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

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

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

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

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
Classic400	400	6205
KOS	400	6906

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

Dataset	α	β
Classic400	16.67	0.01
KOS	16.67	0.01

Table 2: Hyperparaters used in training the LDA model using Gibbs sampling.

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]. 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

$\alpha = 50/K$ and $\beta = 0.01$ where K is the number of topics. For all experiments we set $K = 3$ and therefore $\alpha = 16.67$ (see Table 2). Suggested originally by Griffiths and Steyvers (2004). May want to try other values.

Recall: large α for many topics per document and large β for many topics per word. We only use 3 topics so α will probably be small.

3.3 Convergence of Gibbs

When do we decide to stop Gibbs

3.4 Results

3.4.1 Clustering

Clustering of documents into three topics (reference simplex plots).

3.4.2 Most Common Words Per Topic

Ten most common words (in tables).

4 Findings and Lessons Learned

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

	Classic400			KOS		
Rank	**Topic 1**	**Topic 2**	**Topic 3**	**Topic 1**	**Topic 2**	**Topic 3**
1	patients	boundary	wing			
2	case	layer	mach			
3	ventricular	velocity	supersonic			
4	system	field	effects			
5	research	solution	ratio			
6	scientific	plate	wings			
7	fatty	problem	shock			
8	nickel	free	numbers			
9	acids	heat	jet			
10	aortic	cylinder	lift			

Table 3: Ten most commonly occured words for each topic classification for the classic400 and KOS datasets. Rank corresponds to how frequently a word appears in each topic (1=most occurred).

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