Dirichlet-vMF mixture model

Shaohua Li
shaohua@gmail.com
National University of Singapore

February 19, 2017

We adopt a simplification of the Bayesian vMF mixture model proposed in [2][1]. For computational efficiency, the priors on the vMF mean $\{\boldsymbol{\mu}_k\}$ and on the vMF concentration $\{\kappa_k\}$ are removed.

1 Model Specification

The generative process is as follows:

1. $\boldsymbol{\theta}_i \sim \mathrm{Dir}(\alpha)$;
2. $z_{ij} \sim \mathrm{Cat}(\boldsymbol{\theta}_i)$;
3. $\boldsymbol{x}_{ij} \sim \mathrm{vMF}(\boldsymbol{\mu}_{z_{ij}}, \kappa_{z_{ij}})$.

Here α is a hyperparameter, $\{\boldsymbol{\mu}_k, \kappa_k\}$ are parameters of mixture components to be learned.

2 Model Likelihood and Inference

Given parameters $\{\boldsymbol{\mu}_k, \kappa_k\}$, the complete-data likelihood of a dataset $\{\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Theta}\} = \{\boldsymbol{x}_{ij}, z_{ij}, \boldsymbol{\theta}_i\}$ is:

$$p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Theta} | \alpha, \{\boldsymbol{\mu}_k, \kappa_k\}) = \prod_i \mathrm{Dir}(\boldsymbol{\theta}_i | \alpha) \prod_j \theta_{i,z_{ij}} \mathrm{vMF}(\boldsymbol{x}_{ij} | \boldsymbol{\mu}_{z_{ij}}, \kappa_{z_{ij}}). \tag{1}$$

The incomplete-data likelihood of $\{\boldsymbol{X}, \boldsymbol{\Theta}\} = \{\boldsymbol{x}_{ij}, \boldsymbol{\theta}_i\}$ is obtained by integrating out the latent variables $\boldsymbol{Z}, \boldsymbol{\Theta}$:

$$p(\boldsymbol{X} | \alpha, \{\boldsymbol{\mu}_k, \kappa_k\}) = \int d\boldsymbol{\Theta} \cdot \prod_i \mathrm{Dir}(\boldsymbol{\theta}_i | \alpha) \prod_j \sum_k \theta_{ik} \mathrm{vMF}(\boldsymbol{x}_{ij} | \boldsymbol{\mu}_k, \kappa_k). \tag{2}$$

[1] This model reappears in [3] under the name “mix-vMF topic model”. But the inference scheme in [3] is sampling-based, which is presumed to be less accurate than the EM algorithm presented here.

(2) is apparently intractable, and instead we seek its variational lower bound:

$$\begin{aligned}\log p(\boldsymbol{X}|\alpha, \{\boldsymbol{\mu}_k, \kappa_k\}) \geq & E_{q(\boldsymbol{Z},\boldsymbol{\Theta})}[\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Theta}|\alpha, \{\boldsymbol{\mu}_k, \kappa_k\}) - \log q(\boldsymbol{Z}, \boldsymbol{\Theta})]. \\ = & \mathcal{L}(q, \{\boldsymbol{\mu}_k, \kappa_k\}) \end{aligned} \tag{3}$$

It is natural to use the following variational distribution to approximate the posterior distribution of $\boldsymbol{Z}, \boldsymbol{\Theta}$:

$$q(\boldsymbol{Z}, \boldsymbol{\Theta}) = \prod_i \Big\{ \mathrm{Dir}(\boldsymbol{\theta}_i|\boldsymbol{\phi}_i) \prod_j \mathrm{Cat}(z_{ij}|\boldsymbol{\pi}_{ij}) \Big\}. \tag{4}$$

Then the variational lower bound is

$$\begin{aligned}&\mathcal{L}(q, \{\boldsymbol{\mu}_k, \kappa_k\}) \\ =&C_0 + \mathcal{H}(q) + E_{q(\boldsymbol{Z},\boldsymbol{\Theta})} \Big[(\alpha - 1) \sum_{i,k} \log \theta_{ik} \\ &+ \sum_{i,j,k} \delta(z_{ij} = k)(\log \theta_{ik} + \log c_d(\kappa_k) + \kappa_k \boldsymbol{\mu}_k^\top \boldsymbol{x}_{ij}) \Big] \\ =&C_0 + \mathcal{H}(q) + \sum_{i,k} (\alpha - 1 + n_{i\cdot k}) \Big(\psi(\phi_{ik}) - \psi(\phi_{i0}) \Big) \\ &+ \sum_k \Big(n_{\cdot\cdot k} \cdot \log c_d(\kappa_k) + \kappa_k \boldsymbol{\mu}_k^\top \boldsymbol{r}_k \Big), \end{aligned} \tag{5}$$

where

$$n_{i\cdot k} = \sum_j \pi_{ijk}, \quad n_{\cdot\cdot k} = \sum_{i,j} \pi_{ijk}, \tag{6}$$

$$\boldsymbol{r}_k = \sum_{i,j} \pi_{ijk} \cdot \boldsymbol{x}_{ij}, \tag{7}$$

and $\mathcal{H}(q)$ is the entropy of $q(\boldsymbol{Z}, \boldsymbol{\Theta})$:

$$\begin{aligned}\mathcal{H}(q) =& -E_q[\log q(\boldsymbol{Z}, \boldsymbol{\Theta})] \\ =& \sum_i E_q \Big[\sum_k \log \Gamma(\phi_{ik}) - \log \Gamma(\phi_{i0}) - \sum_k (\phi_{ik} - 1) \log \theta_{ik} \\ &- \sum_{j,k} \delta(z_{ij} = k) \log \pi_{ijk} \Big] \\ =& \sum_i \Big(\sum_k \log \Gamma(\phi_{ik}) - \log \Gamma(\phi_{i0}) - \sum_k (\phi_{ik} - 1) \psi(\phi_{ik}) \Big) \\ &+ (\phi_{i0} - K) \psi(\phi_{i0}) - \sum_{j,k} \pi_{ijk} \log \pi_{ijk}. \end{aligned} \tag{8}$$

By taking the partial derivative of (5) w.r.t. $\pi_{ijk}, \phi_{ik}, \boldsymbol{\mu}_k, \kappa_k$, respectively, we can obtain the following variational EM update equations [1, 2, 3].

2.1 E-Step

$$\begin{aligned} \pi_{ijk} &\sim e^{\psi(\phi_{ik})} \cdot \mathrm{vMF}(\boldsymbol{x}_{ij} | \boldsymbol{\mu}_k, \kappa_k), \\ \phi_{ik} &= n_{i \cdot k} + \alpha. \end{aligned} \tag{9}$$

2.2 M-Step

$$\begin{aligned} \boldsymbol{\mu}_k &= \frac{\boldsymbol{r}_k}{\|\boldsymbol{r}_k\|}, \\ \bar{r}_k &= \frac{\|\boldsymbol{r}_k\|}{n_{\cdot\cdot k}}, \\ \kappa_k &\approx \frac{\bar{r}_k D - \bar{r}_k^3}{1 - \bar{r}_k^2}. \end{aligned} \tag{10}$$

The update equation of κ_k adopts the approximation proposed in [1].

References

[1] Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*, 6(Sep):1345–1382, 2005.

[2] Siddharth Gopal and Yiming Yang. Von mises-fisher clustering models. In *ICML*, pages 154–162, 2014.

[3] Ximing Li, Jinjin Chi, Changchun Li, Jihong OuYang, and Bo Fu. Integrating topic modeling with word embeddings by mixtures of vmfs. In *COLING*, 2016.