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.

Topic 1	Topic 2	Topic 3
patients	boundary	wing
case	layer	mach
ventricular	velocity	supersonic
system	field	effects
research	solution	ratio
scientific	plate	wings
fatty	problem	shock
nickel	free	numbers
acids	heat	jet
aortic	cylinder	lift

Table 1: Ten most commonly occured words for each topic classification of the classic 400 and KOS datasets.

3 Design of Experiments

3.1 Datasets

Two datasets: classic400 and KOS blog posts (from UCI Machine Learning database). KOS dataset was reduced from 3430 documents to 400 documents (vocabulary unchanged from original 6906 words).

3.2 Convergence of Gibbs

When do we decide to stop Gibbs

3.3 Results

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.