A Convex Exemplar-based Approach to MAD-Bayes Dirichlet Process Mixture Models

Ian E.H. Yen1, Xin Lin1, Kai Zhong2, Pradeep Ravikumar1 and Inderjit S. Dhillon1

1Computer Science Dept., 2Institute for Computational Engineering and Sciences, University of Texas at Austin

Abstract

- MAD-Bayes: estimating Bayesian non-parametric models via combinatorial optimization; caveat: computationally expensive to find global optimum.
- We consider an exemplar-based formulation of MAD-Bayes (for Dirichlet Processes (DP) and Hierarchical DP (HPD)).
- We then consider corresponding "convex relaxations" that, under separation condition, recover Integer optimal solutions.
- We also provide a simple ADMM based algorithm to solve the convex relaxations.

MAD-Bayes Dirichlet Process Mixture

- MAD-Bayes derives the asymptotic log-likelihood of DP mixture when variance $\sigma^2$ of dist. $P(x|\mu_k)$ approaches 0.

$\begin{align*}
D_{\text{KL}}(\pi, \mu_0) := \int \log \left( \frac{\pi(z)}{\mu_0(z)} \right) \pi(z) dz
\end{align*}$

where $D_{\text{KL}}(\pi, \mu_0)$ is the Kullback-Leibler divergence.

The approach extends to MAD-Bayes: estimating Bayesian non-parametric models via Exemplar-based modeling.

- The program becomes convex when relaxing $w_{ij}$ from $(0, 1]$ to $[0, 1]$.

Convex Exemplar-based Approach: Hierarchical DPM

For HDP:

$\begin{align*}
&\sum_{d=1}^{D} \sum_{i=1}^{K_d} W_{dij} z_{dij} + \theta \sum_{d=1}^{D} K_d + \lambda K
\end{align*}$

is equivalent to

$\begin{align*}
&\min_{W, \theta, \lambda} \frac{1}{2} \| W \|_F^2 + \lambda \| W \|_{\infty, 1}
\end{align*}$

subject to $W_{ij} = 0$ for $i, j = 1, \ldots, K_i$.

assuming there is an Exemplar Set $E = \{\mu_j\}$ of candidate centroids.

The program becomes convex when relaxing $w_{ij}$ from $(0, 1]$ to $[0, 1]$.

Theorem (Optimality Guarantee)

Suppose there exists a clustering $\{S_k\}_{k=1}^K$ for which we can find $\lambda$ such that

$\begin{align*}
\max \min_{\lambda} \lambda \leq \min_{\lambda} \lambda
\end{align*}$

Then $W^* \{\mu_j\}$ is optimal solution to the convex DPM program.

Optimality Guarantee: Hierarchical DP

- Similar guarantee can be obtained for hanging formulation, where clusters that do not share a data group can have smaller separation requirement.

Optimality Guarantee: Hierarchical DPM

- According to the theorem, the larger extent of separation a clustering has, the wider range of $\lambda$ producing that clustering.

Optimality Guarantee: Hierarchical DPM (HPD)

- For HPD:

$\begin{align*}
&\sum_{d=1}^{D} \sum_{i=1}^{K_d} W_{dij} z_{dij} + \theta \sum_{d=1}^{D} K_d + \lambda K
\end{align*}$

is equivalent to

$\begin{align*}
&\min_{W, \theta, \lambda} \frac{1}{2} \| W \|_F^2 + \lambda \| W \|_{\infty, 1}
\end{align*}$

subject to $W_{ij} = 0$ for $i, j = 1, \ldots, K_i$.

assuming there is an Exemplar Set $E = \{\mu_j\}$ of candidates centroids.

The program becomes convex when we relax $w_{ij}$ from $(0, 1]$ to $[0, 1]$.

An ADMM solver for Structural-Regularized Program

$\begin{align*}
&\min_{W, \theta, \lambda} \frac{1}{2} \| W \|_F^2 + \lambda \| W \|_{\infty, 1}
\end{align*}$

subject to $W_{ij} = 0$ for $i, j = 1, \ldots, K_i$.

and (ii) sub-problem with only structural regularizer

$\begin{align*}
&\min_{W, \theta, \lambda} \frac{1}{2} \| W \|_F^2 + \lambda \| W \|_{\infty, 1}
\end{align*}$

Since the quadratic and linear terms are separable w.r.t. coordinates, the first problem can be solved via simple projection, while the second problem can be solved via proximal mapping of the Group Norm. Then it updates $Z^{(t+1)} = (W^{(t+1)} + W^{(t-1)})/2$, $Y^{(t+1)} = \alpha(W^{(t+1)} - Z^{(t+1)})$.

The procedure has linear convergence to the optimum of convex program.

Results

The MAD-Bayes objective function achieved by different approaches.

Future Works

- Extend our analysis to other exemplar-based unsupervised learning problems such as Multiple Sequence Alignment and Sequence Motif Finding, which can be seen as the exemplar-based version of Hidden Markov Models.
- Develop a greedy Column Generation method that reduces computational complexity of the current algorithm from $O(N^2)$ to $O(NK)$.

Reference

