Topic Modeling Research

Melody Jiang

May 2019

1 Possible Research Directions

1.1 Two main approaches to clustering

(i) **Distance-based** clustering. Using this approach, we analyze matrix of pairwise distances. Namely, suppose y_i and y_j are data points, then D is a distance matrix whose d_{ij} entry represents distance between y_i and y_j.

(ii) **Model-based** clustering. For example, $y_i \sim \sum_{h=1}^{k} \pi_h \mathcal{K}(\theta_h)$.

1.2 Two main problems in clustering

(i) sensitivity to kernel

(ii) issues in high dimensions (large p)

1.3 Semi-solutions

1. **C-Bayes**. All derivations from assumed models (e.g. kernel misspecification). See [Miller and Dunson, 2018].
2. **Model plus distance-based clustering**. See [Duan and Dunson, 2018].
3. **Calculating better distances**. E.g., geodesic or intrinsic distace (Didong Li & Dunson, in preparation).
4. **To address issues in high dimensions**, cluster on the latent variable level or varational autoencoder (VAE).

2 Literature Review

Three main tasks in textmining are clustering, classification, and information extraction [Allahyari et al., 2017]. Topic modeling can be applied to all of these tasks [Lu et al., 2011]. We would like to focus on the most commonly used topic model, Latent Dirichlet Allocation (LDA), and LDA's robustness in the documeng clustering task. Lu et al. investigated LDA's task performance in

document clustering and found LDA's performance is quite sensitive to the setting of its hyper-parameter and parameter [Lu et al., 2011]. In terms of robustness, Wang et al. proposed a model-based approach to make LDA more robust by using localization and empirical Bayes [Wang et al., 2018].

2.1 Latent Dirichlet Allocation (LDA)

In this section, we reproduce the LDA model described in [Wang et al., 2018] and [Blei and Lafferty, 2009].

Let:

- K be a specified number of topics,
- D the number of documents,
- N_d the number of words in a document,
- V the size of the vocabulary,
- $\boldsymbol{\alpha}$ a positive K-vector,
- η a scalar.
- $Dir_K(\boldsymbol{\alpha})$ a K-dimensional Dirichlet distribution with vector parameter $\boldsymbol{\alpha}$,
- $Dir_V(\eta)$ a V-Dimensional symmetric Dirichlet distribution with scalar parameter η.

A *symmetric Dirichlet* is a Dirichlet distribution where each component of the parameter is equal to the same value.

We define each topic $\boldsymbol{\beta}_k$ to be a distribution over a fixed vocabulary and we fix the number of topics K. LDA assumes that a collection of documents comes from the following process:

1. Draw each topic $\boldsymbol{\beta}_k \sim Dir_V(\eta)$ for k = 1, 2, ..., K.
2. For each document d ($d = 1, ..., D$),
   (a) Draw a vector of topic proportions $\boldsymbol{\theta}_d \sim Dir_K(\boldsymbol{\alpha})$
   (b) For each word w_n ($n = 1, ..., N_d$) in document d,
      i. Draw topic assignment $z_{dn} \sim Mult(\boldsymbol{\theta}_d)$, where $z_{dn} \in \{1, ..., K\}$.
      ii. Draw word $w_{dn} \sim Mult(\boldsymbol{\beta}_{z_{dn}})$, where $w_{dn} \in \{1, ..., V\}$.

See Figure 1 for a directed graphical model of LDA.

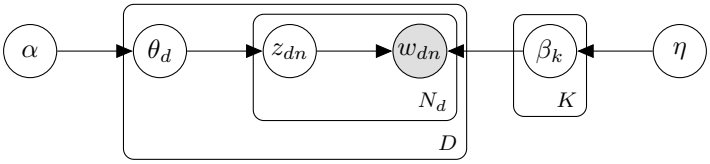

Figure 1: The directed graphical model of LDA. Adapted from [Blei, 2012] and [Blei and Lafferty, 2009]. Nodes denote random variables; edges denote dependence between random variables. Shaded nodes denote observed random variables; unshaded nodes denote hidden random variables. The rectangular boxes are "plate notation", which denote replication.

3 Pilot Experiment

In this section, we perform experiments similar to those in [Lu et al., 2011]. We investigate how the hyperparameter α and the total number of topics K affect LDA's performance in document clustering.

We use the Reuters-21578 R8 [1] dataset. Reuters-21578 R8 is a pre-processed subset of Reuters-21578, [2] a very widely used dataset in textmining research. Training data and testing data are provided in two separate files.

All analyses are done in R.

3.1 Preprocessing

First, we subset training data to 3 document classes (topics) for computational expense. We do the same for testing data.

Similar to what was done in [Lu et al., 2011], we study LDA in the standard clustering setting, where each document belongs to exactly one cluster. Hence, we remove documents appearing in two or more categories. Now we end up with 679 documents for training data and 279 documents for testing data.

As we have mentioned, Reuters-21578 R8 is a pre-processed subset of Reuters-21578. Preprocessing that has already been done for us are:

1. Substituteing TAB, NEWLINE and RETURN characters by SPACE;
2. Keeping only letters (i.e., turn punctuation, numbers, etc. into SPACES);
3. Turning all letters to lowercase.
4. Substituting multiple SPACES by a single SPACE.
5. The title/subject of each document is simply added in the beginning of the document's text.

Preprocessing steps performed by us are stopword removal and stemming. Tokenization is automatically performed when we create document-term matrices.

[1] https://www.cs.umb.edu/ smimarog/textmining/datasets/
[2] http://www.daviddlewis.com/resources/testcollections/reuters21578/

3.2 Evaluation Metric

We use perplexity of testing set as our evaluation metric, since the documents in the corpora are treated as if they were unlabeled. Perplexity is commonly used in topic modeling literature [Blei et al., 2003, Blei et al., 2007]. Other different evaluation metrics worth considering for future experiments include per-word predictive log likelihood used in [Wang et al., 2018] and normalized mutual information used in [Lu et al., 2011].

As described in [Blei et al., 2003], the perplexity is monotonically decrasing in the likelihood of the testing data, and perplexity is algebratically equivalent to the inverse of the geometric mean per-word likelihood. A lower perplexity score indicates better generalization performance of the model. For a testing dataset that consists of M documents, the perplexity is:

$$\text{perplexity } (D_{\text{ test }}) = \exp\left\{-\frac{\sum_{d=1}^{M} \log p\left(\mathbf{w}_d\right)}{\sum_{d=1}^{M} N_d}\right\}$$

.

3.3 Results

First, using Gibbs sampling as estimation method, we calculate perplexity for $K = 2, 3, ..., 10$. α is estimated by the algorithm. See Figure 2 for result.

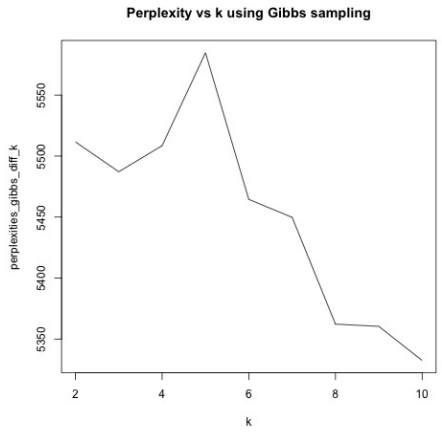

Figure 2: Perplexity versus number of topics, Gibbs sampling

Then, using variational expectation-maximization (VEM) instead of Gibbs sampling as estimation method, we again calculate perplexity for $K = 2, 3, ..., 10$. α is still estimated by the algorithm. See Figure 3 for result.

At last, we use VEM as estimation method, fix the number of topics K to be 3, and calculate perplexity for $\alpha = 0.01, 0.1, 1, 5, 10, 25$. Recall that we subsetted both training and testing data to be having 3 topics during preprocessing, so 3

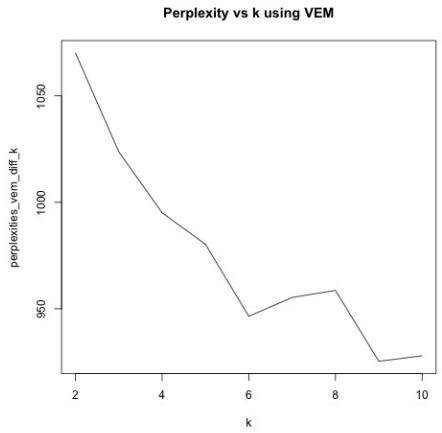

Figure 3: Perplexity versus number of topics, VEM

is the actual number of topics for both training and testing data. See Figure 4 for results.

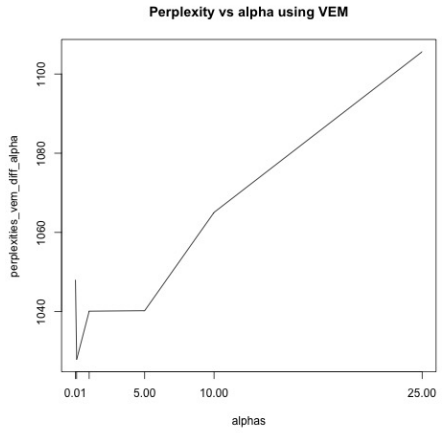

Figure 4: Perplexity versus α, VEM

3.4 Discussion

LDA is sensitive to the number of topics K and the hyperparameter α.

From Figure 2 and Figure 3, we observe that the model tends to perform better with a larger number of topics, despite that the actual number of topics is 3. It may be worth investigating what causes this result - whether it is the nature of LDA, the evaluation method, or something else.

Figure 4 shows that LDA performs best when α is 0.1, and the performance worsen as α increases. This results agrees with the result in [Lu et al., 2011]. The reason we obtain such result might be, as [Lu et al., 2011] explained, while a larger value of α leads to more smoothed topics, α smaller than 1 would cause the modes of the Dirichlet distribution to concentrate at corners of the simplex, thus favoring more sparse topics. Since we limited each document in our data to be in only one cluster, we would expect LDA to assgin a skewed topic distribution to a document. Therefore, a smaller α should result in better performance.

4 Annotated Bibliography

[Wang et al., 2018] Title: A general method for robust Bayesian Modeling. This apaper proposes a general model-based approach to robustify Bayesian models.

[Doyle and Elkan, 2009] Bursty Bayesian models.

[Blei et al., 2003] Title: Latent dirichlet allocation. An introduction to latent dirichlet allocation.

[Lu et al., 2011] Title: Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. Investigates how alpha and k affect LDA's and PLSA's performance.

[Allahyari et al., 2017] Title: A brief survey of text mining: Classification, clustering and extraction techniques

[Blei et al., 2007] Title: A correlated topic model of science. Used for justifying using perplexity as evalutation metric.

5 Thoughts (for my own record)

1. document belonging to more than one category possible? Definitely - LDA is a mixture model. Other uses of topic model in text clustering? How are things often done in clustering research?

2. Outlier categories?

3. To be investigated: More specific goal of analysis - what does document clustering really looklike?

References

[Allahyari et al., 2017] Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*.

[Blei, 2012] Blei, D. M. (2012). Probabilistic topic models. *Commun. ACM*, 55(4):77–84.

[Blei and Lafferty, 2009] Blei, D. M. and Lafferty, J. D. (2009). Topic models. In *Text Mining*, pages 101–124. Chapman and Hall/CRC.

[Blei et al., 2007] Blei, D. M., Lafferty, J. D., et al. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[Doyle and Elkan, 2009] Doyle, G. and Elkan, C. (2009). Accounting for burstiness in topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 281–288. ACM.

[Duan and Dunson, 2018] Duan, L. L. and Dunson, D. B. (2018). Bayesian distance clustering. *arXiv preprint arXiv:1810.08537*.

[Lu et al., 2011] Lu, Y., Mei, Q., and Zhai, C. (2011). Investigating task performance of probabilistic topic models: an empirical study of plsa and lda. *Information Retrieval*, 14(2):178–203.

[Miller and Dunson, 2018] Miller, J. W. and Dunson, D. B. (2018). Robust bayesian inference via coarsening. *Journal of the American Statistical Association*, pages 1–13.

[Wang et al., 2018] Wang, C., Blei, D. M., et al. (2018). A general method for robust bayesian modeling. *Bayesian Analysis*, 13(4):1159–1187.