Advanced Probabilistic Machine Learning and Applications

Martina Contisciani and Caterina De Bacco

November 8, 2021

Contents

1 Tutorial 3: Bayesian Mixture Models and Gibbs sampling

1.1 Exercise 1: Categorical Mixture Model (CMM)

In this tutorial, we will continue working with the CMM and the twitter dataset presented in Tutorial 2. Here, we will use different versions of the Gibbs sampling algorithm to find the posterior distributions of the cluster assignments $\mathbf{Z}$ and model parameters $(\boldsymbol{\pi}, \boldsymbol{\theta})$.
As a recap, the dataset $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}^\top$ describes a set of N documents, here tweets generated by U users. Each tweet has been cleaned and pre-processed (lemmatisation, lowerization, and stemming) using a dictionary of words I, and it is represented as $\mathbf{x}_n = (x_{n1}, \dots, x_{nW_n})$, i.e. as a vector of W_n words. Each word $x_{nj} \in \{1, \dots, |I|\}$ is described by its position in the dictionary. We introduce also a set $\mathbf{Z}$ of latent variables which represent the cluster assignments.
We work with the following Bayesian Mixture Model:

$$p(\mathbf{x}, \mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\theta}) = p(\boldsymbol{\pi}|\boldsymbol{\alpha}) p(\boldsymbol{\theta}|\boldsymbol{\gamma}) \prod_{n=1}^{N} p(z_n|\boldsymbol{\pi}) p(\mathbf{x}_n|z_n, \boldsymbol{\theta}),$$

where

$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = Dir(\boldsymbol{\pi}|\boldsymbol{\alpha}) \quad p(\boldsymbol{\theta}|\boldsymbol{\gamma}) = \prod_{k=1}^{K} Dir(\boldsymbol{\theta}_k|\boldsymbol{\gamma})$$

$$p(z_n|\boldsymbol{\pi}) = Cat(z_n|\boldsymbol{\pi}) \quad p(\mathbf{x}_n|z_n, \boldsymbol{\theta}) = \prod_{j=1}^{W_n} Cat(x_{nj}|\boldsymbol{\theta}_{z_n}).$$

Remember that the conjugate prior of a Categorical distribution is the Dirichlet distribution, and notice the prior distributions for each $\boldsymbol{\theta}_k$ share the same set of parameters.

1. Algorithm 1 approximates (using samples) the posterior distribution $p(\mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\theta}|\mathbf{x})$. Derive the conditional distributions needed to sample in steps (1), (2), and (3) of the Gibbs sampling algorithm.

Algorithm 1: Gibbs sampling algorithm

```
Initialize cluster assignments $\mathbf{Z}$ and model parameters $\boldsymbol{\pi}$, $\boldsymbol{\theta}$;
for $\tau = 1, \ldots, N_{it}$ do
    Sample $\boldsymbol{\pi} \sim p(\boldsymbol{\pi}|\mathbf{x}, \mathbf{z}, \boldsymbol{\theta}) = p(\boldsymbol{\pi}|\mathbf{z})$;    (1)
    for $k = 1, \ldots, K$ do
        Sample $\boldsymbol{\theta}_k \sim p(\boldsymbol{\theta}_k|\mathbf{x}, \mathbf{z}, \boldsymbol{\pi}) = p(\boldsymbol{\theta}_k|\mathbf{x}, \mathbf{z})$;    (2)
    end
    for $n = 1, \ldots, N$ do
        Sample $z_n \sim p(z_n|\mathbf{x}, \mathbf{z}_{-n}, \boldsymbol{\pi}, \boldsymbol{\theta}) = p(z_n|\mathbf{x}_n, \boldsymbol{\pi}, \boldsymbol{\theta})$;    (3)
    end
end
```

Let's start with the posterior distribution over $\boldsymbol{\pi}$:

$$\begin{aligned} p(\boldsymbol{\pi}|\mathbf{z}) \propto p(\boldsymbol{\pi})p(\mathbf{z}|\boldsymbol{\pi}) &= p(\boldsymbol{\pi}) \prod_n p(z_n|\boldsymbol{\pi}) \\ &= Dir(\boldsymbol{\pi}|\boldsymbol{\alpha}) \prod_n \prod_k \pi_k^{[z_n=k]} \\ &= Dir(\boldsymbol{\pi}|\boldsymbol{\alpha}) \prod_k \pi_k^{m_k} \\ &= Dir(\boldsymbol{\pi}|\boldsymbol{\alpha}') \end{aligned} \tag{1}$$

where $m_k = \sum_n [z_n = k]$ and $\alpha'_k = \alpha_k + m_k$.

Let's compute now the posterior distribution over $\boldsymbol{\theta}_k$:

$$\begin{aligned} p(\boldsymbol{\theta}_k|\mathbf{x}, \mathbf{z}) &\propto p(\boldsymbol{\theta}_k)p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}_k) \\ &= Dir(\boldsymbol{\theta}_k|\boldsymbol{\gamma}) \prod_n \prod_j Cat(x_{nj}|\boldsymbol{\theta}_k)^{[z_n=k]} \\ &= Dir(\boldsymbol{\theta}_k|\boldsymbol{\gamma}) \prod_n \prod_j \prod_m \boldsymbol{\theta}_{km}^{[x_{nj}=m][z_n=k]} \\ &= Dir(\boldsymbol{\theta}_k|\boldsymbol{\gamma}) \prod_m \boldsymbol{\theta}_{km}^{c_{km}} \\ &= Dir(\boldsymbol{\theta}_k|\boldsymbol{\gamma}'_k) \end{aligned} \tag{2}$$

where $c_{km} = \sum_n [z_n = k] \sum_j [x_{nj} = m]$ represents the number of occurrences of the m-th word in the cluster k, and $\gamma'_{km} = \gamma_m + c_{km}$.

Finally, the posterior distribution over z_n is given by:

$$p(z_n = k|\mathbf{x}_n, \boldsymbol{\pi}, \boldsymbol{\theta}) = \frac{p(z_n = k, \mathbf{x}_n|\boldsymbol{\pi}, \boldsymbol{\theta})}{\sum_{k'} p(z_n = k', \mathbf{x}_n|\boldsymbol{\pi}, \boldsymbol{\theta})} = \frac{\pi_k p(\mathbf{x}_n|z_n = k, \boldsymbol{\theta}_k)}{\sum_{k'} \pi_{k'} p(\mathbf{x}_n|z_n = k', \boldsymbol{\theta}_{k'})}. \tag{3}$$

2. Algorithm 2 approximates (using samples) the posterior distribution $p(\mathbf{z}, \boldsymbol{\theta}|\mathbf{x})$. Derive the conditional distributions needed to sample in steps (1) and (2) of the $\boldsymbol{\pi}$ collapsed Gibbs sampling algorithm.

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

```
Initialize cluster assignments $\mathbf{Z}$ and model parameters $\boldsymbol{\theta}$;
for $\tau = 1, \dots, N_{it}$ do
    for $k = 1, \dots, K$ do
        Sample $\boldsymbol{\theta}_k \sim p(\boldsymbol{\theta}_k|\mathbf{x}, \mathbf{z})$ ;    (1)
    end
    for $n = 1, \dots, N$ do
        Sample $z_n \sim p(z_n|\mathbf{x}, \mathbf{z}_{-n}, \boldsymbol{\theta}) = p(z_n|\mathbf{x}_n, \mathbf{z}_{-n}, \boldsymbol{\theta})$;    (2)
    end
end
```

The posterior distribution over $\boldsymbol{\theta}_k$ is the same as the one computed in equation (2).

Let's now compute the posterior distribution over z_n:

$$p(z_n = k|\mathbf{x}_n, \mathbf{z}_{-n}, \boldsymbol{\theta}) \propto p(z_n = k, \mathbf{x}_n|\mathbf{z}_{-n}, \boldsymbol{\theta}) = p(\mathbf{x}_n|z_n = k, \boldsymbol{\theta})p(z_n = k|\mathbf{z}_{-n}) \tag{4}$$

which must be normalized. The likelihood term $p(\mathbf{x}_n|z_n = k, \boldsymbol{\theta})$ is given by the usual product over Categorical distributions, i.e. $\prod_j Cat(x_{nj}|\boldsymbol{\theta}_k)$. The prior term, instead, can be computed as following:

$$\begin{aligned}
p(z_n = k|\mathbf{z}_{-n}) &= \int p(z_n = k, \boldsymbol{\pi}|\mathbf{z}_{-n})d\boldsymbol{\pi} \\
&= \int p(z_n = k|\boldsymbol{\pi})p(\boldsymbol{\pi}|\mathbf{z}_{-n})d\boldsymbol{\pi} \\
&= \int \prod_k \pi_k^{[z_n=k]} Dir(\boldsymbol{\pi}|\boldsymbol{\alpha}'')d\boldsymbol{\pi} \\
&= \int \frac{1}{B(\boldsymbol{\alpha}'')} \prod_k \pi_k^{\alpha''_k + [z_n=k] - 1} d\boldsymbol{\pi} \\
&= \frac{B(\{\alpha''_k + [z_n = k]\})}{B(\boldsymbol{\alpha}'')} \\
&= \frac{\prod_k \mathit{\Gamma}(\alpha''_k + [z_n = k])}{\mathit{\Gamma}(\sum_k \alpha''_k + 1)} \frac{\mathit{\Gamma}(\sum_k \alpha''_k)}{\prod_k \mathit{\Gamma}(\alpha''_k)} \\
&= \frac{\mathit{\Gamma}(\alpha''_k + 1)}{\sum_k \alpha_k \mathit{\Gamma}(\sum_k \alpha''_k)} \frac{\mathit{\Gamma}(\sum_k \alpha''_k)}{\mathit{\Gamma}(\alpha''_k)} \\
&= \frac{\alpha''_k \mathit{\Gamma}(\alpha''_k)}{\sum_k \alpha''_k \mathit{\Gamma}(\sum_k \alpha''_k)} \frac{\mathit{\Gamma}(\sum_k \alpha''_k)}{\mathit{\Gamma}(\alpha''_k)} \\
&= \frac{\alpha''_k}{\sum_k \alpha''_k} = \frac{\sum_{i \neq n}[z_i = k] + \alpha_k}{N - 1 + \sum_k \alpha_k}
\end{aligned} \tag{5}$$

where we have defined $\alpha''_k = \alpha_k + \sum_{i \neq n}[z_i = k]$. We have also used the standard definition of the Beta function and the property of the Gamma function $\mathit{\Gamma}(x + 1) = x\mathit{\Gamma}(x)$.

3. Algorithm 3 approximates (using samples) the posterior distribution $p(\mathbf{z}|\mathbf{x})$. Derive the conditional distribution needed to sample in step (1) of the $\boldsymbol{\pi}, \boldsymbol{\theta}$ collapsed Gibbs sampling algorithm.

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

```
Initialize cluster assignments $\mathbf{Z}$;
for $\tau = 1, \ldots, N_{it}$ do
    for $n = 1, \ldots, N$ do
        Sample $z_n \sim p(z_n|\mathbf{x}, \mathbf{z}_{-n})$;    (1)
    end
end
```

The posterior distribution over z_n is the following:

$$p(z_n = k|\mathbf{x}, \mathbf{z}_{-n}) \propto p(z_n = k, \mathbf{x}_n|\mathbf{x}_{-n}, \mathbf{z}_{-n}) = p(z_n = k|\mathbf{z}_{-n})p(\mathbf{x}_n|z_n = k, \mathbf{x}_{-n}, \mathbf{z}_{-n}) \tag{6}$$

which must be normalized. The prior term $p(z_n = k|\mathbf{z}_{-n})$ is the same as the one computed in equation (5), while the posterior predictive can be computed marginalizing the likelihood parameters:

$$\begin{aligned}
p(\mathbf{x}_n|z_n = k, \mathbf{x}_{-n}, \mathbf{z}_{-n}) &= \int p(\mathbf{x}_n, \boldsymbol{\theta}_k|z_n = k, \mathbf{x}_{-n}, \mathbf{z}_{-n}) d\boldsymbol{\theta}_k \\
&= \int p(\mathbf{x}_n|z_n = k, \boldsymbol{\theta}_k) p(\boldsymbol{\theta}_k|\mathbf{x}_{-n}, \mathbf{z}_{-n}) d\boldsymbol{\theta}_k \\
&= \int \prod_j Cat(x_{nj}|\boldsymbol{\theta}_k) Dir(\boldsymbol{\theta}_k|\boldsymbol{\gamma}''_k) d\boldsymbol{\theta}_k \qquad (7) \\
&= \int \prod_j \prod_m \boldsymbol{\theta}_{km}^{[x_{nj}=m]} \frac{1}{B(\boldsymbol{\gamma}''_k)} \prod_m \boldsymbol{\theta}_{km}^{\gamma''_{km}-1} d\boldsymbol{\theta}_k \\
&= \frac{1}{B(\boldsymbol{\gamma}''_k)} \int \prod_m \boldsymbol{\theta}_{km}^{c_{nm}+\gamma''_{km}-1} d\boldsymbol{\theta}_k \qquad (8) \\
&= \frac{B(\boldsymbol{\gamma}''_k + \boldsymbol{c}_n)}{B(\boldsymbol{\gamma}''_k)}
\end{aligned}$$

where we have defined $\gamma''_{km} = \gamma_m + \sum_{i \neq n} [z_i = k] \sum_j [x_{ij} = m]$ and the quantity $c_{nm} = \sum_j [x_{nj} = m]$ represents the number of occurrences of the m-th word in document n. We can further develop the ratio between the two Beta functions as following:

$$\begin{aligned}
\frac{B(\boldsymbol{\gamma}''_k + \boldsymbol{c}_n)}{B(\boldsymbol{\gamma}''_k)} &= \frac{\prod_m \Gamma(\gamma''_{km} + c_{nm})}{\Gamma(\sum_m \gamma''_{km} + c_{nm})} \frac{\Gamma(\sum_m \gamma''_{km})}{\prod_m \Gamma(\gamma''_{km})} \\
&= \frac{\prod_m \prod_{i=0}^{c_{nm}-1} (\gamma''_{km} + i) \Gamma(\gamma''_{km})}{\Gamma(\sum_m \gamma''_{km} + W_n)} \frac{\Gamma(\sum_m \gamma''_{km})}{\prod_m \Gamma(\gamma''_{km})} \\
&= \frac{\prod_m \prod_{i=0}^{c_{nm}-1} (\gamma''_{km} + i) \Gamma(\gamma''_{km})}{\prod_{j=0}^{W_n-1} (\sum_m \gamma''_{km} + j) \, \Gamma(\sum_m \gamma''_{km})} \frac{\Gamma(\sum_m \gamma''_{km})}{\prod_m \Gamma(\gamma''_{km})} \\
&= \frac{\prod_m \prod_{i=0}^{c_{nm}-1} (\gamma''_{km} + i)}{\prod_{j=0}^{W_n-1} (\sum_m \gamma''_{km} + j)}. \qquad (9)
\end{aligned}$$

4. Open the jupyter notebook, and play around with the dataset.
5. Implement the EM algorithm.
6. Show the (approximated) log-likelihood, the ten most representative words for each topic using a wordcloud, and the ten most relevant documents for each topic.

Solution in the notebook *L3_ solution.ipynb*.