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 23, 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.

3 Design of Experiments

3.1 Datasets

Two datasets: classic400 and a second source. Where are they both from, where is the data derived from and how, and why did we chose both for the experiment (plus any pre-processing we may have done).

3.2 Convergence of Gibbs

When do we decide to stop Gibbs

3.3 Results

3.3.1 Classic 400

3.3.2 Second source

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.