

## 4F13 Machine Learning: Coursework #3: Latent Dirichlet Allocation

Zoubin Ghahramani

Due: 4pm Dec 7th, 2015 via Moodle (or to Division F office, room BNO-37)

In this assignment, we will give you two short pieces of matlab code, which implement the main ingredients of Gibbs sampling for a Mixture of Multinomials `bmm.m` and for LDA `lda.m`. Before you start answering questions, you should spend some time understanding in detail, what this code does. This will enable you to answer all the questions with very little programming effort on your part.

Your answers should contain an explanation of what you do, and 2-4 central commands to achieve it (but complete listings are unnecessary). You must also give an interpretation of what the numerical values and graphs you provide *mean* – why are the results the way they are? Hand in a maximum of 5 pages; each question should be labelled and answered separately and distinctly.

- a) 10% : load the data from `kos_doc_data.mat`. The word counts are in the matrix variables `A` and `B` for training and testing respectively, both matrices with 3 columns: document ID, word ID and word count. The words themselves are the variable `V`, such that eg. `V(841) = 'bush'`. How many documents, how many words and how many unique words are there in `A`, in `B` and in the union of `A` and `B`?
- b) 10% : Using the training data in `A`, find the maximum likelihood multinomial over words, and show the 20 largest probability items in a histogram. You may use the `barh` command followed by `set(gca, 'YTickLabel', V(s), 'Ytick', 1:20)`, where `s` is an array of appropriate indices.
- c) 10% : Using the model from question b), what will the test set log probability be if the test set `B` contains a word which is not contained in the training set `A`? Explain the implications of this.
- d) 10% : Instead of the maximum likelihood fit in question b), do Bayesian inference using a symmetric Dirichlet prior with a concentration parameter  $\alpha = 0.1$  on the word probabilities. What is the expression for the predictive distribution?
- e) 10% : What is the log probability for the test document with ID 2001? Explain whether, when computing the log probability of a test document, you would use the multinomial with or without the “combinatorial factor”. What is the per-word perplexity for the document with ID 2001? What is the per-word perplexity over all documents in `B`?
- f) 10% : What would the perplexity be for a uniform multinomial? Compare this value to the previously computed perplexity and explain.
- g) 10% : The `bmm.m` script implements Gibbs sampling for a mixture of multinomials model. Use and modify the script to plot the evolution of the mixing proportions as a function of the number of Gibbs sweeps up to 10 iterations. The mixing proportions are the posterior probabilities of each of the mixture components. Compute perplexity for the final state reached after 10 Gibbs sweeps, and compare to perplexities from question e) and f).
- h) 10% : Explain carefully, what does it mean for an MCMC method to converge to the stationary distribution? Does the Gibbs sampler from question g) converge? Restart with different random seed. Does the Gibbs sampler explore the posterior distribution? Why/why not? Explain.
- i) 10% : Use and modify `lda.m`. Plot topic posteriors as a function of the number of Gibbs sweeps, up to 10 sweeps. Comment on these. Compute the perplexity for the documents in `B` for the state after 10 Gibbs sweeps, and compare to previously computed perplexities. Are 10 Gibbs sweeps adequate?
- j) 10% : For LDA, plot the word entropy for each of the topics as a function of the number of Gibbs sweeps. Explain what you see.
- k) 0% : Explore how performance depends on the number of Gibbs sweeps and how performance depends on the number of topics, `K`.