Homework 5: EM for a Simple Topic Model

Writeup due 23:59 on Friday 17 April 2015

You will do this assignment individually and submit your answers as a PDF via the Canvas course website. There is a mathematical component and a programming component to this homework. Do not submit code.

In this homework, you will implement a very simple kind of topic model. Latent Dirichlet allocation, as we discussed in class, is a topic model in which each document is composed of multiple topics. Here we will make a simplified version in which each document has just a single topic. As in LDA, the vocabulary will have V words and a topic will be a distribution over this vocabulary. Let's use K topics and the kth topic is a vector $\boldsymbol{\beta}_k$, where $\beta_{k,v} \geq 0$ and $\sum_v \beta_{k,v} = 1$. Each document can be described by a set of word counts $\boldsymbol{w}_d$, where $w_{d,v}$ is a nonnegative integer. Document d has N_d words in total, i.e., $\sum_v w_{d,v} = N_d$. Let's have the unknown overall mixing proportion of topics be $\boldsymbol{\theta}$, where $\theta_k \geq 0$ and $\sum_k \theta_k = 1$. Our generative model is that each of the D documents has a single topic $z_d \in \{1, \ldots, K\}$, drawn from $\boldsymbol{\theta}$; then, each of the words is drawn from $\boldsymbol{\beta}_{z_d}$.

1. Complete Data Log Likelihood [4pt]

Write the complete-data log likelihood $\ln p(\{z_d, \boldsymbol{w}_d\}_{d=1}^D \mid \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^K)$. It may be convenient to write z_d as a one-hot coded vector $\boldsymbol{z}_d$.

2. Expectation Step [5pts]

Introduce estimates $q(\boldsymbol{z}_d)$ for the posterior over the hidden variables $\boldsymbol{z}_d$. What did you choose and why? Write down how you would determine the parameters of these estimates, given the observed data $\{\boldsymbol{w}_d\}_{d=1}^D$ and the parameters $\boldsymbol{\theta}$ and $\{\boldsymbol{\beta}_k\}_{k=1}^K$.

3. Maximization Step [5pts]

With the $q(\boldsymbol{z}_d)$ estimates in hand from the E-step, derive an update for maximizing the expected complete data log likelihood in terms of $\boldsymbol{\theta}$ and $\{\boldsymbol{\beta}_k\}_{k=1}^K$.

4. Implementation [10pts]

Implement this expectation maximization algorithm and try it out on some text data. You may have to do a little preprocessing. Try different numbers of topics. Report what topics you find by, e.g., listing the most likely words. Some potential data sets to explore are:

- 20 Newsgroups:
  http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html

- NSF Award Abstracts:
  http://kdd.ics.uci.edu/databases/nsfabs/nsfawards.html
- Reuters:
  http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html
- Any other interesting data set you feel like tracking down or scraping!

5. Calibration [1pt]

Approximately how long did this homework take to complete?

Changelog

- **v1.0** – 9 April 2015 at 23:40