TDA 231 Machine Learning: Homework 4

Goal: LDA & Gibbs sampling
Grader: Fredrik Johansson

Due Date: February 22, 2015

General guidelines:

1. All solutions to theoretical problems, and discussion regarding practical problems, should be submitted in a single file named *report.pdf*
2. All matlab files have to be submitted as a single zip file named *code.zip*.
3. The report should clearly indicate your name, personal number and email address
4. All datasets can be downloaded from the course website.
5. All plots, tables and additional information should be included in *report.pdf*

1 Theoretical problems

Problem 1.1 [Gibbs sampling from a 2D Gaussian, 5 points]

Suppose $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = [1, 1]^T$ and $\boldsymbol{\Sigma} = [1, -0.5; -0.5, 1]^T$. In this problem you will design a Markov Chain and a Gibbs sampler to sample from this distribution (even though it can be done directly too).

(a) Give the state space of the Markov chain.

(b) Give the transition rule for the Gibbs sampler by deriving the conditional distributions $p(x_1|x_2)$

(c) State how to obtain samples from the marginal distribution $p(x_1)$ using this Markov chain.

2 Practical problems

Introduction

Latent Dirichlet Allocation (LDA) [1] is a probabilistic topic model used to automatically associate documents with abstract topics, or more generally - to cluster data. It is built on the assumption that documents are made up of a number of different topics, each associated with its own set of words. Here, we will follow the notation of the lecture slides.

We will consider what is referred to as the *smoothed* LDA in the original paper, used in the lecture slides. This simply means that each topic is assigned a Dirichlet prior of its own. That is, the probabilistic model is,

- $\theta \sim \text{Dirichlet}(\alpha)$
- $\beta \sim \text{Dirichlet}(\eta)$
- $Z_i \mid \theta^{(d_i)} \sim \text{Categorical}(\theta^{(d_i)})$
- $W_i \mid z_i, \beta^{(z_i)} \sim \text{Categorical}(\beta^{(z_i)})$.

In general, statistical inference involves determining properties of random variables based on data or other random variables. In the case of LDA, this typically amounts to finding the topic distribution, θ_d for a certain document d and the word distribution $\beta_k$for a certain topic k.

Data specification

LDA is used typically used with corpora (collections of texts). Under the assumptions of LDA, a corpora is a set of D documents, where each document is a "bag of words", i.e. an unordered collection of words from a known vocabulary. We can represent each document as a list of numbers, one for each word w in the vocabulary, counting the number of times w occurs in the document. Typically, such a vocabulary exludes stop words such as *a, an, in, and* etc.

In this assignment, for each document, the number of times every word, from a specified vocabulary occurs, is counted and stored in the sparse format *word-id:count.* For example, a document with the text *The cat is in a hat* and the vocabulary *Cat, Dog, Hat, Car, Paper* will be stored as "*1:1 3:1*" (words are indexed from 1). The corpus[1] used in this assignment comes from the proceedings of Neural Information Processing Systems (NIPS), a leading Machine Learning conference[2].

Assignment

The assignment is partially a programming assignment and partially an exercise in analysing topic models. The programming part is done in MATLAB, for ease of implementation and because MATLAB is available on Chalmers computers. Furthermore, if you would like to do the assignment on your own machine, MATLAB is available freely for students via the Chalmers website.

Problem 2.1 [Implementation, 10 points]

Implement a Gibbs sampler[2] performing inference in LDA and run on the dataset included in the assignment, see *code_template.zip.* This, the programming part, includes

1. Initialization of count variables, initial sampling of topics
2. Sampling of topics, Z, and update of count variables

[1] http://archive.ics.uci.edu/ml/datasets/Bag+of+Words

[2] http://nips.cc/

3. Estimating β and θ

We will not estimate α and η, but will assume that they are known.

Important: The sampling of topics should actually sample, *not* take the most likely topic.

The solution to this part (an implementation of the Gibbs sampler) should be contained in the file *ldaGibbs.m*, provided as part of the assignment. The function *ldaGibbs* is run by the main function *ldaExercise*. In *ldaExercise.m*, the parameters for the model and the inference procedure are specified. These may be modified for testing purposes (lowering the number of Gibbs iterations for instance), but this is not required. The other files need not be modified, and this part of the assignment will be graded based on the contents of *ldaGibbs* alone.

In summary: to train LDA, run *ldaExercise.m*, after having implemented the missing parts of *ldaGibbs*.

By integrating out θ and β [2], forming a so-called *collapsed Gibbs sampler*, the topics Z can be sampled, in step 2, according to,

$$p(Z_i = k \mid Z_{\neg i}, W) \propto \frac{\eta + N_{k,\neg i}^{(w_i)}}{V\eta + \sum_{w'=1}^{V} N_{k,\neg i}^{(w')}} \frac{\alpha + M_{d,\neg i}^{(k)}}{T\alpha + \sum_{k'=1}^{T} M_{d,\neg i}^{(k')}} \tag{1}$$

where

- V is the number of words in the vocabulary and T is the number of topics.
- Z_i is the topic of word $i = (d, n)$ where d denotes a document and n the word index inside document d,
- $N_{k,\neg i}^{(w)}$ denotes the number of times word w has occurred with topic k, excluding word $i = (d, n)$,
- $M_{d,\neg i}^{(k)}$ denotes the number of times topic k has occurred with a word from document d, excluding word $i = (d, n)$.

Note that, in principle, different occurrences of the same word in one document could be assigned different topics! That is, word n refers to a single occurrence.

By the conjugacy of the multinomial and Dirichlet distributions, we can estimate β and θ by their expected values according to,

$$\beta_{k,w} = \frac{N_k^{(w)} + \eta_w}{\sum_{w'=1}^{V}(N_k^{(w')} + \eta_{w'})} \quad , \quad \theta_{d,k} = \frac{M_d^{(k)} + \alpha_k}{\sum_{k'=1}^{T}(M_d^{(k')} + \alpha_{k'})} \tag{2}$$

Problem 2.2 [Analysis, 5p] Because the dataset comes from a machine learning conference, you should expect topics to reflect different areas of machine learning. A correct solution typically includes topics that represent for example *neural networks*, *signal processing*, *probability* and *neuroscience* respectively. As an example, the top words of one of the topics could be (neuron, cell, input, network, neural, signal, system, visual, connection, layer).

Analyze your results and discuss the following,

(a) Explain how we arrive at Equation (1). (No complete derivation required)

(b) What do the values of β and θ represent?

(c) For each individual topic, which words are ranked as most probable? Can you see distinct topics?

(d) How does the number of topics influence the result?

(e) Why can't we expect to come up with an exact inference algorithm for LDA?

(f) Why doesn't it matter so much that we can't?

Deliverables

You should hand in the completed *ldaGibbs.m* and *ldaExercise.m* (modified or otherwise) and answers to the questions in part B, complying with the general guidelines at the top.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[2] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.