Dirichlet-vMF mixture model

Shaohua Li
shaohua@gmail.com
National University of Singapore

February 20, 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. This model is referred to as **VMFMix**.

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 [4] under the name “mix-vMF topic model”. But [4] only offers a sampling-based inference scheme, which is usually less accurate than the EM algorithm presented in this document.

(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, 4].

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].

3 Evaluation

The performance of this model was evaluated on two text classification tasks that are on 20 Newsgroups (**20News**) and **Reuters**, respectively. The experimental setup for the compared methods were identical to that in [3]. Similar to TopicVec, VMFMix learns an individual set of K topic embeddings from each category of documents, and all these sets are combined to form a bigger set of topic embeddings for the whole corpus. This set of topic embeddings are used to derive the topic proportions of each document, which are taken as features for the SVM classifier. The K for 20News and Reuters are chosen as 15 and 12, respectively, which are identical to TopicVec.

The macro-averaged precision, recall and F1 scores of all methods are presented in Table 1.

	20News			Reuters		
	Prec	Rec	F1	Prec	Rec	F1
BOW	69.1	68.5	68.6	92.5	90.3	91.1
LDA	61.9	61.4	60.3	76.1	74.3	74.8
sLDA	61.4	60.9	60.9	88.3	83.3	85.1
LFTM	63.5	64.8	63.7	84.6	86.3	84.9
MeanWV	70.4	70.3	70.1	92.0	89.6	90.5
Doc2Vec	56.3	56.6	55.4	84.4	50.0	58.5
TWE	69.5	69.3	68.8	91.0	89.1	89.9
TopicVec	**71.3**	**71.3**	**71.2**	**92.5**	**92.1**	**92.2**
VMFMix	63.8	63.9	63.7	87.9	88.7	88.0

Table 1: Performance on multi-class text classification. Best score is in boldface.

We can see from Table 1 that, VMFMix achieves better performance than Doc2Vec, LDA, sLDA and LFTM. However, its performance is still inferior to BOW, Mean word embeddings (MeanWV), TWE and TopicVec. The reason might be that by limiting the embeddings in the unit hypersphere (effectively normalizing them as unit vectors), certain representational flexibility is lost.

An empirical observation we have is that, VMFMix approaches convergence very quickly. The variational lower bound increases only slightly after 10~20 iterations. By manually checking the intermediate parameter values, we see that after so many iterations, the parameters change very little too. It suggests that VMFMix might easily get stuck in local optima.

Nonetheless, VMFMix might still be relevant when the considered embedding vectors are infinite and continuously distributed in the embedding space, as opposed to the finite vocabulary of word embeddings[2]. Such scenarios include the neural encodings of images from a convolutional neural network (CNN).

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] Shaohua Li, Tat-Seng Chua, Jun Zhu, and Chunyan Miao. Generative topic embedding: a continuous representation of documents. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016.

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

[2] Each set of word embeddings can be viewed as a finite and *discrete* sample from a *continuous* embedding space.