CMSC 25025 / STAT 37601

Machine Learning and Large Scale Data Analysis

Assignment 4

Due: Thursday, May 9, 2013

This assignment consists of four problems. Three of them are "pencil and paper" problems to help solidify your understanding of topic modeling. The fourth asks you to calculate topic models on a portion of the wishes data from LSD Project 1, and on the State of the Union data. For extra credit, you can do the fifth problem, which is to implement collapsed Gibbs sampling for LDA. In all of these problems we adopt the notation used in class.

1. *Local Topic Variable Sampling* (20 points)

(a) Let $z_{1:N}$ denote the topic indicator variables for document d in a K-topic LDA model. The topics are denoted β_k, for $k = 1, \ldots, K$; each of these is a multinomial over a V-word vocabulary.

Derive the conditional probability distribution

$$\mathbb{P}(z_n \,|\, z_{-n}, \beta_{1:K}, w_{1:N}, \alpha)$$

where the topic mixture proportions $\theta_d \sim$ Dirichlet(α) are integrated out. Give a detailed derivation and explanation of each step. Include all normalizing constants.

(b) Explain how the distribution above can be used to approximate $\mathbb{E}(\theta_d \,|\, \beta_{1:K}, w_{1:N}, \alpha)$ with a Gibbs sampling algorithm.

(a) Solution:

We calculate the joint probability $P(z_n, w_n | z_{-n}, w_{-n}, \beta_{1:K}, \alpha)$ of the topic assignment z_n and the word w_n conditioned on the other parameters first and then use Bayes rule to find the required conditional probability distribution.

We have

$$\begin{aligned}\mathbb{P}(w_n, z_n = t | z_{-n}, w_{-n}, \beta_{1:K}, \alpha) &= \int \mathbb{P}(w_n, z_n | \theta, \beta_{1:K}) \mathbb{P}(\theta | \alpha, z_{-n}, w_{-n}) d\theta \\ &= \beta_{t,w_n} \int \theta_t \mathbb{P}(\theta | \gamma) d\theta \qquad (1)\end{aligned}$$

where $\gamma = (\gamma_1, \gamma_2, \ldots \gamma_K)$ and

$$\gamma_i = \alpha + n_i(z_{-n}). \qquad \text{Thus}$$
$$\theta | \gamma \sim Dir(\alpha + n_1(z_{-n}), \alpha + n_2(z_{-n}), \ldots \alpha + n_K(z_{-n}))$$

where the last term is obtained as the posterior of the Dirichlet distribution when you have seen the rest of the words and the topic assignments besides the n^{th} one. Note that $n_i(z_{-n})$ refers to the count of the i^{th} topic among all the words except the n^{th} one.

Note that (1) is just the expectation of θ_t under the posterior distribution. Since the posterior is a Dirichlet distribution, the mean is given by

$$\int \theta_t \mathbb{P}(\theta|\gamma) d\theta = \frac{\alpha + n_t(z_{-n})}{K\alpha + \sum_j n_j(z_{-n})}$$

Since $\sum_j n_j(z_{-n}) = N - 1$ (sum of topic counts for all the words except the n^{th} one, plugging this back we get

$$\mathbb{P}(w_n, z_n = t | z_{-n}, w_{-n}, \beta_{1:K}, \alpha) = \beta_{t,w_n} \frac{\alpha + n_t(z_{-n})}{K\alpha + N - 1} .$$

The required conditional distribution is obtained by

$$\mathbb{P}(z_n = t | z_{-n}, w_{-n}, \beta_{1:K}, \alpha) = \frac{\mathbb{P}(w_n, z_n = t | z_{-n}, w_{-n}, \beta_{1:K}, \alpha)}{\mathbb{P}(w_n | z_{-n}, w_{-n}, \beta_{1:K}, \alpha)}$$

using Bayes Rule. The denominator is obtained as

$$\begin{aligned}\mathbb{P}(w_n | z_{-n}, w_{-n}, \beta_{1:K}, \alpha) &= \sum_{z_n} \mathbb{P}(w_n, z_n | z_{-n}, w_{-n}, \beta_{1:K}, \alpha) \\ &= \sum_{z_n} \frac{\beta_{z_n,w_n}(\alpha + n_{z_n}(z_{-n}))}{K\alpha + N - 1}\end{aligned}$$

Taking the ratio of the two probabilities we get

$$\mathbb{P}(z_n = t | z_{-n}, w_{1:N}, \beta_{1:K}, \alpha) = \frac{\beta_{t,w_n}(\alpha + n_t(z_{-n}))}{\sum_j \beta_{j,w_n}(\alpha + n_j(z_{-n}))}$$

(b) Solution:

As is obvious from (1), sampling from the distribution in the first part, requires you to take the mean of a particular θ_t with respect to the posterior distribution. Thus we can approximate $\mathbb{E}(\theta_d | \beta_{1:K}, w_{1:N}, \alpha)$ by sampling from the scaled distribution $\frac{1}{\beta_{z_n,w_n}} \mathbb{P}(z_n, w_n | z_{-n}, w_{-n}, \beta_{1:K}, \alpha)$.

2. *Big Data, Big Integrals* (20 points)

Assuming the same latent Dirichlet allocation model as above, derive a closed form expression for the integral

$$\mathbb{P}(w_{1:N}, z_{1:N} \,|\, \beta_{1:K}, \alpha) = \int \mathbb{P}(w_{1:N}, z_{1:N} \,|\, \theta_d, \beta_{1:K}) \, p(\theta_d \,|\, \alpha) \, d\theta_d.$$

Solution:

We have

$$\begin{aligned}
\mathbb{P}(w_{1:N}, z_{1:N} \mid \beta_{1:K}, \alpha) &= \int \mathbb{P}(w_{1:N}, z_{1:N} \mid \theta_d, \beta_{1:K}) p(\theta_d \mid \alpha)\, d\theta_d \\
&= \int \prod_{n=1}^{N} \beta_{z_n,w_n} \theta_{z_n} C Dir(\alpha) d\theta_d \qquad C = \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \text{ = Constant of proportionality} \\
&= \prod_{n=1}^{N} \beta_{z_n,w_n} \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \int \prod_{z_n} \theta_{z_n} \prod_{j=1}^{K} \theta_j^{\alpha-1} d\theta_d \\
&= \prod_{n=1}^{N} \beta_{z_n,w_n} \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \int \prod_{j=1}^{K} \theta_j^{\alpha+n_j(z_{1:N})-1} d\theta_d \\
&= \prod_{n=1}^{N} \beta_{z_n,w_n} \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \frac{\prod_{j=1}^{K} \Gamma(n_j(z_{1:N}) + \alpha)}{\Gamma(N + K\alpha)}
\end{aligned}$$

The last expression follows from noting that $\sum_j n_j(z_{1:N})$ is the sum of the topic counts over all the topics in the document and evaluates to N, the total number of words in the document.

3. *Global Topic Variable Sampling* (20 points)

   (a) Let $\boldsymbol{z}$ denote the topic indicator variables for entire collection of words $\boldsymbol{w}$ across all documents in the collection. Derive the conditional probability distribution

   $$\mathbb{P}(z_n \mid \boldsymbol{z}_{-n}, \boldsymbol{w}, \alpha, \eta)$$

   of the topic z_n for a specific word w_n in a document, where the topics $\beta_{1:K}$ are integrated out with respect to a prior Dirichlet(η). Give a detailed derivation and explanation of each step. Include all normalizing constants.

   (b) Explain how the distribution above can be used to approximate $\mathbb{E}(\beta_k \mid \boldsymbol{w}, \alpha, \eta)$ with a Gibbs sampling algorithm.

   (a) Solution:

   We proceed in a similar manner to problem 1. The joint distribution is given by

   $$\mathbb{P}(w_{n,d}, z_{n,d} = t | \boldsymbol{z}_{-n}, \boldsymbol{w}_{-n}, \alpha, \eta) = \int\int \mathbb{P}(w_{n,d}, z_{n,d} = t | \theta, \boldsymbol{\beta}) \mathbb{P}(\theta | \alpha, z_{-n}) \mathbb{P}(\boldsymbol{\beta} | \boldsymbol{z}_{-n}, \boldsymbol{w}_{-n}, \eta) d\theta d\boldsymbol{\beta}$$

   $$= \int\int \beta_{t,w_n} \theta_t \mathbb{P}(\theta|\gamma) \prod_{i=1}^{K} \mathbb{P}(\beta_i|\lambda_i) d\theta d\beta_i$$

   where $\gamma = (\gamma_1, \gamma_2, \dots \gamma_K)$ and $\gamma_i = \alpha + n_i(z_{-n})$ and $\lambda_j = (\lambda_{j,1}, \lambda_{j,2} \dots \lambda_{j,V})$ for all $j = \{1, 2, \dots K\}$. Thus

   $$\theta|\gamma \sim Dir(\alpha + n_1(z_{-n}), \alpha + n_2(z_{-n}), \dots \alpha + n_K(z_{-n})) \qquad \text{and}$$
   $$\beta_i|\lambda_i \sim Dir(\lambda_{i,1}, \lambda_{i,2} \dots, \lambda_{i,V})$$

Here $\boldsymbol{\beta} = \beta_{1:K}$ and the second dirichlet distribution is just the posterior distribution of $\{\beta_i\}_{i=1}^K$ after seeing all the words in D documents. As mentioned in the lecture notes, $\beta_i \in \mathbb{R}^V$ is the word distribution corresponding to the i^{th} topic and

$$\lambda_{k,v} = \eta + m_{k,v}(\boldsymbol{z}_{-(n,d)}, \boldsymbol{w}_{-(n,d)})$$

where $m_k(\boldsymbol{z}_{-(n,d)}, \boldsymbol{w}_{-(n,d)})$ refers to the number of words having value v among topic k excluding $w_{n,d}$. Using the fact that

$$\int \int \beta_{t,w_n} \theta_t \mathbb{P}(\theta|\gamma) \prod_{i=1}^{K} \mathbb{P}(\beta_i|\lambda_i) d\theta d\beta_i = \mathbb{E}[\theta_t|\gamma]\mathbb{E}[\beta_{t,w_n}|\lambda]$$

and plugging in the expected values of the posterior Dirichlet distribution we get

$$\mathbb{P}(w_{n,d}, z_{n,d} = t|\boldsymbol{z}_{-n}, \boldsymbol{w}_{-n}, \alpha, \eta) = \frac{(\alpha + n_t(z_{-n}))}{(K\alpha + \sum_j n_j(z_{-n}))} \frac{\eta + m_{t,w_n}(\boldsymbol{z}_{-(n,d)}, \boldsymbol{w}_{-(n,d)})}{(V\eta + \sum_v m_{t,v}(\boldsymbol{z}_{-(n,d)}, \boldsymbol{w}_{-(n,d)}))}$$

Note that if $z_{n,d} = t$ then $\sum_v m_{t,v}(\boldsymbol{z}_{-(n,d)}, \boldsymbol{w}_{-(n,d)}) = m_t(\boldsymbol{w}) - 1$ where $m_t(\boldsymbol{w})$ is the word count of the t^{th} topic among all the words in all the documents.

Calculating the marginal $\mathbb{P}(w_{n,d}|\boldsymbol{z}_{-n}, \boldsymbol{w}_{-n}, \alpha, \eta)$ would require summing over all possible values of $z_{n,d} = 1, 2, \ldots K$ which can be shown easily to be

$$\mathbb{P}(w_{n,d}|\boldsymbol{z}_{-n}, \boldsymbol{w}_{-n}, \alpha, \eta) = \frac{1}{(K\alpha + \sum_j n_j(z_{-n}))} \sum_{k=1}^{K} \frac{(\alpha + n_k(z_{-n}))\eta + m_{k,w_n}(\boldsymbol{z}_{-(n,d)}, \boldsymbol{w}_{-(n,d)})}{(V\eta + \sum_v m_{k,v}(\boldsymbol{z}_{-(n,d)}, \boldsymbol{w}_{-(n,d)}))}$$

Dividing, we get

$$\mathbb{P}(z_{n,d} = t|\boldsymbol{z}_{-n}, \boldsymbol{w}, \alpha, \eta) = \frac{\frac{(\alpha+n_t(z_{-n}))(\eta+m_{t,w_n}(\boldsymbol{z}_{-(n,d)},\boldsymbol{w}_{-(n,d)}))}{(V\eta+m_t(\boldsymbol{w})-1)}}{\sum_{k=1}^{K} \frac{(\alpha+n_k(z_{-n}))(\eta+m_{k,w_n}(\boldsymbol{z}_{-(n,d)},\boldsymbol{w}_{-(n,d)}))}{(V\eta+m_k(\boldsymbol{w})-1)}}$$

(b) Solution:

As in the first problem, it is easy to notice that in this case, we are calculating the expectation $\mathbb{E}[\beta_{k,w_n}|\lambda]$ with respect to the posterior distribution. Thus we can approximate the mean of β_k by sampling from the distribution $\mathbb{P}(w_{n,d}, z_{n,d}|\boldsymbol{z}_{-n}, \boldsymbol{w}{-}n, \alpha, \eta)$ and scaling it with the appropriate terms corresponding to the posterior mean of θ.