Topic Classification using Latent Dirichlet Allocation

Adrian Guthals (aguthals@cs.ucsd.edu),
David Larson (dplarson@ucsd.edu),

CSE 250B: Project #3
University of California, San Diego

February 28, 2013

Abstract

Collapsed Gibbs Sampling is evaluated as training algorithm for Latent Dirichlet Allocation (LDA), an unsupervised learning model. LDA models are trained to classify text documents from two datasets: Classic400, a colelction of English documents from three research areas (library science, aeronautics, and medicine); and KOS, a collection of English blog posts from dailykos.com. For both datasets, the documents were classified into three topics and the LDA is evaluated based on the clustering of documents and lists of the most commonly occurred words for each topic. LDA of the Classic400 dataset showed distinct clustering of topics, and the list of most common words for each topic were cohesive (topic 1: system, research, scientific; topic 2: boundary, layer, wing; topic 3: patients, ventricular, fatty). Meanwhile, LDA of the KOS dataset show less distinct clustering and the list of most common words for each topic showed overlap, specifically, all topics contained words related to the 2008 US President campaign and election (topic 1: bush, kerry, war; topic 2: republican, race, house; topic 3: november, kerry, voting). The poor results of the KOS dataset indicate a lack of variety in the the studied dataset, which is a 400 document subset of the original 3430 document KOS dataset, rather than an issue with the LDA or Gibbs Sampling method. Overall, Gibbs Sampling proved to be a good learning algorithm for LDA models.

1 Introduction

1.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is an unsupervised learning model for determining topic classification of documents based on their vocabulary. It assumes that the generation of the documents was a probalistic process, and therefore, the parameters used in the generation can be learned based on the results. Specifically, LDA assumes thegeneration used the Dirichlet distribution, which is defined as a probability density function of all multinomial parameter vectors. Mathematically it is defined as:

$$p(\gamma|\alpha) = \frac{1}{D(\alpha)} \prod_{s=1}^{m} \gamma_s^{\alpha_s - 1} \tag{1}$$

where α is the parameter vector of length m and D a normalizing constant

$$D(\alpha) = \int_{\gamma} \prod_{s=1}^{m} \gamma_s^{\alpha_s - 1} \tag{2}$$

D can also be defined explicitly as

$$D(\alpha) = \frac{\prod_{s=1}^{m} \Gamma(\alpha_s)}{\Gamma(\sum_{s=1}^{m} \alpha_s)} \tag{3}$$

The LDA generative process is defined in [1] as:

1. Given: Dirichlet distribution using α (parameter vector length K)
2. Given: Dirichlet distribution using β (parameter vector length V)
3. for topic 1 to K:
   (a) draw multinomial distribution using ϕ_k based on β
4. for document 1 to M:
   (a) draw multinomial θ based on α
   (b) for each word in document
      i. draw topic z using θ
      ii. draw word w using ϕ_z

1.2 Collapsed Gibbs Sampling

Collapsed Gibbs Sampling is a learning method for indirectly determining θ for each document and ϕ_k for each topic by inferring the hidden value z for each word in each document [1]. Learning of θ and ϕ_k is based on:

$$p(z_i = j|\bar{z}', \bar{w}) \propto \frac{q'_{jw_i} + \beta_{w_i}}{\sum_t q'_{jt} + \beta_t} \frac{n'_{mj} + \alpha_j}{\sum_k n'_{mk} + \alpha_k} \tag{4}$$

where $\bar{w}$ is the sequence of words in the vocabulary.

2 Design and Analysis of Algorithms (Adrian)

We implemented the LDA generative process as outlined in [1]. First we randomly initialized an array z to hold a single topic assignment for each of the T total word occurences in the corpus. Then the following arrays were populated based upon these topic assignments: q_{jw} being a KXV array specifying the count of how many times the word w occurs with topic j in the whole corpus; Q_j being a Kx array specifying the the count of how many times a word occured with topic j in the whole corpus; n_{mj} being a MxK array specifying the the count of how many times a word within document m occured with topic j; N_m being a $Mx1$ array specifying the the count of how many word occurences are in each document m.

For each iteration of the Gibbs Sampling, z_i was populated for each word occurence $i = 1, 2, ..., T$. If $j' = z_i$ was the previous topic assignment for word i, then array positions that counted word i with topic j were decremented: $q_{j'w_i} - = 1$, $n_{mj'} - = 1$, $Q_{j'} - = 1$, $N_m - = 1$. Then the new value of z_i was chosen by sampling a multinomial distribution with parameters $p(z_i = j|\bar{z}', \bar{w})$ from the following equation. Given the new assignment $j'' = z_i$, correspnding array counts were then incremented: $q_{j''w_i} + = 1$, $n_{mj''} + = 1$, $Q_{j''} + = 1$, $N_m + = 1$.

$$p(z_i = j|\bar{z}', \bar{w}) \propto \frac{q_{jw_i} + \beta}{Q_j + V\beta} \frac{n_{mj} + \alpha}{N_m + K\alpha} \tag{5}$$

Dataset	Documents	Vocabulary	α	β
Classic400	400	6205	0.01	0.1
KOS	400	6906	0.01	0.1

Table 1: Composition of the two datasets used in this study and the chosen Gibbs hyperparameters α and β. Note that the KOS dataset used is a reduced version of the original KOS dataset, which contains 3430 documents.

Convergence was reached after a predetermined number of iterations (100). At this point, ϕ and θ values were output by the following equations:

$$\phi_{kw} = \frac{q_{kw} + \beta}{Q_j + V\beta} \tag{6}$$

$$\theta_{mk} = \frac{n_{mj} + \alpha}{N_m + K\alpha} \tag{7}$$

3 Design of Experiments

3.1 Datasets

Two datasets were classified using LDA: Classic400, a collection of English documents from three research areas (aeronautics, medicine, and library science); and KOS, a collection of English blog posts from dailykos.com (see Table 1 for details on their sizes) [2, 3]. While the Classic400 dataset provided the true labels of its source topics, the KOS dataset provided no such information. Also, to reduce run times, we elected to use a reduced version of the KOS dataset containing only the first 400 of the original 3430 documents.

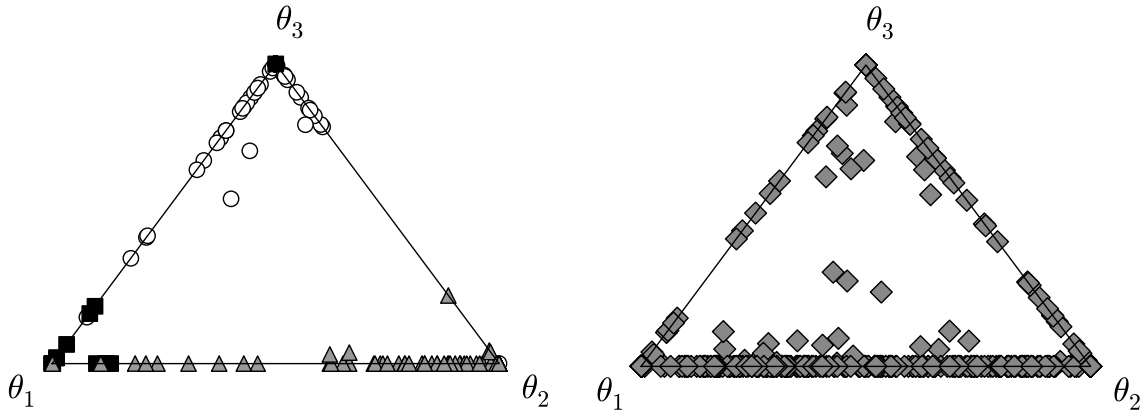
3.2 Hyperparameters

For both datasets we apply LDA with three topics ($K = 3$). Since α is postively correlated to the number of topics present per document and we are only considering three topics, we chose values of $\alpha << 1$, which implies that every document contains only a few topics. Likewise, we chose values of $\beta << 1$, which implies that every word in every document corresponds to a small number of topics. Previous studies have looked into the effect of α and β on the cluserting results for LDA models, some going so far as to suggest likely values based on the number of topics in a dataset (see [4]).

For this study we chose to use a grid search method to determine the optimal values of α and β, anticipating that both values will be much less than one. We trained the LDA model using Gibbs sampling for α and β values of $10^2, 10^1, ..., 10^{-4}$, and evaluated the values based on the coherence of each topics' most commonly occurred words. From this grid search we chosen to set $\alpha = 0.01$ and $\beta = 0.1$ for both datasets (see Table 1).

3.3 Stopping Criteria of Gibbs Sampling

For both datasets we chose the Gibbs Sampling stopping criteria as 100 epochs. We tried 50, 75, ..., 200 epochs as the stopping criteria, but found that the Euclidean distance between the ϕ vectors to not change signifcantly for more than 100 epochs.



(a) Classic400 (b) KOS

Figure 1: Clustering of documents based based on their θ_1, θ_2 and θ_3, projected on a triangular simplex. Documents in Classic400 are labeled based on their true labels (topic 1: white circles with black edges, topic 2: black solid squares, topic 3: gray solid triangles) while documents in KOS are have the same marker label as the true labels are unknown.

3.4 Goodness of Fit

A sensible way to define goodness of fit for a given dataset is to compare lists of most commonly occurred words for each topic. A more quantitative method would involve statistically analyzing the clustering of each document classification.

3.5 Overfitting

As with many models, LDA is susceptible to overfitting. One approach to determine overfitting of a LDA model is to calculate the perplexity of the model (see [5]).

3.6 Results

3.6.1 Document Topic Clustering

Figure 1 shows the clustering of documents relative to each topic as projected onto a triangular simplex, where the each vertex indicates $\theta = 1$ for the denoted topic.

3.6.2 Most Common Words Per Topic

If a LDA model is setup to classify more than 4 topics, then a simplex plot will not be a viable method for checking the clustering of the results. Instead, lists of the most commonly occurred words per topic can be analyzed to gauge the performance of the LDA. Table 2 shows the most commonly occurred words for each topic of the two datasets. The most common words from Classic400 for each topic are all related and correspond well to the true topic labels (topic 1: medicine, topic 2: aeronautics, and topic 3: library science). Meanwhile, the most common words for KOS for each topic seem to overlap and be derived from same general topic, namely politics, or more specifically, the 2008 US Presidential campaign between George Bush and John Kerry. As the source of KOS is a political blog site, it is not surprising that politics shows up prominetly as a topic. However, the lack of diversity between the most common words of each topic and therefore lack of unique topics is more likely the result of selecting only the first 400 documents from the originald dataset rather than an issue with our LDA model implementation.

	Classic400			KOS		
Rank	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 1**	**Topic 2**	**Topic 3**
1	system	boundary	patients	bush	republican	november
2	research	layer	ventricular	kerry	race	kerry
3	scientific	wing	fatty	war	house	voting
4	retrieval	mach	cases	iraq	party	vote
5	problems	supersonic	nickels	people	elections	bush
6	language	wings	left	president	campaign	polls
7	science	ratio	acids	general	democratic	poll
8	methods	velocity	aortic	voters	state	governor
9	systems	effects	blood	administration	democrats	senate
10	subject	shocks	normal	time	senate	republicans

Table 2: Ten most commonly occured words for each topic classification for the classic400 and KOS datasets. Rank indicates how frequently a word appears in each topic, with 1 being the most occurred.

4 Findings and Lessons Learned (Adrian)

The simplex plot of the Classic400 topic clustering showed that the LDA performed well on the dataset. We already new the number of categories and did not have to pick K. It could be interesting to tweak this parameter as it could lead to more precise clustering by picking sub-groupings within each cluster. In the KOS dataset for example, we chose $K = 3$, but analysis of the simplex plots reveals that the documents likely fit into just two topics. But setting $K = 2$ might also influence the α and β values as well and would have required a separate grid search over these values, which would have proved difficult since there was no correct grouping of the KOS data known in advance. To better train α and β for the KOS data, we would require a set of blog posts with a similar number of defined topics.

References

[1] C. Elkan, "Text mining and topic models," February 2013.

[2] "Classic400 dataset." [Online]. Available: ftp://ftp.cs.cornell.edu/pub/smart/

[3] A. Frank and A. Asuncion, "UCI machine learning repository," 2010. [Online]. Available: http://archive.ics.uci.edu/ml

[4] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, pp. 5228–5235, 2004. [Online]. Available: http://www.pnas.org/content/101/suppl.1/5228.abstract

[5] G. Heinrich, "Parameter estimation for text analysis," http://www.arbylon.net/publications/text-est.pdf, vsonix GmbH and University of Leipzig, Germany, Tech. Rep., 2005.