CS281B/Stat241B: Advanced Topics in Learning & Decision Making

Gibbs Sampling (ctd) & Properties of Dirichlet Distribution

Lecturer: Michael I. Jordan *Scribes: Xiaoyue Zhao*

1 Gibbs sampling example: Ising model

In Ising model,

$$P(X|\theta) = \frac{1}{z}\exp\{\sum_{j<k}\theta_{jk}x_jx_k + \sum_i \theta_i x_i\},$$

where $\theta_{jk} = 0$, if $j \neq k$ and z is a normalizing constant.

$$\begin{aligned}
P(X_i = 1|x_{v\setminus i}) &= \frac{P(X_i = 1, x_{v\setminus i})}{P(x_{v\setminus i})} \\
&= \frac{\exp\{\sum_{j\neq i}\theta_{ij}x_j + \theta_i + \sum_{j<k, j\neq i, k\neq i}\theta_{jk}x_jx_k + \sum_{j\neq i}\theta_j x_j\}}{\exp\{\sum_{j\neq i}\theta_{ij}x_j + \theta_i + \sum_{j<k, j\neq i, k\neq i}\theta_{jk}x_jx_k + \sum_{j\neq i}\theta_j x_j\} + \exp\{\sum_{j<k, j\neq i, k\neq i}\theta_{jk}x_jx_k + \sum_{j\neq i}\theta_j x_j\}} \\
&= \frac{\exp\{\sum_{j\neq i}\theta_{ij}x_j + \theta_i\}}{\exp\{\sum_{j\neq i}\theta_{ij}x_j + \theta_i\} + 1} \\
&= \frac{1}{1 + \exp\{-\sum_{j\neq i}\theta_{ij}x_j - \theta_i\}}.
\end{aligned}$$

Here v is the set of all the vertices. The conditional probability is in the form of logistic function.

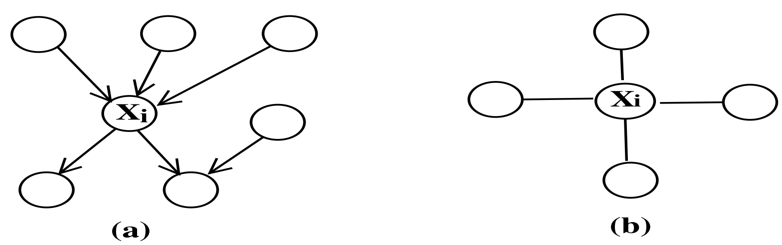

Figure 1: Markov Blanket.

Figure 1(b) gives an example of the Markov blanket for node X_i for the Ising model. The conditional probability of a node X_i depends only on its neighbors.

(For directed graphical models, the Markov blanket includes all parents, all children and all co-parents, as shown in Figure 1(b)).

Gibbs sampling can be viewed as a special case of Metropolis-Hastings. In Gibbs sampling, the proposal distribution is the posterior probability and the acceptance probability is always 1.

2 Dirichlet Processes

2.1 Dirichlet distribution

The Dirichlet distribution is a member of the exponential family. It is a conjugate distribution to the multinomial. In a non-minimal representation, its density can be written:

$$P(p_1, p_2, \cdots, p_k | \alpha_1, \alpha_2, \cdots, \alpha_k) = \frac{\Gamma(\sum \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} \cdots p_k^{\alpha_k - 1},$$

where $\alpha_i > 0$ and $\sum p_i = 1$.

2.2 Examples

1. $k = 2$. There is only one p.
2. $k = 3$. It corresponds to a simplex embedded to a 3-dimension space, see Figure 2. For $0 < \alpha < 1$, multiple modes appear in the corners.

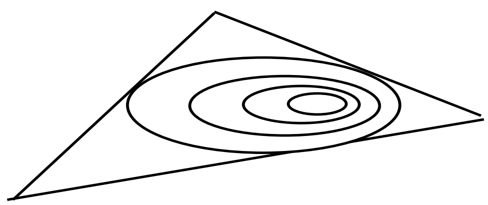

Figure 2: Geometric description of Dirichlet distribution when $k = 3$.

Special case of $k = 2$:
The Dirichlet distribution $Beta(\alpha_1, \alpha_2)$ is conjugate to $X_i \sim Bernoulli(p_1)$. Figure 3 gives the corresponding graphical model representation.

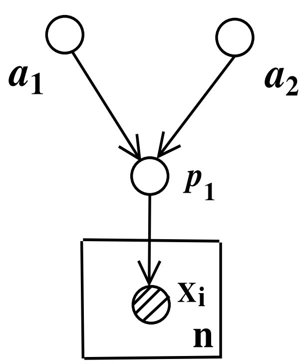

Figure 3: Graphical model representation.

The posterior distribution of p_1 given the data is:

$$P(p_1 | X_1, \cdots, X_n) \quad \propto \quad p_1^{\alpha_1 - 1} (1 - p_1)^{\alpha_2 - 1} \prod_{i=1}^{n} p_1^{X_i} (1 - p_1)^{1 - X_i}$$

$$\begin{aligned} &= p_1^{\alpha_1+\sum_{i=1}^{n}\delta_1(X_i)-1}(1-p_1)^{\alpha_2+\sum_{i=1}^{n}\delta_2(X_i)-1} \\ &\sim Beta(\alpha_1+\sum_{i=1}^{n}\delta_1(x_i), \alpha_2+\sum_{i=1}^{n}\delta_2(x_i)) \end{aligned}$$

p_1 has prior parameters α_1 and α_2. After having data, add spikes $\sum_{i=1}^{n}\delta_1(X_i)$ to α_1, $\sum_{i=1}^{n}\delta_2(X_i)$ to α_2. By averaging out $\mathbf{p}$, we have

$$\begin{aligned} P(X_1,\cdots,X_n|\alpha_1,\alpha_2) &= \int dp_1 P(X_1,\cdots,X_n,p_1|\alpha_1,\alpha_2) \\ &= \int dp_1 P(X_1,\cdots,X_n|p_1)P(p_1|\alpha_1,\alpha_2) \\ &= \int dp_1 \prod_i p_1^{X_i}(1-p_1)^{1-X_i}\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}p_1^{\alpha_1-1}(1-p_1)^{\alpha_2-1} \\ &= \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\frac{\Gamma(\alpha_1+\sum_{i=1}^{n}\delta_1(X_i))\Gamma(\alpha_2+\sum_{i=1}^{n}\delta_2(X_i))}{\Gamma(\alpha_1+\alpha_2+n)} \\ &= \frac{\alpha_1^{[\sum_i\delta_1(X_i)]}\alpha_2^{[\sum_i\delta_2(X_i)]}}{(\alpha_1+\alpha_2)^{[n]}}, \end{aligned}$$

where $X^{[n]} = X(X+1)\cdots(X+n-1)$. Here we use the recursion formula $\Gamma(x+1) = x\Gamma(x)$. So

$$\begin{aligned} P(X_{n+1}|X_1,\cdots,X_n,\alpha_1,\alpha_2) &= \frac{\Gamma(\alpha_+ + n)}{\Gamma(\alpha_+ + n+1)}\frac{\Gamma(\alpha_1+\sum_{i=1}^{n+1}\delta_1(X_i))}{\Gamma(\alpha_1+\sum_{i=1}^{n}\delta_1(X_i))}\frac{\Gamma(\alpha_2+\sum_{i=1}^{n+1}\delta_2(X_i))}{\Gamma(\alpha_2+\sum_{i=1}^{n}\delta_2(X_i))} \\ &= \begin{cases} \frac{\alpha_1+\sum_{i=1}^{n}\delta_1(X_i)}{\alpha_+ + n}, & X_{n+1}=1 \\ \frac{\alpha_2+\sum_{i=1}^{n}\delta_2(X_i)}{\alpha_+ + n}, & X_{n+1}=2 \end{cases} \end{aligned}$$

which has the form of the urn model.

$P(X_{n+1}=1|X_1,\cdots,X_n,\alpha_1,\alpha_2)$ can be written as the weighted summation of prior mean of p_1 and maximum likelihood estimate (MLE) of p_1, i.e.,

$$\frac{\alpha_1+\sum_{i=1}^{n}\delta_1(X_i)}{\alpha_+ + n} = \frac{\alpha_1}{\alpha_+}(\frac{\alpha_+}{\alpha_+ + n}) + (\frac{n}{\alpha_+ + n})\frac{1}{n}\sum\delta_1(X_i).$$

General case of Dirichlet

The mean has the form of

$$E(P_i) = \frac{\alpha_i}{\alpha_+}, \quad \text{where} \quad \alpha_+ = \sum_{i=1}^{k}\alpha_i.$$

The Dirichlet distribution is conjugate to multinomial. The posterior still follows the Dirichlet:

$$P(P_1, \cdots, P_k | X_1, \cdots, X_n) \sim Dir(\alpha_1 + \sum_{i=1}^{n} \delta_1(X_i), \cdots, \alpha_k + \sum_{i=1}^{n} \delta_k(X_i)).$$

The Marginal distribution is:

$$P(X_1, \cdots, X_n | \alpha_1, \cdots, \alpha_k) = \frac{\alpha_1^{[\sum_i \delta_1(X_i)]} \cdots \alpha_k^{[\sum_i \delta_k(X_i)]}}{\alpha_+^{[n]}},$$

and

$$P(X_{n+1} = j | X_1, \cdots, X_n, \alpha) = \frac{\alpha_j + \sum_{i=1}^{n} \delta_i(X_i)}{\alpha_+ + n},$$

which again can be viewed in terms of an urn model.

Let $B_1, B_2, \cdots, B_l$ be a partition at $\{1, 2, \cdots, k\}, l < k$. Let $1 < r_1 < \cdots < r_l = k$ be integers. Then

$$P(\sum_{i=1}^{r_1} P_i, \sum_{i=r_1+1}^{r_2} P_i, \cdots, \sum_{i=r_{l-1}+1}^{k} P_i) \sim Dir(\sum_{i=1}^{r_1} \alpha_i, \sum_{i=r_1+1}^{r_2} \alpha_i, \cdots, \sum_{i=r_{l-1}+1}^{k} \alpha_i).$$

This property is both necessary and sufficient for a distribution to be Dirichlet.

2.3 Dirichlet process: $DP(G_0, \alpha)$

Let G_0 be a probability measure on (X, Ω). Then $\exists$ a unique probability measure $P(G_0, \alpha)$ on the space of measures on X, such that

$$(P(B_1), P(B_2), \cdots, P(B_k)) \sim Dir(\alpha G_0(B_1), \cdots, \alpha G_0(B_k)),$$

where $B_1, \cdots, B_k$ is a partition of X.