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 27, 2013

Abstract

LDA, Gibbs sampling, topic classification of documents, datasets used, results and their meaning, conclusions

1 Introduction

Introduce topic classification of documents. Then transition into formal definitions of LDA and Gibbs Sampling.

Elkan's lecture notes [1]

1.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a multinomial

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

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

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

1.2 Gibbs Sampling

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

2 Design and Analysis of Algorithms

Discuss how we're implementing LDA and Gibbs Sampling.

Dataset	**Documents**	**Vocabulary**	α	β
Classic400	400	6205	0.01	1.0
KOS	400	6906	1.0	0.01

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.

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 $\alpha \leq 1$, which implies that every document contains only a few topics. Likewise, we chose values of $\beta \leq 1$, 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 clusterting results for LDA models, some going so far as to suggest likely values based on the number of topics in a dataset (see [4]). However, for the purposes of this study our selection criteria of α and β is only that they are both less than or equal to one, and we have therefore chosen to omit any sort of optimization of their values (see Table 1).

3.3 Convergence of Gibbs

When do we decide to stop Gibbs

3.4 Results

3.4.1 Clustering

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

3.4.2 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.

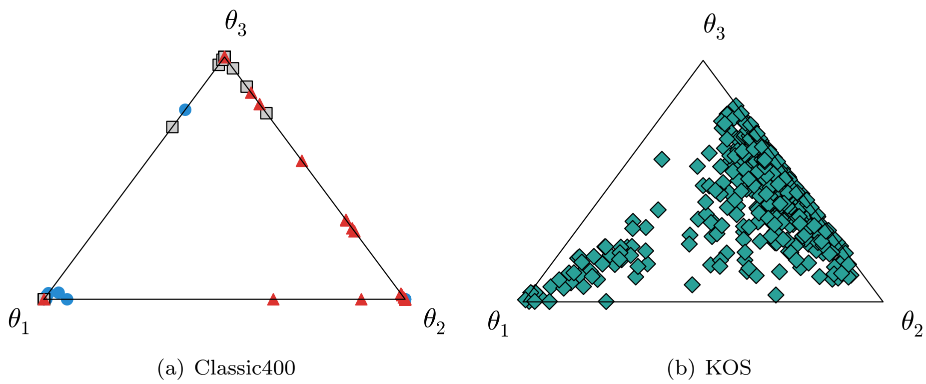

Figure 1: Clustering of documents based on their θ_1, θ_2 and θ_3, projected on a triangular simplex. Documents in Classic400 are labeled based on their true labels (topic 1: blue solid circles, topic 2: black empty squares, topic 3: red solid triangles) while documents in KOS are all labeled the same way as the true labels are unknown.

	Classic400			KOS		
Rank	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 1**	**Topic 2**	**Topic 3**
1	patients	boundary	system	november	bush	kerry
2	ventricular	layer	retrieval	kerry	campaign	time
3	cases	wing	scientific	vote	republican	bush
4	fatty	mach	research	voting	race	people
5	left	supersonic	language	polls	war	jul
6	acids	wings	systems	house	general	party
7	nickel	velocity	journals	poll	elections	john
8	aortic	ratio	field	bush	percent	democratic
9	blood	shock	methods	senate	state	media
10	glucose	effects	subjects	governor	news	voters

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 (1=most occurred).

4 Findings and Lessons Learned

Thoughts on: LDA as a model, Gibbs Sampling as a training method, performance issues, results of the experiments

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