STAT 383C: Statistical Modeling I **Fall 2016**

Lecture 21 — November 10

Lecturer: Purnamrita Sarkar *Scribe: Joowon Cho*

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

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

21.1 Markov Blanket

The Markov blanket for a node X_i in a Bayesian network may be denoted by $MB(X_i)$. $MB(X_i) = \{Parents(X_i),\ Children(X_i),\ CoParents(X_i)\}$ The Markov blanket of a node contains all the variables that shield the node from the rest of the network. This means that the Markov blanket of a node is the only knowledge needed to predict the behavior of that node. $P(X_i \mid ....) = P(X_i \mid MB(X_i))$

21.2 Bayesian Document Model

Consider the problem of classifying documents into topics based on their contents. The basic idea is: if a document contains, for instance, several words related to statistics (e.g., data, model, inference, likelihood), then it is likely that it belongs to Statistics topic.

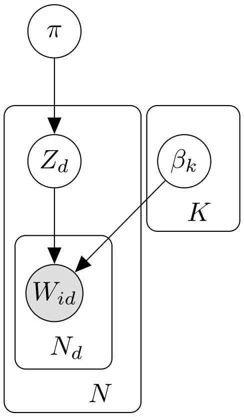
We will see a model that allows more than one topic per document, but for now, the first model we are going to consider in this lecture supposes that a given document belongs to only one topic.

According to this model, a document is randomly generated as follows: first we randomly pick one topic Z_d, from the distribution over topics $\pi \sim$ Dirichlet$(\lambda_1, ..., \lambda_K)$. The latent topic Z_d is just a distribution over words, and such distribution is fully specified by the probabilities $\beta = (\beta_1, ..., \beta_V)$ over the words in the vocabulary. Finally, the words for document d (denoted by $W_{id},\ i = 1, ..., N_d$) are then randomly selected from the vocabulary with probabilities β. The words are the observed data, and all the other variables of the model are latent.

In summary, the model can be defined by the following set of statements:

- Z_d : topic of document $d \in \{1, ..., N\}$ (N: number of documents)
- $Z_d \in \{1, ..., K\}$ (K: number of topics)
- $\pi = (\pi_1, ..., \pi_K) \sim \text{Dir}(\alpha_1, ..., \alpha_K)$
  $\pi_k = P(Z_d = k), k \in \{1, ..., N\}$.
- W_i^d : word i in document $d \in \{1, ..., N\}$
  $W_{id} \in \{1, ...N_d\} \rightarrow$ vocabulary of the document d

- Distribution of W_{id} varies with the topic of document d.
  $P(W_{id} = w \mid Z_d = k) = \beta_{kw}, \quad \beta_k = (\beta_{k1}, ..., \beta_{zV})$
  V : total of words in the vocabulary
  $\beta_k \sim \mathrm{Dir}(\lambda_1, ..., \lambda_{N_d}), \quad z \in \{1, ..., K\}$.



21.2.1 Full-conditional distributions for Gibbs sampler

Now we derive the full-conditional distributions for running a Gibbs sampler and get to the posterior distribution of the parameter of the model described before. Notice that all full-conditionals are analytically available and sampling from them is straightforward.

1.

$$\begin{aligned}\pi^{(t+1)} &\sim p(\pi \mid \{Z_d\}, \{W_i^d\}, \{\beta_k\}; \alpha, \lambda) \\ &= p(\pi \mid \{Z_d\}; \alpha) \\ &\propto p(\{Z_d\} \mid \pi; \alpha) p(\pi; \alpha) \\ &= \prod_{d=1}^{D} p(\{Z_d\} \mid \pi; \alpha) p(\pi; \alpha) \\ &\propto \left(\prod_{d=1}^{D} \prod_{k=1}^{K} \pi_k^{1(Z_d = k)} \right) \prod_{k=1}^{K} \pi_k^{d_k - 1} \\ &= \prod_{k=1}^{K} \pi_k^{n_k + d_k - 1} \\ &= \mathrm{Dir}(n_1 + \alpha_1, ..., n_K + \alpha_K)\end{aligned}$$

where $n_i = \sum_{d=1}^{N} 1(Z_d = i)$.

2.

$$\begin{aligned}
\beta_k^{(t+1)} &\sim p(\beta_k \mid \{Z_d\}, \{W_{id}\}, \{\beta_{-k}\}, \pi; \lambda, \alpha) \\
&= p(\beta_k \mid \{Z_d\}, \{W_{id}\}, \{\beta_{-k}\}; \lambda, \alpha) \\
&= p(\beta_k \mid \{Z_d; Z_d = k\}, \{W_{id}; Z_d = k\}, \{\beta_{-k}\}; \lambda, \alpha) \\
&\propto p(\beta_k, \{Z_d; Z_d = k\}, \{W_{id}\}; Z_d = k\}, \{\beta_{-k}\}; \lambda, \alpha) \\
&\propto p(\{Z_d; Z_d = k\} \mid \beta_k, \{\beta_{-k}\}\{Z_d; Z_d = k\}, ; \lambda, \alpha) \\
&= \prod_{Z_d=k}^{N} \prod_{i=1}^{N_d} \prod_{v=1}^{V} \beta_{kv}^{1(W_{id}=v)} \prod_{v}^{V} \beta_{kv}^{\lambda_v - 1} \\
&= \prod_{v=1}^{V} \beta_{kv}^{m_{dv}+\lambda_k-1} \\
&= \mathrm{Dir}(m_{k1} + \lambda_1, ..., m_{kV} + \lambda_V)
\end{aligned}$$

where $m_{dv} = \{(i, d) : W_{id} = v, z_d = k\}$, i.e. total number of occurrences of word w in documents of topic k.

3.

$$\begin{aligned}
Z_d^{(t+1)} &\sim P(Z_d = k \mid \{Z_{-d}\}, \{W_{id}\}, \{\beta_k\}, \pi; \lambda, \alpha) \\
&= P(Z_d = k \mid \{W_{id}\}, \{\beta_k\}, \pi; \lambda, \alpha), i = 1..N_d \\
&\propto p(\{W_{id}\} \mid Z_d = k, \beta_k, \pi) P(Z_d = k, \{\beta_k\}, \pi), i = 1..N_d \\
&= \prod_{i=1}^{N_d} \beta_{kW_{id}} \times \pi_k
\end{aligned}$$

Hence the full conditional distribution for Z_d is discrete given by

$$P(Z_d = k \mid \{Z_{-d}\}, \{W_{id}\}, \{\beta_k\}, \pi; \lambda, \alpha) = \frac{\prod_{i=1}^{N_d} \beta_{kW_{id}} \times \pi_k}{\sum_{k=1}^{N_d} \prod_{i=1}^{N_d} \beta_{kW_{id}} \times \pi_k}$$

21.2.2 Collapsed Gibbs Sampler

Sometimes it is possible to marginalize the likelihood over a set of parameters, therefore reducing the number of nodes in the MCMC chain. This can save time during the simulation of the chains, since we have less nodes to sample from.

Here we exemplify how it works by showing the calculations for marginalizing out π from the full conditional of Z_d. Notice that this is enough to marginalize π out of the likelihood, since it does not appear in the full-conditional distribution of any of the β_k's.

$$
\begin{aligned}
&P(Z_d = k \mid \{Z_d\}, \{W_{id}\}, \{\beta_k\}) \propto \\
&\qquad \propto P(\{W_{id}\} \mid \{Z_d\}, \{Z_{-d}\}, \{\beta_k\}) \times P(Z_d = k, \{Z_{-d}\}, \{\beta_k\})
\end{aligned}
$$

References

[1] Blei, David. Probabilistic topic models. *Communications of the ACM, 55(4):7784, 201*, 2012.

[1] Resnik, P. and Hardisty, E. Gibbs Sampling for the Uninitiated. *CS-TR-4956 UMIACS-TR-2010-04 LAMP-TR-153*, June 2010.