Topic Modeling and Clustering

Melody Jiang

May 2019

Contents

1 Introduction

Three main tasks in textmining are clustering, classification, and information extraction [2]. Topic modeling can be applied to all of these tasks [9]. We would like to focus on the most commonly used topic model, Latent Dirichlet Allocation (LDA), and its performance 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 [9]. In terms of robustness, Wang et al. proposed a model-based approach to make LDA more robust by using localization and empirical Bayes [11].

In this section, we review types of clustering methods in text clustering, Latent Dirichlet Allocation (LDA), and evaluation metrics for clustering and LDA. In the following sections, we would present simulation experiments. We first verified that performance of LDA is sensitive to settings of hyperparameter α. We then attempted incorporating perturbation in posterior samples and tested perturbation on Dirichlet-Multinomial. At last, we discuss the current idea of using proper scoring rules and cross validation to tune the hyperparameter α.

1.1 Types of document clustering

Hard Clustering versus Soft Clustering. Hard clustering computes a hard assignment of each document to a specific cluster. That is, each document is a member of exactly one cluster. In soft clustering, a document's assignment is a distribution over all clusters [10]. Hard clustering can be seen as a special case of soft clustering.

Model-based Clustering versus Distance-based Clustering. Model-based clustering assumes that the data were generated by a model and tries to recover the original model from the data. A commonly used criterion for estimating the model parameters is maximum likelihood, and the expectation-maximization(EM) algorithm is commonly used for such computation [10].

Distance-based clustering algorithms use a similarity function to measure the closeness between text objects [1]. Two most widely used distanced-based clustering algorithms are k-medoid clustering algorithm and k-means clustering algorithms [1, 10].

Flat Clustering versus Hiearchical Clustering Flat clustering creates a flat set of cluster without any explicit structure that would relate clusters to each other while Hierarchical clustering creates a hierarchy of clusters [10].

1.2 Latent Dirichlet Allocation (LDA)

In this section, we reproduce the LDA model described in [11] and [4].

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.

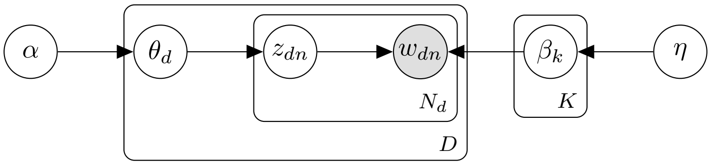

Figure 1: The directed graphical model of LDA. Adapted from [3] and [4]. Nodes denote random variables; edges denote dependence between random variables. Shaded nodes denote observed random variables; unshaded nodes denote hidden random variables. The rectangular boxes are “plate notation”, which denote replication.

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.

1.2.1 Likelihood functions of LDA

$$p_d(w|\{\boldsymbol{\beta}_k\}_{1:K}, \boldsymbol{\theta}_d) = \sum_{k=1}^{K} \theta_{dk} p(w|\boldsymbol{\beta}_k) \tag{1}$$

$$\begin{aligned} & p(d|\boldsymbol{\alpha}, \{\boldsymbol{\beta}_k\}_{1:K}) && (2) \\ & = \int \left\{ \prod_{w \in V} \left[\sum_{k=1}^{K} \theta_{dk} p(w|\boldsymbol{\beta}_k) \right]^{c(w,d)} \right\} p(\boldsymbol{\theta}_d|\boldsymbol{\alpha}) d\boldsymbol{\theta}_d && (3) \end{aligned}$$

1.3 Evaluation

Evaluation of clustering includes using internal criterion and external criterion for evaluation. Most objective functions in clustering formalize the goal of attaining high intracluster similarity and low inter-cluster similarity. That is, in the context of text clustering, documents within a cluster are similar and documents from different clusters are dissimilar. This is an internal criterion for the quality of a clustering. However, good scores on an internal criterion do not necessarily mean effectiveness in applications [10].

To ensure effectiveness in applications, one can compute an external criterion that evaluates how well clustering matches an evaluation benchmark or gold standard [10]. Commonly used external criteria include purity, normalized mutual information, Rand index, and F measure [10]. Purity is simple and transparent. Normalized mutual information has the advantage of being able to be information-theoretically interpreted. The Rand index penalizes both false-positive and false-negative decisions during clustering. The F meassure not only penalizes false-negative and falose-positive decisions, but also supports differential weighting of these two types of errors [10]. The gold standard is ideally produced by human judges with a good level of inter-judge agreement. Historically, the Reuters-21578 collection was the main benchmark for text classification evaluation and is commonly used in evaluation of text clustering [10].

Outside of the contexts in applications, topic models on their own are most commonly evaluted by using perplexity [8]. As described in [6], 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:

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

.

2 Experiment: How α affects performance of LDA

We performed experiments similar to those in [9]. We investigated how the hyperparameter α and the total number of topics K affect LDA's performance in document clustering.

We used 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 were done in R.

[1] https://www.cs.umb.edu/ smimarog/textmining/datasets/

[2] http://www.daviddlewis.com/resources/testcollections/reuters21578/

2.1 Preprocessing

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

Similar to what was done in [9], we studied LDA in the standard clustering setting, where each document belongs to exactly one cluster. Hence, we removed 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 was automatically performed when we create document-term matrices. After preprocessing, there were 5291 terms in the vocabulary.

2.2 Evaluation Metric

We used 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 [6, 5]. Other different evaluation metrics worth considering for future experiments include per-word predictive log likelihood used in [11] and normalized mutual information used in [9].

2.3 Results

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

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

At last, we used VEM as estimation method, fixed the number of topics K to be 3, and calculated perplexity for different values of α. Recall that we subsetted both training and testing data to be having 3 topics during preprocessing, so 3 was the actual number of topics for both training and testing data. See Figure 4 for results. Lu et al. investigated clustering-task peformance of LDA under $\alpha = 0.01, 0.1, 1, 5, 10, 25$ [9] using normalized mutual information. We emulated them and examined perplexity under these settings. However, since the Dirichlet

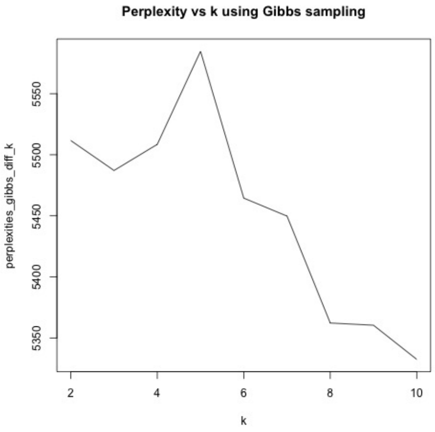

Figure 2: Perplexity versus number of topics, Gibbs sampling

prior should be sensitive to small changes in its parameters in high-dimensional setting, we examine 30 unevenly spaced values between 0.001 and 1 as well.

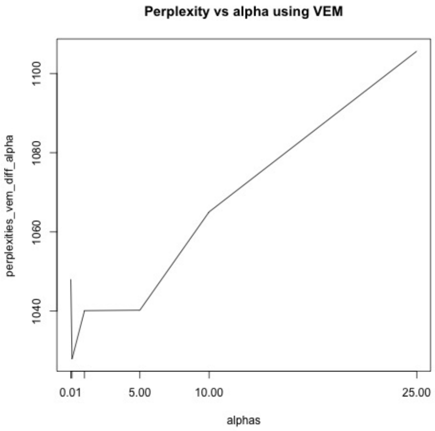

Figure 4: Perplexity versus α, VEM

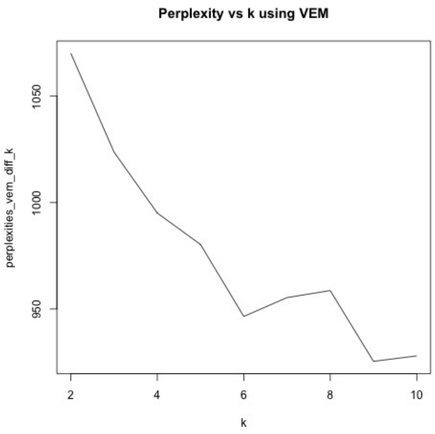

Figure 3: Perplexity versus number of topics, VEM

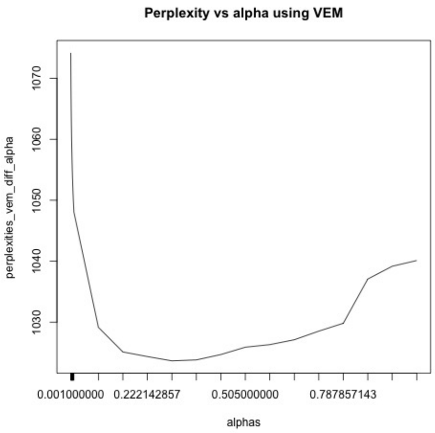

Figure 5: Perplexity versus α, VEM

2.4 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 on test data with a larger number of topics, despite that the actual number of topics is 3. This contradicts our expectation that large number of topics would cause the model to overfit and perform poorly on test data.

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 [9]. The reason we obtain such result might be, as [9] 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. However, as shown in Figure 5, we also observe that α too close to 0 ($\alpha \in [0.001, 0.1]$) results in significant decline of performance. This indicates sensitivity of LDA to settings of α.

For future improvements on this experiment, we would add measures other than perplexity to evaluate task performance of LDA on clustering. We could also investigate the relationship between perplexity and settings of parameters and propose better evaluation metrics.

3 Experiment: Perturbation

To address the issue that Dirichlet is sensitive to small changes in its parameters in high dimensional settings, our first idea was to add perturbation to posterior samples. To examine what happens when we add perturbation, we started with the simple model of Dirichlet multinomial in a high dimensional setting.

3.1 Dirichlet Multinomial

The multinomial sampling distribution is used to describe data for which each observation is one of k possible outcomes.

If y is the vector of counts of the number of observations of each outcome, then

$$p(y|\theta) \propto \prod_{j=1}^{k} \theta_j^{y_j}$$

, where the sum of the probabilities, $\sum_{j=1}^{k} \theta_j = 1$.

The distirbution is typically thought of as implicitly conditioning on the number of observations, $\sum_{j=1}^{k} y_j = n$.

The conjugate piror distribution is the Dirichlet,

$$p(\theta|\alpha) \propto \prod_{j=1}^{k} \theta_j^{\alpha_j - 1}$$

, where $\sum_{j=1}^{k} \theta_j = 1$ and $\theta_j \geq 0$.

The posterior distribution is Dirichlet with parameters

$$\alpha_j + y_j$$

.

3.2 Perturbation

Suppose $(\theta_1, ..., \theta_p)$ is a sample from the Dirichlet posterior. We perturb the sample as follows:

$$\Delta_j \sim_{iid} Gamma(\kappa, 1), j = 1, ..., p$$

$$\theta'_j = \frac{\theta_j \Delta_j}{\sum_{i=1}^{p} \theta_i \Delta_i}, j = 1, ..., p$$

3.3 Simulation 1: Fixed probability vector

Let there be n observations and p number of categories, where p is large. We would choose $n < p$ because it makes sense that when there are a huge number of features we do not have that many data collected. Let $\boldsymbol{\theta}$ denote the probability vector associated with categories.

In our simulation, we chose the number of observations $n = 50$ and the number of categories $p = 100$. We chose $\boldsymbol{\theta} = (\frac{1}{p}, ..., \frac{1}{p})$. We generated simulated data by sampling from $Multinomial(n, \boldsymbol{\theta})$. We chose parameters of the Dirichlet prior $\alpha = (10^{-6}, ..., 10^{-6})$ to be close to 0. After obtaining the posterior distribution, we calculated 95% credible intervals for each entry of the probability vector $\boldsymbol{\theta}$.

We define "coverage rate" as the percentage of categories whose corresponding true probability is covered by credible interval. In this simulation study, coverage rate was 36%. Setting $\kappa = 2$, perturbation improved coverage rate to 39%. However, given that we would like 95% credible intervals, this result was of not very much practical value. The poor performance was due to sparsity of data. Categories with observations "attracts" probability mass in the posterior distribution, causing credible intervals to be higher than the actual probabilities, while categories with no observations have probability mass taken too much from them, resulting in credible intervals lower than the actual probabilities.

3.4 Simulation 2: Probability vector from Dirichlet

In this simulation, we illustrate that coverage rate is poor under model misspecification, even when data is generated from Dirichlet Multinomial.

As in Section 3.3, let there be n observations and p number of categories, where p is large. Let $\boldsymbol{\theta}$ denote the probability vector associated with categories.

In this simulation, we chose the number of observations $n = 600$ and the number of categories $p = 5000$. We generated simulated data from Dirichlet-multinomial with $\alpha_{sample} = (0.01, ..., 0.01)$. For our model, we chose the popular setting $\alpha_{prior} = (0.1, ..., 0.1)$ for our Dirichlet prior.

We performed the following for 30 interations. During each iteration t, we first sampled probability vector $\boldsymbol{\theta}_t$ from $Dir(\alpha_{sample})$. Then, using probability vector $\boldsymbol{\theta}_t$, we generated simulated data as what we did in Section 3.3. Then, we calculate coverage rate for this iteration.

For all iterations, coverage rate is around 30%, supporting our conjecture that coverage rate is poor under model misspecification.

4 Experiment: Proper scoring rules and cross validation

Scoring rules, as a method to assess the quality of probabilisitic forecasts, can be used to tune parameters of priors [7]. In our case, we can use a scoring rule to tune α of the Dirichlet prior in the Dirichlet-multinomial model. We hypothesize that by obtaining better α, interval estimates for the underlying probability vector would be improved.

Scoring rules provide evaluation of probabilistic forecasts by assigning numerical score based on the predictive distribution and on the event or value that materializes [7]. Scoring rules are taken as positively oriented rewards that a forecaster wishes to maximize. That is, the higher the score is, the better the fore cast is. If the forecaster quotes predictive distribution P and event x materializes, then the reward, or the score, is $S(P, x)$.

The function $S(P, \cdot)$ takes values in the real line or in the extended real line, and we write $S(P, Q)$ for the expected value of $S(P, \cdot)$ under Q. Scoring rules are designed to have the property that $S(Q, Q) \geq S(P, Q)$, so that the forecaster is incentivized to quote $P = Q$. If a scoring rule has the property that $S(Q, Q) \geq S(P, Q)$ with equality if and only if $P = Q$, then the scoring rule is said to be strictly proper. If $S(Q, Q) \geq S(P, Q)$ for all P and Q, then the scoring rule is said to be proper.

In estimation probelms, strictly proper scoring rules have attractive properties [7]. Suppose we would like to fit a predictive distribution parameterized by θ, P_θ, based on a sample $X_1, ..., X_n$. To estimate θ, we can measure the goodness-of-fit by the mean score

$$S_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} S(P_\theta, X_i)$$

, where S is a strictly proper scoring rule. If θ_0 denotes the true parameter value, then asymptotic arguments indicate that

$$argmax_\theta S_n(\theta) \rightarrow \theta_0 \text{ as } n \rightarrow \infty$$

.

Hence, in our situation of Dirichlet-multinomial, we would be able to tune α by maximizing the mean score. In this case, the predictive distribution would be a posterior predictive distribution parameterized by α.

To prevent over-fitting, we incorporate cross-validation. Using cross-validation, we compute

$$argmax_\alpha \sum_{t=1}^{K} \frac{1}{n} \sum_{i=1}^{n} S(P_\alpha, X_i)$$

, where K is the number of folds in a K-fold cross validation and P_α is posterior predictive distribution parameterized by α.

5 Summary and Future Directions

We started with LDA and clustering, and along the way we discovered that Dirichlet-Multinomial performs poorly on providing interval estimates for sparse multinomial data. We are currently working on using proper scoring rules and cross validation to improve interval estimates of underlying probabilities for sparse multinomial data. For future work, we would like to develop better priors for sparse multinomial data. Although our focus right now is to improve interval estimates, it would also be of interest to see if our results improve performance of LDA.

References

[1] Charu C Aggarwal and ChengXiang Zhai. *Mining text data*. Springer Science & Business Media, 2012.

[2] Mehdi Allahyari et al. "A brief survey of text mining: Classification, clustering and extraction techniques". In: *arXiv preprint arXiv:1707.02919* (2017).

[3] David M. Blei. "Probabilistic Topic Models". In: *Commun. ACM* 55.4 (Apr. 2012), pp. 77–84. ISSN: 0001-0782. DOI: 10.1145/2133806.2133826. URL: http://doi.acm.org/10.1145/2133806.2133826.

[4] David M Blei and John D Lafferty. "Topic models". In: *Text Mining*. Chapman and Hall/CRC, 2009, pp. 101–124.

[5] David M Blei, John D Lafferty, et al. "A correlated topic model of science". In: *The Annals of Applied Statistics* 1.1 (2007), pp. 17–35.

[6] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[7] Tilmann Gneiting and Adrian E Raftery. "Strictly proper scoring rules, prediction, and estimation". In: *Journal of the American Statistical Association* 102.477 (2007), pp. 359–378.

[8] Thomas L. Griffiths and Mark Steyvers. "Finding scientific topics". In: *Proceedings of the National Academy of Sciences* 101.suppl 1 (2004), pp. 5228–5235. ISSN: 0027-8424. DOI: 10.1073/pnas.0307752101. eprint: https://www.pnas.org/content/101/suppl_1/5228.full.pdf. URL: https://www.pnas.org/content/101/suppl_1/5228.

[9] Yue Lu, Qiaozhu Mei, and ChengXiang Zhai. "Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA". In: *Information Retrieval* 14.2 (2011), pp. 178–203.

[10] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. "Introduction to Information Retrieval". In: New York, NY, USA: Cambridge University Press, 2008, pp. 327–331. ISBN: 0521865719, 9780521865715.

[11] Chong Wang, David M Blei, et al. “A general method for robust Bayesian modeling”. In: *Bayesian Analysis* 13.4 (2018), pp. 1159–1187.