Robust Bayesian Max-Margin Clustering Discussion

Changyou Chen, Jun Zhu and Xinhua Zhang

Discussion by: R. Henao

Duke University

November 21, 2014
Motivation:

- Loss-function based clustering methods are excellent performance-wise.
- In loss-function based clustering methods estimating the number of clusters is challenging.
- In loss-function based clustering methods is difficult to impose complex distributions to mixture components.
- Adding constraints to Bayesian clustering is difficult.

Proposed model: Bayesian max-margin clustering (BMC)

- Fully Bayesian specification.
- Probabilistic inference of the number of clusters.
- First extension of RegBayes to unsupervised clustering.
- Efficient MCMC.
- 2 flavors:
  - Dirichlet Process Max-Margin Gaussian Mixture (DPMMGM).
  - Max-Margin Clustering Topic Model (MMCTM).
Posterior inference via Bayes theorem is equivalent to solving the following optimization problem:

$$\inf_{q(\Theta) \in P} \ KL(q(\Theta) \| p(\Theta)) - \mathbb{E}_{\Theta \sim q(\Theta)} [\log p(X|\Theta)] .$$

where $\Theta$ are latent variables, $X$ is a dataset, $p(X|\Theta)$ is the likelihood and $P$ is the space of probability distributions.

We can shape $P$ by imposing constraints on $q(\Theta)$. If constraints are parameterized via auxiliary (slack) variables $\xi$ we have

$$\inf_{\xi, q(\Theta) \in P} \ KL(q(\Theta) \| p(\Theta)) - \mathbb{E}_{\Theta \sim q(\Theta)} [\log p(X|\Theta)] + U(\xi) .$$

s.t. $q(\Theta) \in P(\xi)$.

where $U(\cdot)$ is a penalization function.

The optimal $q(\Theta)$ is called post-data posterior\(^1\).

\(^1\)J. Zhu et al, 2014 (JMLR)
Robust Bayesian Max-Margin Clustering

Assume \( x_i \in \mathbb{R}^p \), \( \eta_k \in \mathbb{R}^p \) is a latent projector for cluster \( k \) and \( \eta_k \in \Theta \). Define the compatibility score of \( x_i \) w.r.t. cluster \( k \) as

\[
F_k(x_i) = \mathbb{E}_{q(\Theta)}[\eta_k^\top x_i].
\]

For each \( x_i \), let \( y_i \in \mathbb{Z}^+ \) be a random variable denoting a cluster assignment for \( x_i \) and \( y_i \in \Theta \). BMC amounts to solve

\[
\inf_{\xi_i \geq 0, q(\Theta)} \mathcal{L}(q(\Theta)) + 2c \sum_i \xi_i \\
\text{s.t. } F_{y_i}(x_i) - F_k(x_i) \geq \ell \mathbb{1}(y_i \neq k) - \xi_i, \forall i, k,
\]

where

\[
\mathcal{L}(q(\Theta)) = KL(q(\Theta) || p(\Theta)) - \mathbb{E}_{\Theta \sim q(\Theta)}[\log p(\mathbf{X}|\Theta)].
\]
Robust Bayesian Max-Margin Clustering

The constrained optimization problem is equivalent to

$$\inf_{q(\Theta)} L(q(\Theta)) + 2c \sum_i \max(0, \max_{k: k \neq y_i} E_{\Theta \sim q(\Theta)}[\zeta_{ik}]),$$  

(1)

where $\zeta_{ik} := \ell(y_i \neq k) - (\eta_{y_i} - \eta_k)^T x_i$.

The objective in (1) can be upper bounded by

$$\inf_{q(\Theta)} L(q(\Theta)) + 2c \sum_i E_{\Theta \sim q(\Theta)}[\max(0, \max_{k: k \neq y_i} \zeta_{ik})].$$  

(2)

Introducing auxiliary variables $s_i := \arg\max_{k: k \neq y_i} \zeta_{ik}$ and applying standard derivations in calculus of variations, an analytic form of the optimal solution of (2) can be written as

$$q(\Theta, \{s_i\}) \propto p(\Theta|X) \prod_i \exp(-2c\max(0, \zeta_{i s_i})).$$

Furthermore

$$q(\Theta, \{s_i\}, \{\lambda_i\}) \propto p(\Theta|X) \prod_i \tilde{\phi}_i(\lambda_i|\Theta),$$

where $\tilde{\phi}_i(\lambda_i|\Theta) := \lambda_i^{-1/2} \exp(-\frac{1}{2\lambda_i}(\lambda_i + c\zeta_{is_i})^2)$. 
A quick summary:

- Model parameters: $\Theta = \{\Theta_k, \eta_k, y_i, \}$.
- Auxiliary variables: $\{s_i\}$ and $\{\lambda_i\}$.
- Free parameters: $c$ and $\ell$.

where $\Theta_k$ denotes the parameters of each mixture component.

Inference: MCMC via Gibbs sampling.
Let \( p(X|\Theta) \) be a Gaussian mixture model so \( \Theta_k = \{\mu_k, \Lambda_k\} \) and

\[
\begin{align*}
\mu_k &\sim \mathcal{N}(\mu_k; m, (r\Lambda_k)^{-1}), \\
\Lambda_k &\sim \mathcal{IW}(\Lambda_k; S, \nu), \\
y_i &\sim \text{DP}(\alpha), \\
\eta_k &\sim \mathcal{N}(0, vI).
\end{align*}
\]

Note that \( p(x_i|\Theta) = \mathcal{N}(x_i; \mu_{y_i}, (r\Lambda_{y_i})^{-1}) \) is independent of \( \eta_k \).

Note also that the post-data posterior includes \( \eta_k \) via \( \tilde{\phi}_i(\lambda_i|\Theta) \).

The hyperparameters of the model now include \( \{m, r, S, \nu, \alpha, v\} \).
The DPMMGM can be seen as a DP mixture with *Normal-inverse Wishart-Normal* base measure and pseudo-likelihood proportional to

\[
f(x_i, \lambda_i | y_i, \mu_{y_i}, \Lambda_{y_i}, \{\eta_k\}) := \mathcal{N}(x_i; \mu_{y_i}, (r\Lambda_{y_i})^{-1}) \tilde{\phi}_i(\lambda | \Theta).
\]

DPMMGM has the following generative process

\[
(\mu_k, \Lambda_k, \eta_k) \sim \mathcal{N}(\mu_k; m, (r\Lambda_k)^{-1}) \times \mathcal{IW}(\Lambda_k; S, \nu) \times \mathcal{N}(\eta; 0, vI),
\]

\[
w \sim \text{SBP}(\alpha),
\]

\[
y_i | w \sim \text{Discrete}(w),
\]

\[
(x_i, \lambda_i | y_i, \{\mu_k, \Lambda_k, \eta_k\}) \sim f(x_i, \lambda_i | y_i, \mu_{y_i}, \Lambda_{y_i}, \{\eta_k\}).
\]

Posterior inference: Gibbs sampling

- Marginalize out \(w\).
- Sample \(\{\mu_k, \Lambda_k, \eta_k\} \cup \{y_i, s_i, \lambda_i\}\).
- Jointly sample \(\{y_i, s_i\}\) using *Reused Algorithm*\(^2\).
- Complexity is roughly \(O(\kappa_1 p^3 + \kappa_2 K^3 + \kappa_3 NK^2)\).  

\(^2\)Favaro and Teh, 2013 (Stat. Sci.)
Max-Margin Clustering Topic Model

Extend LDA by introducing a cluster label, and define each cluster as a mixture of topic distributions.

Let $V$ be the size of the vocabulary, $T$ the number of topics and $K$ the number of clusters. A generative model for cluster-based topic model (CTM) is

- For each topic $t$, generate word distribution $\phi_t|\beta \sim \text{Dirichlet}(\beta 1_V)$.
- Draw a base topic distribution $\mu_0|\alpha_1 \sim \text{Dirichlet}(\alpha_0 1_T)$.
- For each cluster $k$, generate topic distribution $\mu_k|\alpha_1 \sim \text{Dirichlet}(\alpha_1 \mu_0)$.
- Draw a base cluster distribution $\gamma|\omega \sim \text{Dirichlet}(\omega 1_K)$.
- For each document $i$:
  - Generate cluster label $y_i|\gamma \sim \text{Discrete}(\gamma)$.
  - Generate topic distribution $\mu_i|\alpha, \mu_{y_i} \sim \text{Dirichlet}(\alpha \mu_{y_i})$.
  - Generate word assignments $z_{il} \sim \text{Discrete}(\mu_i)$.
  - Generate words $w_{il} \sim \text{Discrete}(\phi_{z_{il}})$.

Model parameters:
\[ \Theta = \{ \phi_t \} \cup \{ \eta_k, \mu_k \} \cup \{ \mu_0, \gamma \} \cup \{ \mu_i, y_i \} \cup \{ z_{il} \}, \text{ for } t = 1, \ldots, T, \]
\[ k = 1, \ldots, K, \ i = 1, \ldots, D \text{ and } l = 1, \ldots, N_i. \]
Max-Margin Clustering Topic Model

Since the raw word space is high-dimensional and sparse, clustering structure is characterized by empirical latent topic distributions, i.e. document $i$ is summarized as $\mathbf{x}_i \in \mathbb{R}^T$, where $x_{ti} = \frac{1}{N_i} \sum_l \mathbb{I}(z_{il} = t)$.

The compatibility score is again $F_k(\mathbf{x}_i) = \mathbb{E}_{q(\Theta)}[\eta_k^\top \mathbf{x}_i]$, however, $x_i$ is not observed.

Posterior inference: Gibbs sampling

- Marginalize out $\{\phi_t\} \cup \{\mu_k\} \cup \{\mu_0, \gamma\} \cup \{\mu_i\}$.
- Sample $\{\eta_k\} \cup \{y_i\} \cup \{z_{il}\}$.

Dealing with vacuous solutions:

The problem: max-margin constraints do not interact with observed data (words), thus making $z_{il}$ easily collapsing into a single cluster.

One solution: add cluster balance constraints (expensive and does not work well in practice).

Alternative solution (semisupervised setting): assign a few documents to their true label (landmarks).
Experiments: DPMMGM

Heuristic approach for model selection:

The problem: free parameters $c$ and $\ell$ are critical but difficult to select without training data. Besides, not feasible via Bayesian sampling because they are not parameters from a proper Bayesian model.

Heuristic 1: assume $K$ is known. Initialize $c = \ell = 0.1$. At each iteration compare inferred $K$ with ground truth, if larger, increase $c$ and $\ell$ by $u/n$, where $u \sim \text{Uniform}(0,1)$ and $n$ is the current iteration (DPMMGM).

Heuristic 2: estimate $K$ via DPGMM, then use heuristic 1 (DPMMGM*).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Data property</th>
<th>NMI</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n$ $p$ $K$</td>
<td>kmeans</td>
<td>nCut</td>
<td>DPGMM</td>
<td>DPGMM</td>
<td>DPGMM*</td>
</tr>
<tr>
<td>Glass</td>
<td>214 10 7</td>
<td>0.37±0.04</td>
<td>0.22±0.00</td>
<td>0.37±0.05</td>
<td><strong>0.46±0.01</strong></td>
<td>0.45±0.01</td>
</tr>
<tr>
<td>Half_circle</td>
<td>300 2 2</td>
<td>0.43±0.00</td>
<td>1.00±0.00</td>
<td>0.49±0.02</td>
<td>0.67±0.02</td>
<td>0.51±0.07</td>
</tr>
<tr>
<td>Iris</td>
<td>150 4 3</td>
<td>0.72±0.08</td>
<td>0.61±0.00</td>
<td>0.73±0.00</td>
<td><strong>0.73±0.00</strong></td>
<td>0.73±0.00</td>
</tr>
<tr>
<td>Letter</td>
<td>1000 16 10</td>
<td>0.33±0.01</td>
<td>0.04±0.00</td>
<td>0.19±0.09</td>
<td><strong>0.38±0.04</strong></td>
<td>0.23±0.04</td>
</tr>
<tr>
<td>MNIST</td>
<td>1000 784 10</td>
<td>0.50±0.01</td>
<td>0.38±0.00</td>
<td>0.55±0.03</td>
<td><strong>0.56±0.01</strong></td>
<td>0.55±0.02</td>
</tr>
<tr>
<td>Satimage</td>
<td>4435 36 6</td>
<td>0.57±0.06</td>
<td>0.55±0.00</td>
<td>0.21±0.05</td>
<td>0.51±0.01</td>
<td>0.30±0.00</td>
</tr>
<tr>
<td>Segment’n</td>
<td>1000 19 7</td>
<td>0.52±0.03</td>
<td>0.34±0.00</td>
<td>0.23±0.09</td>
<td><strong>0.61±0.05</strong></td>
<td>0.52±0.10</td>
</tr>
<tr>
<td>Vehicle</td>
<td>846 18 4</td>
<td>0.10±0.00</td>
<td>0.14±0.00</td>
<td>0.02±0.02</td>
<td><strong>0.14±0.00</strong></td>
<td>0.05±0.00</td>
</tr>
<tr>
<td>Vowel</td>
<td>990 10 11</td>
<td>0.42±0.01</td>
<td>0.44±0.00</td>
<td>0.28±0.03</td>
<td>0.39±0.02</td>
<td>0.41±0.02</td>
</tr>
<tr>
<td>Wine</td>
<td>178 13 3</td>
<td>0.84±0.01</td>
<td>0.46±0.00</td>
<td>0.56±0.02</td>
<td><strong>0.90±0.02</strong></td>
<td>0.59±0.01</td>
</tr>
</tbody>
</table>
Datasets: 20News and Reuters-R8.

- Landmarks $L \in \{5, 10, 15, 20, 25\}$.
- 80/20% split for training ans testing, respectively.
- $T = 50$.
- Set $v = 0.1$, $c = 9$ and $\ell = 0.1$ (MMCTM).
- Set $v = \ell = 0.1$ and $c \in \{0.1, 0.2, 0.5, 0.7, 1, 3, 5, 7, 9, 15, 30, 50\}$, pick best performance (MMCTM*).
- Initialize MMCTM with CTM.

<table>
<thead>
<tr>
<th>$L$</th>
<th>CTM</th>
<th>SVM</th>
<th>S3VM</th>
<th>MMCTM</th>
<th>MMCTM*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20NEWS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>17.22± 4.2</td>
<td>37.13± 2.9</td>
<td>39.36± 3.2</td>
<td>56.70± 1.9</td>
<td>57.86± 0.9</td>
</tr>
<tr>
<td>10</td>
<td>24.50± 4.5</td>
<td>46.99± 2.4</td>
<td>47.91± 2.8</td>
<td>54.92± 1.6</td>
<td>56.56± 1.3</td>
</tr>
<tr>
<td>15</td>
<td>22.76± 4.2</td>
<td>52.80± 1.2</td>
<td>52.49± 1.4</td>
<td>55.06± 2.7</td>
<td>57.80± 2.2</td>
</tr>
<tr>
<td>20</td>
<td>26.07± 7.2</td>
<td>56.10± 1.5</td>
<td>54.44± 2.1</td>
<td>56.62± 2.2</td>
<td>59.70± 1.4</td>
</tr>
<tr>
<td>25</td>
<td>27.20± 1.5</td>
<td>59.15± 1.4</td>
<td>57.45± 1.7</td>
<td>55.70± 2.4</td>
<td>61.92± 3.0</td>
</tr>
<tr>
<td></td>
<td>Reuters-R8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>41.27± 16.7</td>
<td>78.12± 1.1</td>
<td>78.51± 2.3</td>
<td>79.18± 4.1</td>
<td>80.86± 2.9</td>
</tr>
<tr>
<td>10</td>
<td>42.63± 7.4</td>
<td>80.69± 1.2</td>
<td>79.15± 1.2</td>
<td>80.04± 5.3</td>
<td>83.48± 1.0</td>
</tr>
<tr>
<td>15</td>
<td>39.67± 9.9</td>
<td>83.25± 1.7</td>
<td>81.87± 0.8</td>
<td>85.48± 2.1</td>
<td>86.86± 2.5</td>
</tr>
<tr>
<td>20</td>
<td>58.24± 8.3</td>
<td>85.66± 1.0</td>
<td>73.95± 2.0</td>
<td>82.92± 1.7</td>
<td>83.82± 1.6</td>
</tr>
<tr>
<td>25</td>
<td>51.93± 5.9</td>
<td>84.95± 0.1</td>
<td>82.39± 1.8</td>
<td>86.56± 2.5</td>
<td>88.12± 0.5</td>
</tr>
</tbody>
</table>

Figure 4: Accuracy vs. #topic

(a) 20NEWS dataset

(b) Reuters-R8 dataset