Advanced Probabilistic Machine Learning and Applications

Abdullahi Adinoyi Ibrahim and Caterina De Bacco

May 7, 2021

1 Tutorial 3 (Solution Exercise 1)

Recall, a Dirichlet distribution with parameter $\boldsymbol{\alpha}$ has a probability density function defined by

$$\text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})}\prod_k \pi_k^{\alpha_k - 1} \qquad B(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)}$$

Throughout the document it is possible to notice the condicional distributions are independent on some variables. These independency properties can be extracted from the graphical model. The log likelihood of the data can be computed as

$$\begin{aligned}\log p(\boldsymbol{X}|\boldsymbol{\Theta}, \boldsymbol{\pi}, \boldsymbol{Z}) &= \sum_n \log p(\mathbf{x}_n|\boldsymbol{\Theta}, \boldsymbol{\pi}, \boldsymbol{Z}) = \sum_n \log p(\mathbf{x}_n|\boldsymbol{\Theta}, \boldsymbol{\pi}, z_n) = \sum_n \log p(\mathbf{x}_n|\boldsymbol{\theta}_{z_n}) \\ &= \sum_n \sum_{j=1}^{W_n} \log \text{Cat}(x_{nj}|\boldsymbol{\theta}_{z_n}) \\ &= \sum_n \sum_{j=1}^{W_n} \sum_m \log \theta_{z_n m}^{[x_{nj}=m]}\end{aligned}$$

1.1 Notation

- $\boldsymbol{Z}_{-n} = \{\mathbf{z}_i | i \neq n\}_{i=1}^N$
- $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_k\}_{k=1}^K$

Algorithm 1: Gibbs sampling

Algorithm 1: Gibbs sampling algorithm

Initialize cluster assignments $\boldsymbol{Z}$ and the model parameters $\boldsymbol{\pi}$, $\boldsymbol{\Theta}$;
while *not converged* **do**
  Sample $\boldsymbol{\pi} \sim p(\boldsymbol{\pi}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\Theta}) = p(\boldsymbol{\pi}|\boldsymbol{Z})$;
  for $k = 1, \ldots, K$ **do**
    Sample $\boldsymbol{\theta}_k \sim p(\boldsymbol{\theta}_k|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}) = p(\boldsymbol{\theta}_k|\boldsymbol{X}, \boldsymbol{Z})$;
  end
  for $n = 1, \ldots, N$ **do**
    Sample $z_n \sim p(z_n|\boldsymbol{X}, \boldsymbol{Z}_{-n}, \boldsymbol{\pi}, \boldsymbol{\Theta}) = p(z_n|\mathbf{x}_n, \boldsymbol{\pi}, \boldsymbol{\Theta})$;
  end
end

Posterior distribution over $\boldsymbol{\pi}$: We first calculate the form of $p(\boldsymbol{\pi}|\boldsymbol{Z})$

$$
\begin{aligned}
p(\boldsymbol{\pi}|\boldsymbol{Z}) \propto p(\boldsymbol{\pi})p(\boldsymbol{Z}|\boldsymbol{\pi}) &= p(\boldsymbol{\pi}) \prod_{n=1}^{N} p(z_n|\boldsymbol{\pi}) \\
&= \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \prod_n \prod_k \pi_k^{[z_n=k]} \\
&= \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \prod_k \pi_k^{m_k} \qquad (1) \\
&= \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}') \qquad (2)
\end{aligned}
$$

where we have used $m_k = \sum_n [z_n = k]$ in Equation 1 and $\alpha'_k = \alpha_k + m_k$ in Equation 2.

Posterior distribution over $\boldsymbol{\theta}_k$:

$$
\begin{aligned}
p(\boldsymbol{\theta}_k|\boldsymbol{X}, \boldsymbol{Z}) &\propto p(\boldsymbol{\theta}_k)p(\boldsymbol{X}|\boldsymbol{\theta}_k, \boldsymbol{Z}) \\
&= \mathrm{Dir}(\boldsymbol{\theta}_k|\boldsymbol{\gamma}) \prod_n p(\mathbf{x}_n|\boldsymbol{\theta}_k)^{[z_n=k]} \\
&= \mathrm{Dir}(\boldsymbol{\theta}_k|\boldsymbol{\gamma}) \prod_n \prod_j^{W_n} \mathrm{Cat}(x_{nj}|\boldsymbol{\theta}_k)^{[z_n=k]} \\
&= \mathrm{Dir}(\boldsymbol{\theta}_k|\boldsymbol{\gamma}) \prod_n \prod_j^{W_n} \prod_{m=1}^{|I|} \boldsymbol{\theta}_{km}^{[x_{nj}=m][z_n=k]} \\
&= \mathrm{Dir}(\boldsymbol{\theta}_k|\boldsymbol{\gamma}) \prod_{m=1}^{|I|} \boldsymbol{\theta}_{km}^{c_{km}} \qquad (3) \\
&= \mathrm{Dir}(\boldsymbol{\theta}_k|\boldsymbol{\gamma}'_k) \qquad (4)
\end{aligned}
$$

where we have used $c_{km} = \sum_n [z_n = k] \sum_j [x_{nj} = m]$ in Equation 3 which is the number of occurrences of the m-th word in the cluster k; and $\gamma'_{km} = \gamma_m + c_{km}$ in Equation 4.

Posterior distribution over z_n: We use Bayes Rule

$$
p(z_n = k|\mathbf{x}_n, \boldsymbol{\pi}, \boldsymbol{\Theta}) = \frac{p(z_n = k, \mathbf{x}_n|\boldsymbol{\pi}, \boldsymbol{\Theta})}{\sum_{k'} p(z_n = k', \mathbf{x}_n|\boldsymbol{\pi}, \boldsymbol{\Theta})} = \frac{p(z_n = k|\boldsymbol{\pi})p(\mathbf{x}_n|z_n = k, \boldsymbol{\Theta})}{\sum_{k'} p(z_n = k', \mathbf{x}_n|\boldsymbol{\pi}, \boldsymbol{\Theta})} = \frac{\pi_k p(\mathbf{x}_n|\boldsymbol{\theta}_k)}{\sum_{k'} \pi_{k'} p(\mathbf{x}_n|\boldsymbol{\theta}_{k'})}
$$

Algorithm 2: Gibbs sampling with $\boldsymbol{\pi}$ collapsed

Since we have selected conjugate prior distribution for the mixing components $\boldsymbol{\pi}$, we can marginalize them out.

Algorithm 2: $\boldsymbol{\pi}$ collapsed Gibbs sampling algorithm

Initialize cluster assignments $\boldsymbol{Z}$ and the model parameters $\boldsymbol{\Theta}$;
while *not converged* **do**
  for $k = 1, \ldots, K$ **do**
    Sample $\boldsymbol{\theta}_k \sim p(\boldsymbol{\theta}_k|\boldsymbol{X}, \boldsymbol{Z})$;
  end
  for $n = 1, \ldots, N$ **do**
    Sample $z_n \sim p(z_n|\boldsymbol{X}, \boldsymbol{Z}_{-n}, \boldsymbol{\Theta}) = p(z_n|\mathbf{x}_n, \boldsymbol{Z}_{-n}, \boldsymbol{\Theta})$;
  end
end

Posterior distribution over z_n: We first can write the posterior probability of the n-th sample belonging to cluster k is proportional to the joint distribution

$$
p(z_n = k|\mathbf{x}_n, \boldsymbol{Z}_{-n}, \boldsymbol{\Theta}) \propto p(z_n = k, \mathbf{x}_n|\boldsymbol{Z}_{-n}, \boldsymbol{\Theta}) = p(\mathbf{x}_n|z_n = k, \boldsymbol{Z}_{-n}, \boldsymbol{\Theta})p(z_n = k|\boldsymbol{Z}_{-n})
$$

which notice we must normalize the resulting distribution $\sum_k p(z_n = k|\mathbf{x}_n, \boldsymbol{\Theta}, \boldsymbol{Z}_{-n}) = 1$. The prior term (given all previous cluster assignments) can be calculated as

$$
\begin{aligned}
p(z_n = k|\boldsymbol{Z}_{-n}) &= \int p(z_n = k, \boldsymbol{\pi}|\boldsymbol{Z}_{-n}) d\boldsymbol{\pi} \\
&= \int p(z_n = k|\boldsymbol{\pi}, \boldsymbol{Z}_{-n}) p(\boldsymbol{\pi}|\boldsymbol{Z}_{-n}) d\boldsymbol{\pi} \\
&= \int p(z_n = k|\boldsymbol{\pi}) p(\boldsymbol{\pi}|\boldsymbol{Z}_{-n}) d\boldsymbol{\pi} \\
&= \int \prod_k \pi_k^{[z_n=k]} \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}'') d\boldsymbol{\pi} \qquad (5) \\
&= \int \frac{1}{B(\boldsymbol{\alpha}'')} \prod_k \pi_k^{\alpha_k'' + [z_n=k] - 1} d\boldsymbol{\pi} \\
&= \frac{B(\{\alpha_k'' + [z_n = k]\})}{B(\boldsymbol{\alpha}'')} \\
&= \frac{\prod_{k=1}^K \Gamma(\alpha_k'' + [z_n = k])}{\Gamma(\sum_k \alpha_k'' + 1)} \frac{\Gamma(\sum_k \alpha_k'')}{\prod_{k=1}^K \Gamma(\alpha_k'')} \\
&= \frac{\Gamma(\alpha_k'' + 1)}{\sum_k \alpha_k'' \Gamma(\sum_k \alpha_k'')} \frac{\Gamma(\sum_k \alpha_k'')}{\Gamma(\alpha_k'')} \qquad (6) \\
&= \frac{\alpha_k'' \Gamma(\alpha_k'')}{\sum_k \alpha_k'' \Gamma(\sum_k \alpha_k'')} \frac{\Gamma(\sum_k \alpha_k'')}{\Gamma(\alpha_k'')} \qquad (7) \\
&= \frac{\alpha_k''}{\sum_k \alpha_k''} \\
&= \frac{\sum_{i \neq n} [z_i = k] + \alpha_k}{N - 1 + \sum_k \alpha_k}
\end{aligned}
$$

where we have used the result in Exercise 1 to get $\alpha_k'' = \alpha_k + \sum_{i \neq n} [\mathbf{z}_i = k]$ in Equation 6; in the steps 7 and 9 we make use of the property of the gamma function $\Gamma(x + 1) = x\Gamma(x)$. Notice the resulting distribution follows the scheme "rich get richer". Additionally, note that normalization is needed so that $\sum_k p(z_n = k|\boldsymbol{Z}_{-n}) = 1$. Moving on, the likelihood term has the form

$$
p(\mathbf{x}_n|z_n = k, \boldsymbol{Z}_{-n}, \boldsymbol{\Theta}) = p(\mathbf{x}_n|z_n = k, \boldsymbol{\Theta}) = p(\mathbf{x}_n|\boldsymbol{\theta}_k) = \prod_{j=1}^{W_n} \mathrm{Cat}(x_{nj}|\boldsymbol{\theta}_k)
$$

Posterior distribution over $\boldsymbol{\theta}_k$: It was computed in Exercise 1.

Algorithm 3: Gibbs sampling with $\boldsymbol{\pi}$ and $\boldsymbol{\Theta}$ collapsed

Since we have selected the conjugate prior distribution for the likelihood parameters $\boldsymbol{\Theta}$, we can marginalize them out.

Algorithm 3: $\boldsymbol{\pi}, \boldsymbol{\Theta}$ collapsed Gibbs sampling algorithm

```
Initialize cluster assignments Z;
while not converged do
    for n = 1, ..., N do
        Sample z_n ~ p(z_n|X, Z_{-n});
    end
end
```

Posterior distribution over z_n: Firstly, we can write the posterior probability of the n-th sample belonging to cluster k is proportional to the joint distribution

$$
p(z_n = k|\mathbf{x}_n, \boldsymbol{X}_{-n}, \boldsymbol{Z}_{-n}) \propto p(z_n = k, \mathbf{x}_n|\boldsymbol{X}_{-n}, \boldsymbol{Z}_{-n}) = p(z_n = k|\boldsymbol{Z}_{-n}) p(\mathbf{x}_n|z_n = k, \boldsymbol{X}_{-n}, \boldsymbol{Z}_{-n})
$$

which notice we must normalize the resulting distribution $\sum_k p(z_n = k|\mathbf{x}_n, \boldsymbol{X}_{-n}, \boldsymbol{Z}_{-n}) = 1$. The prior term is computed in Exercise 2. The posterior predictive can be computed marginalizing the likelihood parameters

$$\begin{aligned}
p(\mathbf{x}_n|z_n = k, \boldsymbol{X}_{-n}, \boldsymbol{Z}_{-n}) &= \int p(\mathbf{x}_n, \boldsymbol{\theta}_k|z_n = k, \boldsymbol{X}_{-n}, \boldsymbol{Z}_{-n}) d\boldsymbol{\theta}_k \\
&= \int p(\mathbf{x}_n|z_n = k, \boldsymbol{\theta}_k) p(\boldsymbol{\theta}_k|\boldsymbol{X}_{-n}, \boldsymbol{Z}_{-n}) d\boldsymbol{\theta}_k \\
&= \int \prod_{j=1}^{W_n} \text{Cat}(x_{nj}|\boldsymbol{\theta}_k) \text{Dir}(\boldsymbol{\theta}_k|\boldsymbol{\gamma}''_k) d\boldsymbol{\theta}_k \qquad (8) \\
&= \int \prod_{j=1}^{W_n} \prod_{m=1}^{|I|} \boldsymbol{\theta}_{km}^{[x_{nj}=m]} C \prod_{m=1}^{|I|} \boldsymbol{\theta}_{km}^{\gamma''_{km}-1} d\boldsymbol{\theta}_k \\
&= \frac{1}{B(\boldsymbol{\gamma}''_k)} \int \prod_{m=1}^{|I|} \boldsymbol{\theta}_{km}^{c_{nm}+\gamma''_{km}-1} d\boldsymbol{\theta}_k \qquad (9) \\
&= \frac{B(\boldsymbol{\gamma}''_k(n) + \boldsymbol{c}_n)}{B(\boldsymbol{\gamma}''_k(n))}
\end{aligned}$$

where we have used the result in Exercise 1 to get $\gamma''_{km} = \gamma_m + \sum_{i \neq n} [\mathbf{z}_i = k] \sum_j [\mathbf{x}_{ij} = m]$ in Equation 8; the quantity $c_{nm} = \sum_j [x_{nj} = m]$ in Equation 9 represents the number of occurrences of the m-th word in document n; in steps 13 we use $\sum_m c_{nm} = W_n$. We can further develop the ratio between the two Beta functions

$$\begin{aligned}
\frac{B(\boldsymbol{\gamma}''_k + \boldsymbol{c}_n)}{B(\boldsymbol{\gamma}''_k)} &= \frac{\prod_{m=1}^{|I|} \Gamma(\gamma''_{km} + c_{nm})}{\Gamma(\sum_m \gamma''_{km} + c_{nm})} \frac{\Gamma(\sum_m \gamma''_{km})}{\prod_{m=1}^{|I|} \Gamma(\gamma''_{km})} \\
&= \frac{\prod_{m=1}^{|I|} \prod_{i=0}^{c_{nm}-1} (\gamma''_{km} + i) \Gamma(\gamma''_{km})}{\Gamma(\sum_m \gamma''_{km} + W_n)} \frac{\Gamma(\sum_m \gamma''_{km})}{\prod_{m=1}^{|I|} \Gamma(\gamma''_{km})} \\
&= \frac{\prod_{m=1}^{|I|} \prod_{i=0}^{c_{nm}-1} (\gamma''_{km} + i) \Gamma(\gamma''_{km})}{\prod_{j=0}^{W_n-1} (\sum_m \gamma''_{km} + j) \, \Gamma(\sum_m \gamma''_{km})} \frac{\Gamma(\sum_m \gamma''_{km})}{\prod_{m=1}^{|I|} \Gamma(\gamma''_{km})} \\
&= \frac{\prod_{m=1}^{|I|} \prod_{i=0}^{c_{nm}-1} (\gamma''_{km} + i)}{\prod_{j=0}^{W_n-1} (\sum_m \gamma''_{km} + j)}
\end{aligned} \qquad (10)$$

We compute the log posterior predictive to avoid numerical instabilities

$$\begin{aligned}
\log p(\mathbf{x}_n|z_n = k, \boldsymbol{X}_{-n}, \boldsymbol{Z}_{-n}) &= \log \prod_{m=1}^{|I|} \prod_{i=0}^{c_{nm}-1} (\gamma''_{km} + i) - \log \prod_{j=0}^{W_n-1} \left(\sum_m \gamma''_{km} + j \right) \\
&= \sum_{m=1}^{|I|} \sum_{i=0}^{c_{nm}-1} \log(\gamma''_{km} + i) - \sum_{j=0}^{W_n-1} \log \left(\sum_m \gamma''_{km} + j \right)
\end{aligned}$$

Finally, the form of our target posterior distribution is

$$p(z_n = k|\mathbf{x}_n, \boldsymbol{X}_{-n}, \boldsymbol{Z}_{-n}) \propto \frac{\sum_{i \neq n} [z_i = k] + \alpha_k}{N - 1 + \sum_k \alpha_k} \frac{B(\boldsymbol{\gamma}''_k + \boldsymbol{c}_n)}{B(\boldsymbol{\gamma}''_k)}$$