Experiments with Latent Dirichlet Allocation

Phuc Xuan Nguyen(nguyen.phuc.x@gmail.com)
Panqu Wang(pawang@ucsd.edu)
Saekwang Nam(s9nam@ucsd.edu)

March 6, 2012

Abstract

Latent Dirichlet Allocation is a generative topic model for text. In this report, we implement collapsed Gibbs sampling to learn the topic model. We test our implementation on two data sets: classic400 and Psychological Abstract Review. We also discuss the different evaluation of goodness-of-fit of the models how parameter settings interact with the goodness-of-fit.

1 Introduction

Latent Dirichlet Allocation(LDA), originally introduced by Blei et al. (2003), is a generative model for words in documents. It has been widely researched and used in many applications such as text mining, information retrieval, and computer vision. One popular technique to learn an LDA model is via collapsed Gibbs sampling. The unsupervised nature of topic model, however, makes the process of model selection a challenging task. There are many works addressing methods of defining and computing the goodness-of-fit of topic models.

In section 2, we will provide a brief theoretical overview of LDA, collapsed Gibbs sampling, and different methods for evaluating the goodness-of-fit. Section 3 will describe our experiment design and results.

2 Theoretical Overview

2.1 Latent Dirichlet Allocation

In this model, a "topic" z is a latent variable described by the discrete multinomial distribution with parameter ϕ_z. Two Dirichlet priors with concentration parameters α and β are placed over the documents $\Theta = \{\theta_1, ..., \theta_D\}$ and the topics $\Phi = \{\phi_1, ..., \phi_K\}$:

$$p(\Theta) = \prod_d Dir(\theta_d; \alpha)$$

and

$$p(\Phi) = \prod_t Dir(\phi_t; \beta),$$

where D is the number of documents in the corpus.

The word tokens, w, are drawn from the topic's distributions:

$$p(\bar{w}^{(d)}|z^{(d)}, \Phi) = \prod_d \phi_{z_n^{(d)}}.$$

The variables are drawn with the following distributions:

$$\begin{aligned}\theta &\sim Dirichlet(\alpha),\\ \phi &\sim Dirichlet(\beta),\\ z &\sim Multinomial(\theta),\\ w &\sim Multinomial(\phi).\end{aligned}$$

A data set of documents $W = \{\bar{w}^{(1)}, ..., \bar{w}^{(D)}\}$ is observed, but the topic assignment $z = \{z^{(1)}, ..., z^{(D)}\}$ is latent.

2.2 Collapsed Gibbs sampling

We use collapsed Gibbs sampling to learn the LDA model. Gibbs sampling is based on the concept of sequentially re-drawing each topic assignment for slot n, z_n, from the posterior given $\bar{w}$, Φ, α, and $z_{\backslash n}$(all other topic assignments):

$$p(z_n = k|z_{\backslash n}, \bar{w}) \propto \phi_{w_n|k} \frac{\{N_k\}_{\backslash n} + \alpha_k}{N - 1 + \alpha},$$

where $\{N_k\}_{\backslash n}$ is the number of times topic k occurs in the document, excluding position n, and N is the total number of word tokens.

2.3 Goodness-of-fit

2.3.1 Definition

Defining and computing the goodness-of-fit of an LDA model is a tricky task. We will look at how to estimate the goodness-of-fit of the model under different settings.

If the topics ϕ_1 to ϕ_k are given, the goodness-of-fit could be modeled as $p(\bar{w}|\Phi, \alpha)$ where $\Phi = \{\phi_1, ..., \phi_K\}$. Wallach (2009) has described several methods for approximating the likelihood of data, $p(\bar{w}|\Phi, \alpha)$ when the topic distribution $\Phi = \{\phi_1, ..., \phi_K\}$ is given or learned. Furthermore, the author also argues that evaluation methods such as harmonic mean, importance sampling, and document completion methods, are generally inaccurate. He suggests the Chib-style estimator and the "left-to-right" algorithm as better alternatives for selecting topic models.

If the topics ϕ_1 to ϕ_k are not given, the goodness-of-fit could be modeled as $p(\bar{w}|\alpha, \beta)$, where α and β are the concentration of the Dirichlet priors. G. Doyle and C. Elkan (2009) suggests a non-parametric likelihood estimates, called empirical likelihood (EL). In this process, a set of pseudo documents are generated based on the LDA generative process. These pseudo documents are then used to train a tractable model, such as mixture of multinomials. The true likelihood is estimated as its likelihood in the pseudo corpus.

2.3.2 Comparing goodness-of-fit across different number of topics

Given α and β, we would like to choose a topic number, K, that has the best likelihood for the held-out data. We can model the likelihood as,

$$p(\bar{w}|K) = \int p(\bar{w}|\bar{z}, K)p(\bar{z}|K)dz. \tag{1}$$

We can approximate $p(\bar{w}|K)$ by using the general Monte Carlo method for evaluating integrals (Michael A.Newton, 1991)

$$p(\bar{w}|K) = \int g(z|K)p(z|K)dz \tag{2}$$

where $g(z|K) = p(\bar{w}|z, K)$. The author suggests computing estimator $\hat{I}$ of $p(\bar{w}|K)$ as

$$\hat{I} = \hat{p}(\bar{w}|K) = \frac{\sum_{m=1}^{M} t_m g(z^{(m)}|K)}{\sum_{m=1}^{M} t_m} \tag{3}$$

where

$$t_m = \frac{p(z^{(m)})}{p(z^{(m)} \mid w)} = \frac{p(z^{(m)})}{\frac{p(z^{(m)})p(\overline{w}|z^{(m)},K)}{p(\overline{w}|K)}} = \frac{p(\bar{w}|K)}{p(\bar{w}|z^{(m)},K)} \tag{4}$$

and $g(z^{(m)}|K) = p(\bar{w}|z^{(m)}, K)$. m is the document. By substituting t_m and $g(z^{(m)}|K)$ into (3), we get

$$\hat{p}(\bar{w}|K) = \frac{M}{\sum_{m=1}^{M} \frac{1}{p(\bar{w}|z^{(m)},K)}} \tag{5}$$

Therefore, the approximate probability is the harmonic mean of the likelihood value. Griffiths suggests that,

$$p(\bar{w}|z^{(m)}) = \left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\right)^K \prod_{k=1}^{K} \frac{\prod_{t=1}^{V}\Gamma(n_{kt}+\beta)}{\Gamma(n_{k\cdot}+V\beta)} \tag{6}$$

where n_{kt} is the number of times that word t occurs with topic k and $n_{k\cdot}$ is the number of times that all words occur with topic k. By substituting (6) into (5), we can compute the approximated likelihood of the data given K.

2.4 Overfitting

Overfitting occurs when a model memorizes training data rather than learns to generalize from trend. As an extreme example, if the number of parameters is the same as or greater than the number of observations, a simple model can learn to perfectly predict the training data simply by memorizing the training data entirely. Such a model will typically fail drastically on unseen data, as it has not learned to generalize. A common criterion of clustering quality is perplexity (Heinrich, 2005). A common method to evaluate perplexity in topic model is to test the model on held-out set from the main corpus. Higher value of perplexity indicates higher misrepresentation of the words of the test documents, thus indicating overfitting.

Mathematically, perplexity is defined as the reciprocal geometric mean of the likelihood of a test corpus given the model $\{\Phi, \Theta\}$, or $Q = \{\bar{w}, z\}$:

$$\mathrm{P}(\tilde{W} \mid Q) = \prod_{m=1}^{M} p(\tilde{\bar{w}}_{\tilde{m}} \mid Q)^{-\frac{1}{N}} = exp\{-\frac{\sum_{m=1}^{M} \log p(\tilde{\bar{w}}_{\tilde{m}} \mid Q)}{\sum_{m=1}^{M} N_m}\} \tag{7}$$

where

$$\begin{aligned} p(\tilde{\bar{w}}_{\tilde{m}} \mid Q) &= \prod_{n=1}^{N_{\tilde{m}}} \sum_{k=1}^{K} p(w_n = t \mid z_n = k) \cdot p(z_n = k \mid d = \widetilde{m}) \\ &= \prod_{t=1}^{V} (\sum_{k=1}^{K} \phi_{k,t} \cdot \theta_{\underset{\sim}{m},k})^{n_{\underset{\sim}{m}}^{(t)}} \end{aligned} \tag{8}$$

and,

$$log p(\tilde{\bar{w}}_{\tilde{m}} \mid Q) = \sum_{t=1}^{V} n_{\underset{\sim}{m}}^{(t)} log(\sum_{k=1}^{K} \phi_{k,t} \cdot \theta_{\underset{\sim}{m},k}) \tag{9}$$

where $n_{\underset{\sim}{m}}^{(t)}$ is the number of times term t has been observed in document $\tilde{m}$. Equation (7) suggests that increase of perplexity $\mathrm{P}(\widetilde{W} \mid Q)$ indicates the decrease of $\sum_{m=1}^{M} \log p(\tilde{\bar{w}}_{\tilde{m}} \mid Q)$.

With our definition of perplexity, we can detect whether an LDA model is overfitting by computing the perplexity on the test set of the corpus for different values of number of topics, K. We then select a value $\tilde{K}$ that gives that minimum perplexity. If this value is smaller than the number of topics provided by the model in question, then this model is likely overfitting its training data. A justification for this approach is that the larger the number of topics means larger parameter space. In the training stage, a larger parameter

space means a better fit to the data as it has more expressing power. If a smaller value of K has a lower perplexity on the test set, the LDA model is likely overfitting the training data.

3 Experiment

3.1 Hyperparameters

In the Gibbs sampling algorithm, the values of the Dirichlet priors, α and β, are assumed to be known. Many topic modeling papers use the heuristic values for the hyperparameters. In particular, common values are $\alpha = 50/K$ and $\beta = 0.1$ (Griffiths,2004), where K is the total number of topics. G. Doyle and C. Elkan (2009) have argued that fitted values yield better likelihood, especially when the number of topics is small. The optimized values α and β can be learned via performing a grid search with the likelihood of the data as the objective function.

We use the harmonic mean of θ as a goodness-of-fit metric given different values of α and β. We define a better fit as having smaller averaged harmonic mean $\frac{1}{M}\sum_{m=1}^{M} HM\{(\theta_{m,k})\}_{k=1}^{K}$. As a property of harmonic mean, smaller elements have a stronger impact on the mean. When a data point is much closer to its centroids than other centroids, the harmonic mean will be low. As the distance between a data point and its centroid increases, the harmonic means will increase.

We perform grid search for hyperparameter α and β with the harmonic mean as the objective function on the held-out set of 100 randomly picked documents. We set the search intervals as 0.1 to 5 with step size 0.2 for α, and 0.5 to 10 with step size 0.5 for β. Figure 1 shows the visualization of the grid search on the Classic400 data set. We set $K = 3$. Lowest harmonic mean at 0.0043 with $\alpha = 0.1$ and $\beta = 2$. In general, the increase in α and β lead to increase in harmonic mean. This fits our understanding of α and β as smoothing factors in the model.

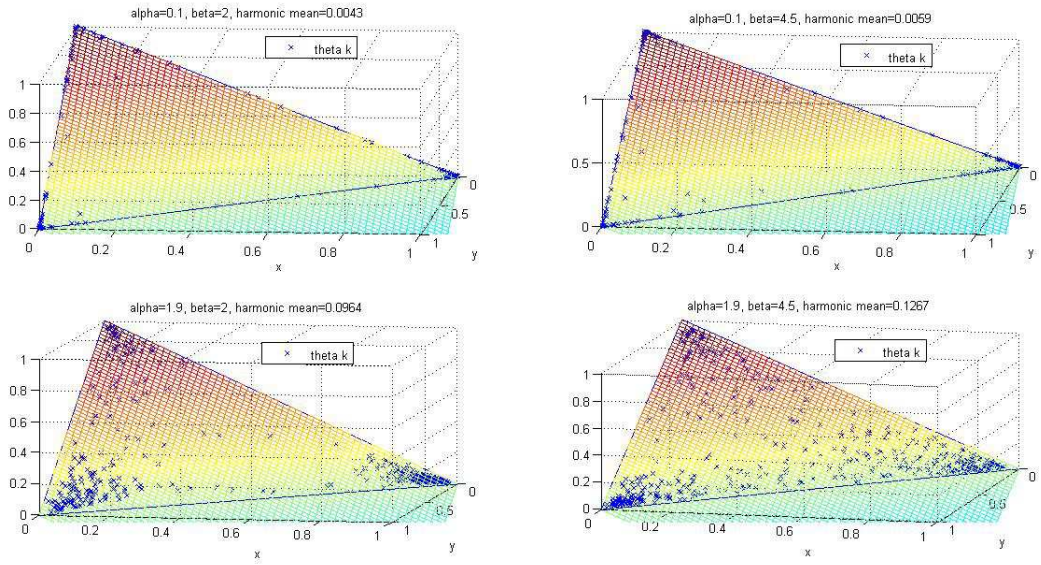

Figure 1: Performance for topic model of choosing different α and β. Best result (the top left figure) is $\alpha = 0.1$ and $\beta = 2$.

3.2 Design

We perform clustering with LDA on two data sets: Classic400 (a subset of the Classic3 dataset) and Psychological Review Abstracts data provided by Griffiths and Steyers (2005). We obtain the second data set as a

Dataset	V	M	$\bar{N}$	SD
Classic400	6205	400	78.79	45.5738
Psychological Review Abstract	9244	2000	66.3786	20.7286

Table 1: Data sets used in experiments. V is vocabulary size. M is the number of documents. $\bar{N}$ is the mean document length, SD is the standard deviation of the document length.

Topic 1 -'medical'	Topic 2 -'scientific methods'	Topic 3 - 'aero-physics'
patients	system	boundary
ventricular	retrieval	layer
heart	scientific	wing
volume	cases	mach
cardiac	nickel	supersonic
ventricle	language	ratio

topic 1	topic 2	topic 3	topic 4	topic 5
memory	culture	visual	evidence	drug
recognition	japanese	motion	speech	tolerance
item	frequent	spatial	sex	stress
recall	individuals	perception	human	aggression
item	interactionist	system	gender	fight
list	attitudes	objects	child	immune
retrieval	attitudes	image	presented	inaction

Table 2: Top common occurred words for some topics inferred from the (top) Classic400 and (bottom) Psychological Review training data by LDA.

Matlab-formatted file. It has been preprocessed with the stop-words list of size 465. Punctuation letters are treated as word separators. Table 1 describes the statistics of both the datasets.

In both experiments, we use $\alpha = 0.1$ and $\beta = 2$. We use $K = 3$ and $K = 50$ for the first and second data set correspondingly.

3.3 Results

One informal, but important, measure of correctness and the success of the topic model is whether the most likely words in the topic forms a coherent patterns according to human understanding. Table 2 shows the top commonly occurred words from a selection of topics for both corpora. For the Classic400 data set, we can map the words into particular "natural" topics. For example, medical for topic 1, scientific methods for topic 2, and aero-physics for topic 3. For the Psychology Review abstracts data set, the top words could be mapped to sub-fields of psychology, for example, social psychology as topic 2, developmental psychology as topic 4, and clinical psychology for topic 5.

4 Conclusion

In this project, we implement the collapsed Gibbs sampling to learn the Latent Dirichlet Allocation topic model. Two data sets, Classic400 and Psychological Review Abstract, are used to see the results of the learned model. We also look at how to evaluate the topic models in different setting.

References

[1] Blei, D., Ng, A. & Jordan, M. (2003). Latent Dirichlet allocation. J. Machine Learning Res., 3, 993-1022.

[2] Hanna Wallach, Iain Murray, Ruslan Salakhutdinov & David Mimno. "Evaluation Methods for Topic Models." Presented at the Learning Workshop (Snowbird), Clearwater, Florida, 2009.

[3] G. Doyle & C. Elkan In Proceedings of the 26th International Conference on Machine Learning (ICML), July 2009

[4] Griffiths,, T. & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences, 104, 5228-5235.

[5] Li, W. & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. Proceedings of the 23rd International Conference on Machine Learning (pp. 577-584).

[6] Heinrich, G. (2005). Parameter estimation for text analysis. Technical report, vsonix GmbH and University of Leipzig, Germany. Available at http://www.arbylon.net/publications/text-est.pdf.

[7] Michael A. Newton & Adrian E. Raftery (1991). Approximate Bayesian Inference by the Weighted Likelihood Bootstrap. Technical Report NO.199, March 1991.