Advanced Probabilistic Machine Learning and Applications

Pablo Sánchez and Isabel Valera

1 Tutorial 4 (Solution Exercise 1)

Now, we are assuming the number of components is unbounded, $K \to \infty$, but it is constrained by the complexity of our dataset to a finite number K^+. Additionaly, we choose $\alpha_k = \frac{\alpha}{K}$ for convenience. We also choose to use a Dirichlet prior over the mixing coefficients of the form:

$$p(\pi|\alpha) = \mathrm{Dir}(\pi|(\alpha/K, \cdots, \alpha/K))$$

1. α: Concentration parameter related to the a priori knowledge about the number of clusters.
2. $H(\cdot)$: Prior distribution from which we are going to sample the likelihood parameters, i.e. base measure.
3. The DP provides a framework to have l samples from some distribution H with probability π_l.
4. The DP is the $\lim\limits_{K\to\infty}$ of a Dirichlet distribution.
5. K^+ grows/decreases over the iterations. When a cluster k disappears we cannot recover it in the future nor its likelihood parameters.

Algorithm 1: Gibbs sampling with π collapsed

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

Algorithm 1: π collapsed Gibbs sampling algorithm

```
Initialize K⁺ = 1, {z_i = 1}_{i=1}^N and θ_1 ~ p(θ|γ);
while not converged do
    for n = 1,...,N do
        Sample z_n ~ p(z_n|X, Z_{-n}, Θ) = p(z_n|x_n, Z_{-n}, Θ);
        if z_n = K⁺ + 1 then
            K⁺ += 1;
            θ_{K⁺} ~ p(θ|x_n)
        end
        If necessary, remove empty clusters;
    end
    for k = 1,...,K⁺ do
        Sample θ_k ~ p(θ_k|X, Z) ;
    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) for a finite number of components was computed in Tutorial 2

$$p(z_n = k|\boldsymbol{Z}_{-n}) = \frac{\sum_{i\neq n}[z_i = k] + \alpha_k}{N - 1 + \sum_k \alpha_k} = \frac{m_k + \alpha_k}{N - 1 + \sum_k \alpha_k} = \frac{m_k + \alpha/K}{N - 1 + \alpha}$$

where m_k is the number of obsevations, except for n, assigned to component k. Notice the resulting distribution follows the scheme "rich get richer". The challenge in this Tutorial is that we are considering $K \to \infty$. The good point is that we still know that $\lim_{K\to\infty} \sum_{k=1}^{K} p(z_n = k|\boldsymbol{Z}_{-n}) = 1$. We can expand this limit in two terms: one with the clusters with observations an other with the clusters without any observation

$$\begin{aligned}
\lim_{K\to\infty} \sum_{k=1}^{K} \frac{m_k + \alpha/K}{N - 1 + \alpha} &= \lim_{K\to\infty} \sum_{k=1}^{K^+} \frac{m_k + \alpha/K}{N - 1 + \alpha} + \sum_{k=K^+ + 1}^{K} \frac{m_k + \alpha/K}{N - 1 + \alpha} \\
&= \sum_{k=1}^{K^+} \frac{m_k}{N - 1 + \alpha} + \lim_{K\to\infty} \sum_{k=K^+ + 1}^{K} \frac{\alpha/K}{N - 1 + \alpha} = 1
\end{aligned}$$

so now we can compute the probability of the infinity number of empty clusters

$$\lim_{K\to\infty} \sum_{k=K^+ + 1}^{K} \frac{\alpha/K}{N - 1 + \alpha} = 1 - \sum_{k=1}^{K^+} \frac{m_k}{N - 1 + \alpha} = 1 - \frac{N - 1}{N - 1 + \alpha}$$

Finally, we have that

$$p(z_n = k|\boldsymbol{Z}_{-n}) = \frac{m_k}{n - 1 + \alpha} \qquad p(z_n = K^+ + 1|\boldsymbol{Z}_{-n}) = \frac{\alpha}{n - 1 + \alpha}$$

Then, the posterior of a non-empty component is of the form

$$\begin{aligned}
p(z_n = k|\mathbf{x}_n, \boldsymbol{Z}_{-n}, \boldsymbol{\Theta}) &\propto \frac{m_k}{n - 1 + \alpha} p(\mathbf{x}_n|z_n = k, \boldsymbol{Z}_{-n}, \boldsymbol{\Theta}) \\
&\propto \frac{m_k}{n - 1 + \alpha} p(\mathbf{x}_n|\boldsymbol{\theta}_k)
\end{aligned}$$

The likelihood term has the form

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

On the contrary, the posterior of an empty component is of the form

$$\begin{aligned}
p(z_n = K_{\mathrm{new}}|\mathbf{x}_n, \boldsymbol{Z}_{-n}, \boldsymbol{\Theta}) &\propto \frac{\alpha}{n - 1 + \alpha} p(\mathbf{x}_n|z_n = K_{\mathrm{new}}, \boldsymbol{Z}_{-n}, \boldsymbol{\Theta}) \\
&= \frac{\alpha}{n - 1 + \alpha} p(\mathbf{x}_n|z_n = K_{\mathrm{new}}) \\
&= \frac{\alpha}{n - 1 + \alpha} \int p(\mathbf{x}_n, \boldsymbol{\theta}|z_n = K_{\mathrm{new}}) d\boldsymbol{\theta} \\
&= \frac{\alpha}{n - 1 + \alpha} \int p(\mathbf{x}_n|z_n = K_{\mathrm{new}}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\gamma}) d\boldsymbol{\theta} \qquad (1)
\end{aligned}$$

where we need to integrate all the possible likelihood parameters,$\boldsymbol{\theta}$,that can be used in the new component K_{new}. Thus, we need to compute the integral

$$
\begin{aligned}
\int p(\mathbf{x}_n|z_n = K_{\text{new}}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\gamma})d\boldsymbol{\theta} = & \\
&= \int \prod_{j=1}^{W_n} \text{Cat}(x_{nj}|\boldsymbol{\theta})\text{Dir}(\boldsymbol{\theta}|\boldsymbol{\gamma})d\boldsymbol{\theta} \\
&= \int \prod_{j=1}^{W_n} \prod_{m=1}^{|I|} \boldsymbol{\theta}_m^{[x_{nj}=m]} \frac{1}{B(\boldsymbol{\gamma})} \prod_{m=1}^{|I|} \boldsymbol{\theta}_m^{\gamma_m - 1} d\boldsymbol{\theta} \\
&= \int \frac{1}{B(\boldsymbol{\gamma})} \prod_{m=1}^{|I|} \boldsymbol{\theta}_m^{\gamma_m + c_{nm} - 1} d\boldsymbol{\theta} \\
&= \frac{B(\boldsymbol{\gamma} + \mathbf{c}_n)}{B(\boldsymbol{\gamma})} \\
&= \frac{\prod_{m=1}^{|I|} \prod_{i=0}^{c_{nm}-1} (\gamma_m + i)}{\prod_{j=0}^{W_n - 1} \left(\sum_m \gamma_m + j\right)}
\end{aligned} \tag{2}
$$

where $c_{nm} = \sum_{j=1}^{W_n} [x_{nj} = m]$. Note we can compute this a priori and save a bunch of computational resources.

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

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

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. Finally, for a new component we will sample $\boldsymbol{\theta}_{K_{\text{new}}}$

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{x}_n) &\propto p(\boldsymbol{\theta})p(\mathbf{x}_n|\boldsymbol{\theta}) \\
&= \text{Dir}(\boldsymbol{\theta}|\boldsymbol{\gamma})p(\mathbf{x}_n|\boldsymbol{\theta}) \\
&= \text{Dir}(\boldsymbol{\theta}|\boldsymbol{\gamma}_n)
\end{aligned}
$$

where d $\gamma_{nm} = \gamma_m + \sum_j [x_{nj} = m]$. We choose to sample from the posterior instead of the prior because otherwise the sampled $\boldsymbol{\theta}$ will probably not explain our sample correctly.

Summary

$$p(\boldsymbol{\theta}_k|\boldsymbol{X},\boldsymbol{Z}) = \text{Dir}(\boldsymbol{\theta}_k|\boldsymbol{\gamma}'_k) \qquad \gamma'_{km} = \gamma_m + \sum_n [z_n = k] \sum_j [x_{nj} = m] \tag{5}$$

$$p(\boldsymbol{\theta}_{K_{\text{new}}}|\mathbf{x}_n) = \text{Dir}(\boldsymbol{\theta}|\boldsymbol{\gamma}_n) \qquad \gamma_{nm} = \gamma_m + \sum_j [x_{nj} = m] \tag{6}$$

$$p(z_n = k|\mathbf{x}_n, \boldsymbol{Z}_{-n}, \boldsymbol{\Theta}) \propto \frac{m_k}{n-1+\alpha} p(\mathbf{x}_n|\boldsymbol{\theta}_k) \qquad m_k = \sum_{i \neq n} [z_i = k] \tag{7}$$

$$p(z_n = K_{\text{new}}|\mathbf{x}_n, \boldsymbol{Z}_{-n}, \boldsymbol{\Theta}) \propto \frac{\alpha}{N-1+\alpha} \frac{B(\boldsymbol{\gamma} + \mathbf{c}_n)}{B(\boldsymbol{\gamma})} \qquad c_{nm} = \sum_{j=1}^{W_n} [x_{nj} = m]$$

Algorithm 2: Gibbs sampling with π and Θ collapsed

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

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

---

```
Initialize K+ = 1 and {z_i = 1}_{i=1}^N;
while not converged do
    for n = 1,...,N do
        Sample z_n ~ p(z_n|X, Z_{-n});
        if z_n = K_new then
            K+ += 1;
        end
        If necessary, remove empty clusters;
    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 $\sum_k p(z_n = k|\mathbf{x}_n, \boldsymbol{X}_{-n}, \boldsymbol{Z}_{-n}) = 1$. We need to distinguish between existing ones. For new clusters it can be computed using Equation 1 in Exercise 2.

$$p(z_n = K_{\text{new}}|\boldsymbol{X}, \boldsymbol{Z}_{-n}) \propto \frac{\alpha}{n-1+\alpha} \frac{B(\boldsymbol{\gamma} + \mathbf{c}_n)}{B(\boldsymbol{\gamma})}$$

For existing clusters, the prior term was computed previously. The posterior predictive can be computed marginalizing the likelihood parameters

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

where we have used the result in Exercise 1 to get $\gamma_{km}''(n) = \gamma_m + \sum_{i\neq n}[\mathbf{z}_i = k]\sum_j[\mathbf{x}_{ij} = m]$ 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

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

We compute the log posterior predictive to avoid numerical instabilities

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

Summary

$$
p(z_n=k|\mathbf{x}_n,\boldsymbol{X}_{-n},\boldsymbol{Z}_{-n}) \propto \frac{m_k}{N-1+\alpha}\frac{B(\boldsymbol{\gamma}_k''(n)+\mathbf{c}_n)}{B(\boldsymbol{\gamma}_k''(n))} \qquad \gamma_{km}''(n) = \gamma_m + \sum_{i\neq n}[\mathbf{z}_i=k]\sum_j[\mathbf{x}_{ij}=m]
$$

$$
p(z_n=K_{\text{new}}|\mathbf{x}_n,\boldsymbol{X}_{-n},\boldsymbol{Z}_{-n}) \propto \frac{\alpha}{N-1+\alpha}\frac{B(\boldsymbol{\gamma}+\mathbf{c}_n)}{B(\boldsymbol{\gamma})}
$$