STAT 383C: Statistical Modeling I **Fall 2015**

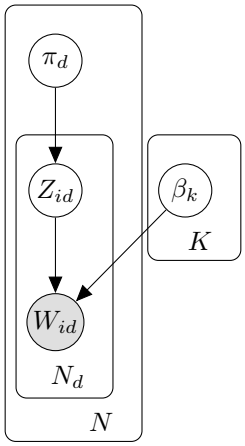
Lecture 22 — November 12

Lecturer: Purnamrita Sarkar *Scribe: Andrew Beck*

Disclaimer: These scribe notes have been slightly proofread and may have typos etc.

Note: The latex template was borrowed from EECS, U.C. Berkeley.

22.1 Latent Dirichlet Application: Gibbs Sampling



Recall that for the LDA:
$\pi_d \sim Dir(\alpha_1, ..., \alpha_K), d = 1, ...N$
$\beta_k \sim Dir(\lambda_1, ..., \lambda_V), k = 1, ..., K$
$Z_{id} \sim \pi_d, i = 1...N_d, d = 1...N$

Furthermore, once we know the topic d, we have $w_{id}|z_{id} = k \sim \beta_k$

22.1.1 Gibbs Sampling for LDA

We begin by looking at:
$\pi_d \sim P(\pi_d|\{Z_d\}, \{W_d\}, \{\beta_k\}; \alpha, \lambda)$

As $\{Z_{\setminus d}\}$ and $\{W_{\setminus d}\}$ are not in the Markov Blanket of π_d, we get:

$$\pi_d \sim P(\pi_d|Z_d; \alpha) = Dir(n_{1d} + \alpha_1, ..., n_{kd} + \alpha_k) \tag{22.1}$$

where Z_d is a vector for document d, and $n_{id} = |\{i : Z_{id} = 1\}|$

Next,

$$\beta_k \sim P(\beta_k|\{\pi_d\},\{Z_d\},\{W_d\},\{\beta_{\setminus k}\};\alpha,\lambda) = Dir(m_{1k}+\lambda_1,...,m_{Vk}+\lambda_V) \tag{22.2}$$

where $m_{ik} = |\{(i,d)|z_{id}=k, w_{id}=1\}|$

This situation is the analogue of the Naive Bayes version, except that every document now has multiple topics, and every word is associated with a topic. The formula is derived from Bayes' rule in the same way as in the derivations shown in previous classes. The difference is that for every document, we look at the distribution π_d, whereas in the Naive Bayes document model we had only a single distribution π for all documents. We look at π_d as a probability vector that is unique for each document, that gives each document its own characteristics.

Finally, we look at $Z_{id} \sim P(Z_{id}=k|\{Z_{\setminus id}\},\{w_d\},\{\beta_k\},\pi;\alpha,\lambda)$ The Markov Blanket here consists of π_d, W_{id}, and β_k. Thus, we can knock off the other Z's and the other W's:

$$\begin{aligned} Z_{id} &\sim P(Z_{id}=k|W_{id},\{\beta_k\},\pi;\alpha,\lambda) \\ &\propto P(W_{id}=w|Z_{id}=k,\{\beta_k\},\pi;\alpha,\lambda)\cdot P(Z_{id}=k|\{\beta_k\},\pi;\alpha,\lambda) \end{aligned}$$

$$P(Z_{id}=k|W_{id}=w,\beta_k) \sim \frac{\beta_{kw}\pi_k}{\sum_i \beta_{iw}\pi_i}$$

Note that the proportionality between (22.3) and (22.4) comes from Bayes' Rule. Also note that the final expression in (22.5) is very similar to what we saw in Naive Bayes: $(\prod_{w\in W_d}\beta_{kw})\pi_k$. In Naive Bayes, the words are all coming from the document topic d, and we must look at every probability. Now, each word is coming from its own topic, so we simply need to look at our specific word's topic, β_{kw}.

22.1.2 Discussion

One issue is that we will need to sample a very large number of parameters:
$\pi_d \; for \; d = 1..N$
$\beta_k \; for \; k = 1..K$
$Z_{id} \; for \; d = 1...N, i = 1...N_d$

Consequently, our sample space will be huge. Since in Gibbs Sampling we're sampling a parameter in each step by conditioning on every other parameter, and then using the result to move one step in the underlying Markov Chain, this will take "forever" to converge. Thus, we can feel the need for collapsing

There are a few ways to get around large sample spaces. As an example, given every paper at a conference, we can pre-process by eliminating all the most common words (ie prepositions) and all the most uncommon words (ie typos). However, the vocabulary will still be huge.

As another example, we can take advantage of sparse vectors by storing them compactly. For example, if we have a vector with a length of a million that only has a thousand data points, we can simply store the non-zero entries along with their positions in the vector.

Nevertheless, Gibbs Sampling may still take "forever!" So we may need to collapse the problem. Either the π's, the β's, or both can be collapsed. See the posted: "Primer to Gibbs for the Uninitiated" posted in the lecture notes on 09/10 for more information.

22.2 Collapsed Gibbs Sampling for LDA

As we saw last class, in the collapsed model:

$$Z_{id} \sim P(Z_{id} = k | \{Z_{\setminus id}\}, \{W_{id}\}; \alpha, \lambda) \tag{22.3}$$

We also recall that integrating out π makes all the Z's dependent on each other. This makes the gibbs sampling steps more dependent on each other, and sometimes Collapsed Gibbs may actually take longer. Thus this is really a tradeoff between smaller sample space and longer mixing time.

We note that the first thing we should do when we something bad is to re-write the problem in a new way so that we know how to do the likelihood, i.e., use Bayes' Rule!

$$Z_{id} \propto \underbrace{P(\{W_{id}\}|Z_{id} = k, \{Z_{\setminus id}\}; \alpha, \lambda)}_{A} \cdot \underbrace{P(Z_{id} = k|\{Z_{\setminus id}\}; \alpha, \lambda)}_{B} \tag{22.4}$$

Now we look at A, and we use Bayes' Rule again:

$$A \propto \underbrace{P(W_{id} = w|Z_{id} = k, \{W_{\setminus id}\}, \{Z_{\setminus id}\}; \alpha, \lambda)}_{A1} \cdot \underbrace{P(\{W_{\setminus id}\}|Z_{id} = k, \{Z_{\setminus id}\}; \alpha, \lambda)}_{A2} \tag{22.5}$$

Now we look at $A1$. If we hadn't collapsed out the β_k's, we could write:

$$\begin{aligned} A1 &= \int_{\beta_k} P(W_{id} = w, \beta_k | Z_{id} = k, \{W_{\setminus id}\}, \{Z_{\setminus id}\}; \alpha, \lambda) d\beta_k \\ &= \int_{\beta_k} P(W_{id} = w | \beta_k, Z_{id} = k, \{W_{\setminus id}\}, \{Z_{\setminus id}\}; \alpha, \lambda) \cdot P(\beta_k | Z_{id} = k, \{W_{\setminus id}\}, \{Z_{\setminus id}\}; \alpha, \lambda) d\beta_k \end{aligned}$$

By using the Markov Blanket and D-Separation, we simplify to:

$$\int_{\beta_k} P(W_{id} = w | \beta_k, Z_{id} = k; \alpha, \lambda) \cdot P(\beta_k | \{W_{\setminus id}\}, \{Z_{\setminus id}\}; \alpha, \lambda) d\beta_k \tag{22.6}$$

Now we note, as we saw last class, that this is not simply one integral, but rather multiple integrals, one for each element of the vector β_k. However, if we look only at the integral over β_{kw}, then the rest becomes a Dirichlet, and we get:

$$\int_{\beta_k} \beta_{kw} \cdot Dir(\lambda_1 + m_{1k}^{-id}, ..., \lambda_v + m_{vk}^{-id}) d\beta_k \tag{22.7}$$

Here, as we saw last time, m_{jk}^{-id} is our original count values m_{jk}, with the single word id omitted. Now, by using the technique we saw last class and on the first homework assignment, we get:

$$\int_{\beta_{kw}} \beta_{kw} \cdot Beta(m_{wk}^{-id} + \lambda_{w1}, \sum_{j \neq w} m_{jk}^{-id} + \lambda_j) d\beta_w \tag{22.8}$$

Finally, we note that this expression is simply the expected value of a Beta, so we conclude that:

$$A1 = \frac{m_{wk}^{-id} + \lambda_w}{\sum_{j=1}^{V} m_{jk}^{-id} + \lambda_j} \tag{22.9}$$

We return to $A2$. This expression looks at every word other than the idth word. If we correctly use the β's and integrate them out, this expression will not depend on the z_{id}! Thus, the entire $A2$ will be absorbed into the proportionality constant. We conclude that:

$$A \propto A1 \cdot A2 \propto A1 \tag{22.10}$$

Now we return to $B = P(Z_{id} = k | \{Z_{\setminus id}\}; \alpha, \lambda)$. As there are no words in this expression, we have no need to work with the β's, so we only use the π's. Thus, using the same techniques as before, we get:

$$B = \int_{\pi_d} P(Z_{id} = k | \pi_d, \{Z_{\setminus id}\}; \alpha, \lambda) \cdot P(\pi_d | \{Z_{\setminus id}\}; \alpha, \lambda) d\pi_d \tag{22.11}$$

As the Z's become independent with the π's:

$$= \int_{\pi_d} P(Z_{id} = k | \pi_d; \alpha, \lambda) \cdot P(\pi_d | \{Z_{\setminus id}\}; \alpha, \lambda) d\pi_d \tag{22.12}$$

And, as before:

$$= \int_{\pi_d} \pi_{dk} \cdot Dir(n_{1d}^{-id} + \alpha_1, ..., n_{kd}^{-id} + \alpha_k) d\pi_d \tag{22.13}$$

$$B = \frac{n_{kd}^{-id} + \alpha_k}{\sum_{j=1}^{K} n_{jd}^{-id} + \alpha_j} \tag{22.14}$$

Thus, to tie it all back together,

$$Z_{id} \sim P(Z_{id} = k | \{Z_{\setminus id}\}, \{W_{id}\}; \alpha, \lambda) \tag{22.15}$$

$$\propto A1 \cdot B \tag{22.16}$$

$$= \left(\frac{m_{wk}^{-id} + \lambda_w}{\sum_{j=1}^{V} m_{jk}^{-id} + \lambda_j} \right) \left(\frac{n_{kd}^{-id} + \alpha_k}{\sum_{j=1}^{K} n_{jd}^{-id} + \alpha_j} \right) \tag{22.17}$$