Donglai Wei and Erik Sudderth
Department of Computer Science
Brown University
Providence, RI 02912
{donglai_wei,sudderth}@cs.brown.edu

Proposal: ME algorithm for Hierarchical Dirichilet Process and its Extensions

Anonymous Author(s)
Affiliation
Address
email

Abstract

We here develop a Maximize-Expectation(ME) learning algorithm for Hierarchical Dirichilet Process(HDP). We first show the advantage of ME algorithm to learn Dirichlet Process Mixture model with synthetic data. Then we use Chinese Restaurant Franchise(CRF) representation to construct ME algorithm for HDP. Later, we extend the ME algorithm to learn Hierarchical Pitman-Yor Process(HPY) model, which generalizes HDP with more flexiblility. Implementing our novel learning method, hopefully we will improve the learning performance of HPY model for shared segmentation problem.

1 BackGround

1.1 Hierarchical Dirichilet Process

During clustering, we may not only want to separate observations into different groups but also wish these groups to share common features. For example, in document modeling, the aim is to cluster words within the documents into different topics. When clustering documents from NIPS in machine learning and computer vision, we may wish to allow topics like “graphical model ”and “optimization”to be shared among them.

Hierarchical Dirichilet Process(HDP), which natually handles the problem above, was formally introduced into unsupervised learning in Teh at.el[2]. Figure 1 shows the graphical model for DP mixture and HDP. But due to the complexity, the extant learning methods developed for HDP are far from maturity.

1.2 ME Algorithm

Consider a probabilistic model P(x,w,α),where x is observed random variable, w hidden variable and α hyperparameter. Given a data set D, a typical task in machine learning is to maximize the likelihood function P(D $|\alpha$) by marginaling out hidden variable w. But the exact leariningis often intractable. Variational Bayesian algorithms can give a reasonable lower bound for the likelihood by approximating the true distribution of hidden variables P (w | D) based on Kullback-Leibler (KL) divergence. For our clustering problems, hidden variables are divided into two classes: cluster assignment variable z and model parameters θ. Heuristically, we factorize the distribution P(w|D,α)=P(θ, z| D,α) $\approx$ q (z)q (θ) and update the estimated distribution for one hidden variable at a time. This results in the update:

$$q(\theta) \propto exp(E[logP(\theta, z, D)]_{q(z)}) \longleftrightarrow q(z) \propto exp(E[logP(\theta, z, D)]_{q(\theta)}).$$

This is the well-known Meanfield algorithm, which maintains a disribution over parameters (known as E-step). Also, we can decide to estimate a maximum a posterior(MAP) value for the hidden variable (known as M-step). Discussed in [4], we can have four combinations of E-step and M-step for z and θ. In this way, K-means belong to MM algorithm while Meanfield belongs to EE algorithm. But for most of the time, people may just want the optimal solution for the cluster assignment z instead of the real distribution of q(z). Also, the cluster assignment variable z is discrete and high dimensional, which makes the update formula hard to compute. So instead of maintaining the huge matrix for q(z), we may just pick out the MAP estimator(M-step) for q(z). Since we don't want to be too greedy to also take M-step for q(θ), we end up with ME algorithm. In the next section, we are going to show the advantage of ME algorithm for learning DP mixture model.

2 Preliminary Result

We test Meanfield algorithm (EE) [1], Collapsed Meanfield algorithm(Collapsed EE)[5] and ME algorithm [3] implemented by Kurihara on the synthetic data, 200 random samples drawn from Gaussian of four mixtures. We set almost same hyperparameters(except the one to generate mixture components) for these three algorithms and test with different initial cluster numbers. Results are summarized in table 1. From figure 3 and 4, we can see that to some degree both EE and Collapsed EE suffer from local minimas and initialization. Three variants of ME algorithms are tested here. Though the hierarchical clustering methods work perfect on this synthetic data, for real data top-down method uses heuristic criterion to split and merge while bottom-up method may have difficulty maintaining the huge distance matrix. We here only try a naive local search. Though it does not perform well enough at first glance, it can be improved with advanced search algorithms and it fits best into Chinese Restaurant Franchise representation for Constructing HDP.

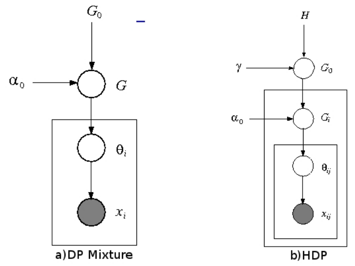

Figure 1: Graphical Model for DP mixture and HDP

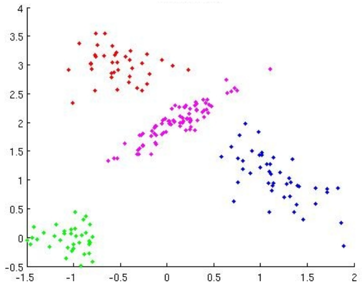

Figure 2: Ground Truth 200 data from Gaussian Mixture

3 Research Plan

– 3.1: **Survey** the literature on **DP mixture** and **HDP**
– 3.25: **Test** ME, EE, Collapsed EE algorithms for **DP Mixture** on a synthetic data set.
– 4.15: **Develop** ME algorithm for **HDP**
– 4.25: **Test** ME algorithm against Collapsed EE and Gibbs sampling for **HDP** model
– 5.1: **Test** ME algorithm for learning **HPY** model for shared segmentation problems.
– 5.10: **Write** documentation and technical report

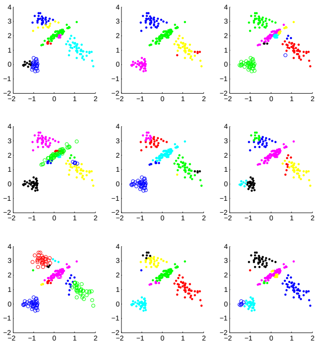

Figure 3: EE with random initialization, 20 inital clusters

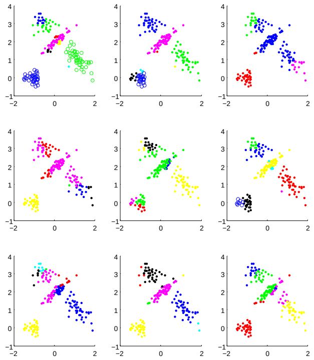

Figure 4: Collapsed EE with random initialization, 20 inital clusters

Table 1: Comparison of EE, Collapsed-EE and ME algorithms for DP mixture(with mean and std)

Learning Algorithm	**Initial Cluster Numbers**	**Number of clusters**	**RandIndex**
EE(sampled initialization)	10	4.1(0.3)	0.99(0.00)
	20	4.1(0.3)	0.99(0.00)
EE(random initialization)	10	6.9(1.2)	0.97(0.01)
	20	7.6(1.7)	0.93(0.02)
Collapsed EE	10	5.6(1.6)	0.90(0.22)
	20	7.4(0.9)	0.87(0.09)
ME	1(Top-down)	4	1
	200(Bottom-up)	4	1
	200(Local search)	15.5(3.9)	0.83(0.09)
	200(Local search+Merge)	6.5(0.9)	0.93(0.09)

4 References

[1] Blei,D,M, Jordan,M.I, and Ng, A. Y. (2003), Hierarchical Bayesian Models for Applications in Information Retrieval, in Bayesian Statistics, vol. 7, pp. 2544

[2] Ghahramani, Z., & Beal, M. J. (2000). Variational inference for Bayesian mixtures of factor analysers. In S. A. Solla, T. K. Leen, & K.-R. Muller (Eds.), Advances in neural information processing systems, 12. Cambridge, MA: MIT Press.

[3] K. Kurihara and M. Welling. Bayesian K-means as a maximization-expectation algorithm. Neural Computation, 2008

[4] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. Journal of the American Statistical Association, 101(476):15661581, 2006.

[5] Y. W. Teh, K. Kurihara and M. Welling. Collapsed Variational Inference for HDP NIPS 2007