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 β_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

The complete data likelihood is given by

$$\begin{aligned}
p(\{z_d, w_d\}_{d=1}^{D}|\theta, \{\beta_k\}_{k=1}^{K}) &= p(\{w_d\}_{d=1}^{D}|\{\beta_k\}_{k=1}^{K})p(\{z_d\}_{d=1}^{D}|\theta) \\
&= (\prod_{d=1}^{D}\prod_{k=1}^{K} p(w_d|z_d = k)^{z_{d,k}})(\prod_{d=1}^{D}\prod_{k=1}^{K} \theta_k^{z_{d,k}}) \\
&\propto (\prod_{d=1}^{D}\prod_{k=1}^{K}\prod_{v=1}^{V} \beta_{k,v}^{w_{d,v}z_{d,k}})(\prod_{d=1}^{D}\prod_{k=1}^{K} \theta_k^{z_{d,k}})
\end{aligned}$$

taking logs gives us

$$\ln p(\{z_d, w_d\}_{d=1}^{D}|\theta, \{\beta_k\}_{k=1}^{K}) = (\sum_{d=1}^{D}\sum_{k=1}^{K}\sum_{v=1}^{V}(w_{d,v}z_{d,k}) \ln \beta_{k,v}) + (\sum_{d=1}^{D}\sum_{k=1}^{K} z_{d,k} \ln \theta_k) + const. \qquad (1)$$

Check 1.1. You gave Eq. 1 as your answer for the complete data log likelihood.

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

In the E-step, we compute the posterior over the hidden variables z_d, given some setting of the parameters β and θ. These parameters come from the M-step, or a random initialization if this is the first E-step performed. For a given vector z_d, its posterior probability is given by:

$$\begin{aligned} q(z_d|\theta,\beta,w_d) &\propto p(w_d|z_d,\theta,\beta)p(z_d|\theta) \\ &\propto \left(\prod_{k=1}^{K}\prod_{v=1}^{V}\beta_{k,v}^{w_{d,v}\cdot z_{d,k}}\right)\left(\prod_{k=1}^{K}\theta_k^{z_{d,k}}\right) \end{aligned}$$

The probabilities of all possible values for z_d must sum to 1. z_d can take K possible values, which we denote by $z_d(k)$, such that $(z_d(k))_i = 1$ iff $i = k$ and $= 0$ otherwise:

$$\sum_{k=1} q(z_d(k)|\theta,\beta,w_d) = 1$$

Hence we can normalize:

$$q(z_d|\theta,\beta,w_d) = \frac{\left(\prod_{k=1}^{K}\prod_{v=1}^{V}\beta_{k,v}^{w_{d,v}\cdot z_{d,k}}\right)\left(\prod_{k=1}^{K}\theta_k^{z_{d,k}}\right)}{\sum_{k=1}^{K}\left(\prod_{v=1}^{V}\beta_{k,v}^{w_{d,v}\cdot z_{d,k}}\right)\left(\theta_k^{z_{d,k}}\right)}$$

And in particular:

$$\forall k \in K, \quad q(z_d(k)|\theta,\beta,w_d) = \frac{\left(\prod_{v=1}^{V}\beta_{k,v}^{w_{d,v}\cdot z_{d,k}}\right)\left(\theta_k^{z_{d,k}}\right)}{\sum_{k=1}^{K}\left(\prod_{v=1}^{V}\beta_{k,v}^{w_{d,v}\cdot z_{d,k}}\right)\left(\theta_k^{z_{d,k}}\right)} \tag{2}$$

Finally, we can write: $\gamma_{d,k} = \mathbb{E}[z_{d,k}] = \mathbb{P}(\text{document d} \in \text{topic k}) = q(z_d(k)|\theta,\beta,w_d)$, which is given to us by the equation above.

Check 2.1. You cited both the initialization step and the previous M-step as the setting of parameters β and θ.

Check 2.2. You gave Eq. 2 with its normalization constant as your answer.

Check 2.3 You explained the link between $q(z_d(k))$ and $\gamma_{d,k}$ for the implementation of your algorithm

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 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

We want to maximize the expected complete data likelihood given fixed γ's. By linearity of expectation, the ECDLL is

$$\begin{aligned}ECDLL &= E(\ln p(\{z_d, w_d\}_{d=1}^D | \theta, \{\beta_k\}_{k=1}^K) \\ &= (\sum_{d=1}^D \sum_{k=1}^K \sum_{v=1}^V (w_{d,v}\gamma_{d,k}) \ln \beta_{k,v}) + (\sum_{d=1}^D \sum_{k=1}^K \gamma_{d,k} \ln \theta_k) + const.\end{aligned}$$

To find maximizing values of θ, we can operate only on the right term. Note that our solution is constrained such that $\sum_k \theta_k = 1$. Our Lagrangian is thus

$$L(\theta_k, \lambda) = \sum_{d=1}^D \sum_{k=1}^K \gamma_{d,k} \ln \theta_k + \lambda(\sum_{k=1}^K \theta_k - 1) \tag{3}$$

and the derivatives are

$$\frac{\partial L}{\partial \theta_k} = \sum_{d=1}^D \frac{1}{\theta_k}\gamma_{d,k} + \lambda$$

$$\frac{\partial L}{\partial \lambda} = \sum_{k=1}^K \theta_k - 1$$

$$\theta_k = \frac{-1}{\lambda} \sum_{d=1}^D \gamma_{d,k}$$

Now we'd like to back out the value of λ. Summing both sides over k allows us to use the second constraint

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

$$-\lambda = \sum_{k=1}^K \sum_{d=1}^D \gamma_{d,k} = D$$

due to the fact that the $\gamma_{d,k}$ values sum to 1 for each d. So our maximization for θ_k is simply the fraction of documents assigned to topic k

$$\hat{\theta}_k = \frac{1}{D} \sum_{d=1}^{D} \gamma_{d,k} \tag{4}$$

Note that the same exact argument (due to the constraint $\sum_{v=1}^{V} \beta_{k,v} = 1$) will lead us to

$$\hat{\beta}_{k,v} = \frac{\sum_{d=1}^{D} w_{d,v} \gamma_{d,k}}{\sum_{d=1}^{D} \sum_{v=1}^{V} w_{d,v} \gamma_{d,k}} \tag{5}$$

Check 3.1. You correctly gave Eq. 3. as the expected complete data log likelihood.

Check 3.2. You set up the Langrangian in Eq. 4.

Check 3.3. You solved for $\hat{\theta}_k$ as shown in Eq. 5 and explained why the optimized θ makes sense.

Check 3.4. You solved for $\hat{\beta}_{k,v}$ as shown in Eq. 6.

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.

[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

Implementing the expectation maximization algorithm, you can verify that the objective function indeed never increases, and that it converges. You should have explained how you are reporting your results and reported the results for several different numbers of topics. The solutions code outputs the most likely 10 words in each of the K topics. The most likely 10 words for $K = 5, 10, 15$ are shown in the tables below.

The most common word in many of the classes is "research", which is not surprising given the dataset. In the $K = 15$ case, we can see academic discplines emerging in several of the topics: biological words "cell", "proteins", "genes", "dna" in Class 9; mathematical words "equations", "geometry", "mathematical", "theory" in Class 8; and computer science words such as "computer", "engineering", "software" in Class 15. You should find similar results. See the Tables 1-3 for our results.

Check 4.1: You implemented the expectation maximization algorithm on the text data. You plotted the objective function and verified that it is never increasing.

Check 4.2: You coded up a reasonable way to report the topics you found and indicated what the method was. An example is to report the most likely words in each topic.

Check 4.3: You ran your code and reported the results for different numbers of topics.

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

Class 1:	research, species, protein, cell, cells, studies, study, proteins, gene, understanding
Class 2:	research, theory, systems, problems, project, methods, design, study, system, analysis
Class 3:	research, students, project, science, program, university, data, support, laboratory, engineering
Class 4:	research, materials, high, project, chemistry, properties, study, university, program, systems
Class 5:	research, data, study, project, provide, ocean, studies, processes, model, field

Table 1: The most likely 10 words in each of the topics with $K = 5$.

Class 1	research, project, data, study, social, support, important, information, provide, science
Class 2	science, students, project, research, program, mathematics, teachers, school, education, university
Class 3	theory, problems, research, study, equations, systems, project, work, methods, analysis
Class 4	research, chemistry, materials, properties, study, project, chemical, studies, metal, surface
Class 5	species, research, study, plant, project, studies, genetic, important, understanding, plants
Class 6	research, students, university, project, laboratory, program, support, equipment, science, chemistry
Class 7	cell, protein, cells, proteins, gene, genes, studies, research, specific, system
Class 8	research, design, systems, system, project, control, data, based, performance, models
Class 9	research, high, materials, project, study, optical, properties, magnetic, university, physics
Class 10	data, research, study, project, ocean, ice, processes, model, studies, models

Table 2: The most likely 10 words in each of the topics with $K = 10$.

Class 1	research, project, data, study, social, economic, information, important, understanding, political
Class 2	research, data, study, project, ocean, ice, water, climate, carbon, processes
Class 3	research, data, study, model, solar, project, observations, field, stars, models
Class 4	research, science, students, project, program, university, teachers, mathematics, school, education
Class 5	research, systems, design, system, project, data, problems, control, algorithms, methods
Class 6	research, materials, high, months, support, phase, optical, devices, project, year
Class 7	species, research, study, data, project, important, provide, studies, understanding, dr
Class 8	theory, research, study, problems, equations, work, systems, project, geometry, mathematical
Class 9	protein, cell, proteins, cells, gene, genes, studies, specific, research, dna
Class 10	research, project, cell, high, study, university, cells, development, phase, water
Class 11	research, data, project, study, seismic, earthquake, results, models, provide, structure
Class 12	research, students, chemistry, laboratory, university, analysis, equipment, project, molecular, undergrad.
Class 13	research, materials, properties, study, chemistry, systems, project, high, studies, surface
Class 14	research, university, conference, support, workshop, program, international, scientists, project, award
Class 15	students, laboratory, project, computer, research, data, science, engineering, courses, software

Table 3: The most likely 10 words in each of the topics with $K = 15$.