Your Name
`email@fas.harvard.edu`
CS181-S16

Assignment #5
Due: 5:00pm April 15, 2016

Collaborators: John Doe, Fred Doe

Homework 5: EM for a Simple Topic Model

There is a mathematical component and a programming component to this homework. Please submit ONLY your PDF to Canvas, and push all of your work to your Github repository. If a question requires you to make any plots, please include those in the writeup.

Background: 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 $z_d \in \{1, \dots, K\}$, drawn from $\boldsymbol{\theta}$; then, each of the words is drawn from $\boldsymbol{\beta}_{z_d}$.

Problem 1 (Complete Data Log Likelihood, 4 pts)
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 $\boldsymbol{z}_d$.

Solution

Problem 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 $\{\boldsymbol{\beta}_k\}_{k=1}^{K}$.

Solution

Problem 3 (Maximization Step, 5pts)

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

(a) Derive an expression for the expected complete data log likelihood for fixed γ's.

(b) Find a value of θ that maximizes the expected complete data log likelihood derived in (a). You may find it helpful to use Lagrange multipliers in order to force the constraint $\sum \theta_k = 1$. Why does this optimized θ make intuitive sense?

(c) Apply a similar argument to find the value of $\beta_{k,v}$ that maximizes the expected complete data log likelihood.

Solution

Problem 4 (Implementation, 10pts)

Implement this expectation maximization algorithm and try it out on some text data. In order for the EM algorithm to work, you may have to do a little preprocessing.

The starter code loads the text data as a numpy array that is 5224951×3 in size. As shown below, the first number in the numpy array represents the document_id, the second number represents a word_id, and the third number is the count the word appears.

$$[\text{doc\_id, word\_id, count}]$$

A dictionary of the mappings between word_ids and words is also provided. The full dataset description can be found at `http://kdd.ics.uci.edu/databases/nsfabs/nsfawards.data.html`.

Plot the objective function as a function of iteration and verify that it never increases. Try different numbers of topics and report what topics you find by, e.g., listing the most likely words.

Solution

Problem 5 (Calibration, 1pt)
Approximately how long did this homework take you to complete?