Yohann Smadja
smadja@post.harvard.edu
CS181-S16

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

Collaborators:

# 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 $k$th 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 $\boldsymbol{\beta}_{z_d}$.

**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

$$\ln p(\{z_d, \boldsymbol{w}_d\}_{d=1}^D \mid \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^K) = \ln \prod_{d=1}^{D} p(z_d, \boldsymbol{w}_d \mid \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^K)$$

since $z_d$ follows a $Categorical_K(\boldsymbol{\theta}_k)$, we get

$$\ln p(\{z_d, \boldsymbol{w}_d\}_{d=1}^D \mid \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^K) = \ln \prod_{d=1}^{D} \prod_{k=1}^{K} \left[\boldsymbol{\theta}_k p(\boldsymbol{w}_d \mid \{\boldsymbol{\beta}_k\}_{k=1}^K)\right]^{z_{dk}}$$

since $w_d$ follows a $Multinomial_{N_d}(\beta_{z_{dk}})$, we get

$$\ln p(\{z_d, \boldsymbol{w}_d\}_{d=1}^D \mid \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^K) = \ln \prod_{d=1}^{D} \prod_{k=1}^{K} \left[\boldsymbol{\theta}_k \frac{\Gamma(N_d+1)}{\prod_i \Gamma(w_{id}+1)} \prod_{i=1}^{N_d} (\beta_{ki}^{w_{id}})\right]^{z_{dk}}$$

$$\ln p(\{z_d, \boldsymbol{w}_d\}_{d=1}^D \mid \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^K) = \sum_{d=1}^{D} \sum_{k=1}^{K} z_{dk} \ln \left[\boldsymbol{\theta}_k \frac{\Gamma(N_d+1)}{\prod_i \Gamma(w_{id}+1)} \prod_{i=1}^{N_d} (\beta_{ki}^{w_{id}})\right]$$

we finally get:

$$\ln p(\{z_d, \boldsymbol{w}_d\}_{d=1}^D \mid \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^K) = \sum_{d=1}^{D} \sum_{k=1}^{K} z_{dk} \left[\ln \boldsymbol{\theta}_k + \ln \Gamma(N_d+1) - \sum_{i=1}^{N_d} \ln \Gamma(w_{id}+1) + \sum_{i=1}^{N_d} w_{vd} \ln \beta_{ki}\right]$$

**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

We have $q(z_{kd}) \propto p(z_{kd} = 1 \mid \theta) \prod_{i=1}^{N_d} p(w_{id} \mid z_{kd} = 1, \beta_{k=1..K})$

It's simply the prior times the likelihood.

We take $p(z_{kd} = 1 \mid \theta) = \theta_k$

$q(z_{kd}) \propto \theta_k \prod_{i=1}^{N_d} p(w_{id} \mid z_{kd} = 1, \beta_{k=1..K})$

$w_d$ is drawn from a $Multinomial(\beta_k)$:

$q(z_{kd}) \propto \theta_k \prod_{i=1}^{N_d} \beta_{ki}^{w_{id}}$

We then normalize and get:

$$q(z_d) = \left( \frac{\theta_k \prod_{i=1}^{N_d} \beta_{ki}^{w_{id}}}{\sum_{k=1}^{K} \theta_k \prod_{i=1}^{N_d} \beta_{ki}^{w_{id}}} \right)_{k=1...K}$$

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

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

(b) Find a value of $\theta$ 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 $\theta$ 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

a- In this 'Maximization' step, we suppose $\gamma$ given (from the 'Expectation' step) and fixed

$$\mathbb{E}_z\big[\ln p(\{z_d, \boldsymbol{w}_d\}_{d=1}^D \mid \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^K)\big] = \sum_{d=1}^{D}\sum_{k=1}^{K} \gamma_{dk} \ln p(\{z_d, \boldsymbol{w}_d\} \mid \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^K)$$

$$\mathbb{E}_z\big[\log p(\{z_d, \boldsymbol{w}_d\}_{d=1}^D \mid \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^K)\big] = \sum_{d=1}^{D}\sum_{k=1}^{K} \gamma_{dk}\Big[\ln\theta_k + \ln\Gamma(N_d+1) - \sum_{i=1}^{N_d}\ln\Gamma(w_{id}+1) + \sum_{v=1}^{V} w_{vd}\ln\beta_{kv}\Big]$$

b - We derive the previous expression by $\theta_k$ including the Lagrangian from the constraint: $\sum\theta_k = 1$

$$\frac{\partial}{\partial\theta_k}\sum_{d=1}^{D}\sum_{k=1}^{K} \gamma_{dk}\Big[\ln\theta_k + \ln\Gamma(N_d+1) - \sum_{i=1}^{N_d}\ln\Gamma(w_{id}+1) + \sum_{v=1}^{V} w_{vd}\ln\beta_{kv} + \lambda(1 - \sum_{k=1}^{K}\theta_k)\Big] = 0$$

$$\frac{\partial}{\partial\theta_k}\sum_{d=1}^{D}\sum_{k=1}^{K} \gamma_{dk}\Big[\ln\theta_k + \lambda(1 - \sum_{k'=1}^{K}\theta'_k)\Big] = 0$$

This implies:

$$\sum_{d=1}^{D}\gamma_{dk}\big(\frac{1}{\theta_k} - \lambda\big) = 0$$

$$\sum_{d=1}^{D}\gamma_{dk}\big(1 - \lambda\theta_k\big) = 0$$

$$\sum_{d=1}^{D}\gamma_{dk} = \sum_{d=1}^{D}\lambda\theta_k$$

$$\sum_{d=1}^{D}\gamma_{dk} = \lambda D\theta_k$$

hence, we get the following equation (1):

$$\theta_k = \frac{1}{\lambda D}\sum_{d=1}^{D}\gamma_{dk}$$

Now let's sum over $k$:

$$\sum_{k=1}^{K}\theta_k = \frac{1}{\lambda D}\sum_{k=1}^{K}\sum_{d=1}^{D}\gamma_{dk}$$

We know that $\sum_{k=1}^{K} \theta_k = 1$ and $\sum_{k=1}^{K} \gamma_{dk} = \sum_{k=1}^{K} q(z_{dk}) = 1$ so $\sum_{k=1}^{K} \sum_{d=1}^{D} \gamma_{dk} = D$

We now have $1 = \frac{D}{\lambda D}$ so $\lambda = 1$

Replacing $\lambda$ in equation (1), we get:

$$\theta_k = \frac{1}{D} \sum_{d=1}^{D} \gamma_{dk}$$

which makes a lot of sense, $\theta_k$ becomes the average probability of documents being of topic $k$.

c- We want to derive the expected log likelihood with respect to $\beta_{kv}$.

We add the Lagrangian for the constraint: $\sum_{v=1}^{V} \beta_{kv} = 1$

$$\frac{\partial}{\partial \beta_{kv}} \Big[ \sum_{d=1}^{D} \sum_{k=1}^{K} \gamma_{dk} \Big( \ln \theta_k + \ln \Gamma(N_d + 1) - \sum_{i=1}^{N_d} \ln \Gamma(w_{id} + 1) + \sum_{v=1}^{V} w_{vw} \ln \beta_{kv} + \lambda (1 - \sum_{v=1}^{V} \beta_{kv}) \Big) \Big] = 0$$

$$\sum_{d=1}^{D} \gamma_{dk} (\frac{w_{vd}}{\beta_{kv}} - \lambda) = 0$$

$$\sum_{d=1}^{D} \gamma_{dk} (w_{vd} - \lambda \beta_{kv}) = 0$$

$$\sum_{d=1}^{D} \gamma_{dk} w_{vd} = \sum_{d=1}^{D} \lambda \beta_{kv}$$

$$\sum_{d=1}^{D} \gamma_{dk} w_{vd} = \lambda D \beta_{kv}$$

We get the following equation (2):

$$\beta_{kv} = \tfrac{1}{\lambda D} \textstyle\sum_{d=1}^{D} \gamma_{dk} w_{vd}$$

Let's sum over $v$:

$$\sum_{v=1}^{V} \beta_{kv} = \frac{1}{\lambda D} \sum_{v=1}^{V} \sum_{d=1}^{D} \gamma_{dk} w_{vd}$$

We know that $\sum_{v=1}^{V} \beta_{kv} = 1$:

$$\sum_{v=1}^{V} \sum_{d=1}^{D} \gamma_{dk} w_{vd} = \lambda D$$

$$\sum_{d=1}^{D} \gamma_{dk} \Big( \sum_{v=1}^{V} w_{vd} \Big) = \lambda D$$

By definition $\sum_{v=1}^{V} w_{vd} = N_d$

$$\lambda = \frac{1}{D} \sum_{d=1}^{D} \gamma_{dk} N_d$$

We plug this into equation (2):

$$\hat{\beta}_{kv} = \frac{D}{\sum_{d=1}^{D} \gamma_{dk} N_d} \frac{1}{D} \sum_{d=1}^{D} \gamma_{dk} w_{vd}$$

Finally,

$$\hat{\beta}_{kv} = \frac{\sum_{d=1}^{D} \gamma_{dk} w_{vd}}{\sum_{d=1}^{D} \gamma_{dk} N_d}$$

**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 $5224951 \times 3$ 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.

[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

After an extensive use of Pandas' pivot_table and merge functions, I was able to get an iteration (step E, step M and computation of the expected complete data log likelihood) under 14 seconds for K=10 and 23 for K=20.

I have initialized theta randomly and used a Dirichelet distribution to initialize beta.

I had to use a exp and a ln function to compute the $\gamma_{dk}$ matrix to avoid Python rounding up to zero when multiplying powers of probabilities ($\prod_{i=1}^{N_d} \beta_{ki}^{w_{id}}$).

My loss function converges and never increases but unfortunately I have noticed that after a few iterations my matrix beta converges towards a matrix where all $beta_k$ are the almost the same. Which means that the most 'likely' word of a topic is always the same one, the word with the most counts across documents: word number 23314: 'researched'.

All of this makes this model useless.

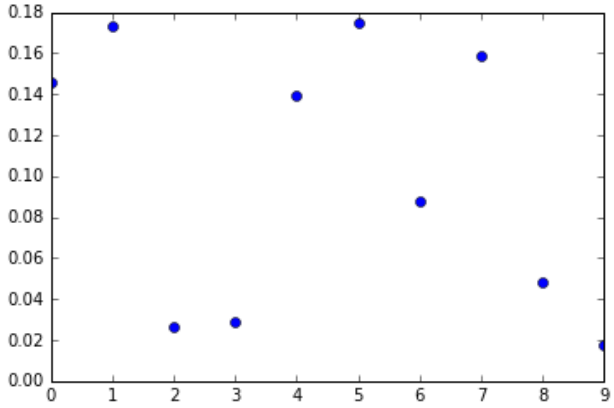


Figure 1: Theta initialization for K=10

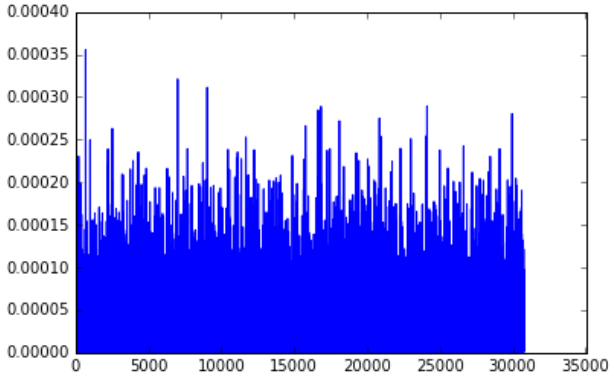


Figure 2: Initialization of a vector beta

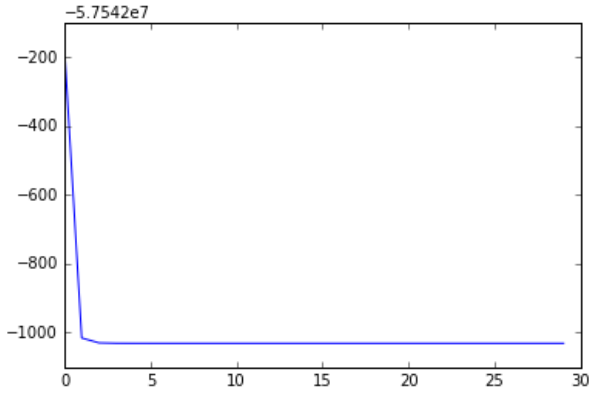


Figure 3: Loss function vs iteration

**Problem 5** (Calibration, 1pt)
Approximately how long did this homework take you to complete? 35h