,
New York City, June 2006.

Hal's research interests lie at the intersection of machine learning and natural language processing. He works primarily on problems in automatic document summarization and information extraction, using a variety of machine learning techniques. As a Bayesian, he has successfully applied variational inference and expectation propagation techniques to unsupervised learning problems in summarization. He has also successfully applied nonparametric infinite Bayesian models to problems in supervised clustering. In December 2005, he co-organized (with Yee Whye Teh, National University of Singapore) a workshop on "Bayesian Methods for NLP" at the Conference for Neural Information Processing Systems.