CSE 250B Project 3

For this project you must work in a team of exactly two students. The joint report for your team must be submitted in hard copy at the start of the lecture on Tuesday February 26, 2013. You only have two weeks for this project, so do start work immediately.

The objective of the project is to understand Gibbs sampling as a training method for latent Dirichlet allocation (LDA) models. First, implement the Gibbs sampling training algorithm as described in class and in the online lecture notes. (See also the explanations in [Heinrich, 2005].) Making the inner loop of your LDA code fast is a challenge, but is doable. Second, for an LDA model trained on a collection of documents, write code to print the words that have highest probability for each topic. Third, write code to create a visualization of the documents based on the topics of the trained model. In Matlab, it is easy to allow a user to rotate a 3D visualization interactively, to see it from multiple points of view.

Do experiments with your LDA implementation on two datasets. The first is the Classic400 dataset, available at `http://cseweb.ucsd.edu/users/elkan/151/classic400.mat` in Matlab format. This dataset consists of 400 documents over a vocabulary of 6205 words. Each matrix entry is how many times a given word appears in a given document.

With LDA, one can learn any number of topics on a dataset. Ideally, each topic is meaningful and not redundant. One way to investigate meaningfulness is to look at the highest-probability words for each topic. These should be words that are semantically related in a way that a human can understand. Separately, when you use the θ vectors to plot the documents in Euclidean space, ideally the true groups of documents will be distinct, assuming that ground-truth groups are known. For the Classic400 dataset, the array named `truelabels` specifies which of three domains each document comes from. Note that LDA does not use this array in any way. It is provided only so that you can investigate whether learned topics and θ vectors are correlated with ground-truth labels.

The second dataset should be one that you select, obtain, and preprocess. It may be an interesting collection of documents. Or, find a non-text dataset for which the LDA model is appropriate, and explain why. In any case, use your

code to train LDA on the dataset, present a selection of results in your report, and explain why these results are interesting and believable.

In your 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 and compute the goodness-of-fit, for a given dataset, of LDA models with different hyperparameters K, α, and β?

2. How can you 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, α and β. You can choose these values informally (you do not need an automated algorithm) but your choices should be sensible and justified.

References

[Heinrich, 2005] Heinrich, G. (2005). Parameter estimation for text analysis. Technical report, vsonix GmbH and University of Leipzig, Germany. Available at `http://www.arbylon.net/publications/text-est.pdf`.