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

Carl Edward Rasmussen

Due: 12:00 noon, Dec 6th, 2019 online via moodle

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? **Each question should be labelled and answered separately and distinctly.** Total combined length of answers must not exceed 1000 words; clearly indicate the actual total number of words in your coursework. All questions carry approximately equal weight.

In this assignment, we will give you two short pieces of matlab or python code, which implement the main ingredients of Gibbs sampling for a Mixture of Multinomials `bmm` and for LDA `lda`. 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.

The data is in the file `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`.

d n N_d

a) 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. For that multinomial model, what is the highest and lowest possible test set log probability (for any possible test set)? Explain the implications of this.

b) Instead of the maximum likelihood fit in question a), do Bayesian inference using a symmetric Dirichlet prior with a concentration parameter α on the word probabilities. Compare the expressions for the predictive word probabilities for these two types of inference, and explain the implications, both for common and rare words for both small and large values of α.

c) For the Bayesian model, 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 or the categorical distribution function? What is the per-word perplexity for the document with ID 2001? What is the per-word perplexity over all documents in `B`? Explain why the perplexities are different for different documents? What would the perplexity be for a uniform multinomial?

d) The `bmm` 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 50 iterations. The mixing proportions are the posterior probabilities of each of the mixture components. Explain carefully, how would you determine whether the Gibbs sampler converges to and explores the stationary distribution (the posterior), does it?

e) Use and modify `lda`. Plot topic posteriors for $\mathtt{K} = 20$ as a function of the number of Gibbs sweeps, up to 50 sweeps. Comment on these. Compute the perplexity for the documents in `B` for the state after 50 Gibbs sweeps, and compare to previously computed perplexities. Are 50 Gibbs sweeps adequate? Plot the word entropy (what units do you use?) for each of the topics as a function of the number of Gibbs sweeps. Explain what you see.

Note that the performance and learning time for LDA depends a lot on the number of topics `K`.