Topic Modelling using Latent Dirichlet Allocation

CSE 250B Project 3

Suvir Jain, Gaurav Saxena

27 February,2014

Abstract

Abstract about LDA, our data, brief results.

1 Introduction

2 Framework

2.1 Dataset

Describe dataset

2.2 Model

Basic Theory of LDA

2.2.1 Inference Algorithm for Linear Chain CRF

How do we interpret results - Theta and Phi. Mention that theory here.

2.2.2 Training Methods for LDA

Theory of Gibbs Sampling and Collapsed Gibbs Sampling

Gibb's Sampling

Collins Perceptron

3 Design and Analysis of Algorithms

3.1 LDA

3.2 Collapsed Gibb's Sampling

4 Design of Experiments

4.1 Dataset Preprocessing

4.2 Expt 1

4.3 Expt 2

4.4 Implementation

Implementation specific notes

4.5 Code Optimization

4.6 Sanity Checks

Comparison with true labels. Can make a table here. Summation of n should be equal to q vector should be equal to classic400.

5 Results of Experiments

5.1 Grid Search for Stochastic Gradient Ascent

5.2 Accuracy for Stochastic Gradient Ascent

5.3 Collins Perceptron

6 Findings and Lessons Learned

6.1 Goodness of Fit

ADDRESS THE FOLLOWING

In the report, try to answer the following questions. The questions are related to each other, and do not have definitive answers. 1. What is a sensible way to define the goodness-of-fit, for the same dataset, of LDA models with different hyperparameters K, ALPHA, and BETA? 2. Given the definition of goodness-of-fit, is it possible to compute it numeri- cally, either exactly or approximately? 3. How can we determine whether an LDA model is overfitting its training data? For the two datasets with which you do experiments, present and justify good values for K, ALPHA and BETA. You can choose these values informally (you do not need an automated algorithm) but your choices should be sensible and justified.