Machine Learning 2 - Homework 5

Sindy Löwe (11594969)

May 6, 2018

Collaborators: Pascal Esser

Problem 1

1. For deriving the update rules, we take the derivative of the expected value of the complete-data log-likelihood with regard to the parameter to be updated and set it to zero. Meanwhile, we keep the term $\gamma(z_{nk})$ fixed and treat it as a constant.

   For deriving the update rule for $\boldsymbol{\pi}$, we need to introduce a Lagrangian Multiplier in order to satisfy the constraint $\sum_{k=1}^{K} \boldsymbol{\pi}_k = 1$:

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\text{posterior}} &= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \ln \boldsymbol{\pi}_k + \lambda(\sum_{k=1}^{K} \boldsymbol{\pi}_k - 1) + \text{const} \\
\frac{\partial \, \mathbb{E}_{\text{posterior}}}{\partial \boldsymbol{\pi}_k} &= \sum_{n=1}^{N} \frac{\gamma(z_{nk})}{\boldsymbol{\pi}_k} + \lambda = 0 \\
N_k &= -\lambda \boldsymbol{\pi}_k \\
\sum_{k=1}^{K} N_k &= -\sum_{k=1}^{K} \lambda \boldsymbol{\pi}_k \\
N &= -\lambda \\
\boldsymbol{\pi}_k &= \frac{N_k}{N}
\end{aligned}
$$

   where $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$ is the effective number of data points associated with component k.

For $\boldsymbol{\mu}$ we get:

$$\begin{aligned}
\underset{\text{posterior}}{\mathbb{E}} &= \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\left[-\frac{1}{2}(\mathbf{x}_n^T\boldsymbol{\Sigma}_k^{-1}\mathbf{x}_n - 2\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\mathbf{x}_n + \boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k)\right] + \text{const} \\
\frac{\partial\,\mathbb{E}_{\text{posterior}}}{\partial\boldsymbol{\mu}_k} &= -\frac{1}{2}\sum_{n=1}^{N}\gamma(z_{nk})(-2\boldsymbol{\Sigma}_k^{-1}\mathbf{x}_n + 2\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k) = 0 \\
\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{\Sigma}_k^{-1}\mathbf{x}_n &= \sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k \\
\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{\Sigma}_k^{-1}\mathbf{x}_n &= N_k\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k \\
\boldsymbol{\mu}_k &= \frac{\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n}{N_k}
\end{aligned}$$

And for $\boldsymbol{\Sigma}$ (using equations (57) and (61) from the Matrix Cookbook and $\boldsymbol{\Sigma}_k\boldsymbol{\Sigma}_k^{-1} = I$):

$$\begin{aligned}
\underset{\text{posterior}}{\mathbb{E}} &= -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\left[\ln(|\boldsymbol{\Sigma}_k|) + (\mathbf{x}_n - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)\right] + \text{const} \\
\frac{\partial\,\mathbb{E}_{\text{posterior}}}{\partial\boldsymbol{\Sigma}_k} &= -\frac{1}{2}\sum_{n=1}^{N}\gamma(z_{nk})\left[\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}\right] = 0 \\
\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{\Sigma}_k^{-1} &= \sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1} \\
\boldsymbol{\Sigma}_k N_k\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_k &= \sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{\Sigma}_k\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_k \\
\boldsymbol{\Sigma}_k &= \frac{\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T}{N_k}
\end{aligned}$$

2. When constraining all covariance matrices to have a common value $\boldsymbol{\Sigma}$, the update rules for $\boldsymbol{\pi}$ and $\boldsymbol{\mu}$ remain the same as they are not dependent on $\boldsymbol{\Sigma}$. For the update rule of $\boldsymbol{\Sigma}$, we get:

$$\begin{aligned}
\underset{\text{posterior}}{\mathbb{E}} &= -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\left[\ln(|\boldsymbol{\Sigma}|) + (\mathbf{x}_n - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)\right] + \text{const} \\
\frac{\partial\,\mathbb{E}_{\text{posterior}}}{\partial\boldsymbol{\Sigma}} &= -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\left[\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}\right] = 0 \\
\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\boldsymbol{\Sigma}^{-1} &= \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1} \\
\boldsymbol{\Sigma} &= \frac{\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T}{N}
\end{aligned}$$

Where $\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk}) = \sum_{k=1}^{K}N_k = N$.

Problem 2

Using the dependencies as indicated in the graphical model, the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$ can be rewritten as:

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{X}) &\propto p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\
&= \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\
&\propto \ln \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})
\end{aligned}
$$

When applying the EM-Algorithm, we maximize the posterior over the latent variables $\mathbf{Z}$ in the E-Step, while keeping the parameters $\boldsymbol{\theta}$ fixed. This gives us:

$$
\begin{aligned}
\arg\max_{\mathbf{Z}} p(\boldsymbol{\theta}|\mathbf{X}) &\propto \arg\max_{\mathbf{Z}} \ln \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) \\
&\propto \arg\max_{\mathbf{Z}} \ln \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})
\end{aligned}
$$

Since the prior $p(\boldsymbol{\theta})$ is independent of $\mathbf{Z}$, we can drop the term and end up with a maximization over the log-likelihood. Therefore, the E-Step remains the same as in the maximum likelihood case.

In the M-Step, we maximize the posterior over the parameters $\boldsymbol{\theta}$, while keeping the latent variables $\mathbf{Z}$ fixed. This gives us:

$$
\arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{X}) \propto \arg\max_{\boldsymbol{\theta}} \ln \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})
$$

Here, we want to maximize over the complete-data log-likelihood. In practice, however, we are not given the complete data set $\{\mathbf{X},\mathbf{Z}\}$. The only knowledge that we have about the latent variables $\mathbf{Z}$ is given only by the posterior distribution $p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})$. As we can not make use of the complete-data log-likelihood directly, we use its expected value under the posterior distribution of the latent variables instead. This gives us the quantity to be maximized in the M-Step as follows:

$$
\begin{aligned}
\arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{X}) &\propto \arg\max_{\boldsymbol{\theta}} \mathop{\mathbb{E}}_{p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}})} \left[\ln \sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) \right] + \ln p(\boldsymbol{\theta}) \\
&\approx \arg\max_{\boldsymbol{\theta}} \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})
\end{aligned}
$$

Problem 3

For the M-Step, we need to derive the update rules for $\boldsymbol{\pi}$ and $\boldsymbol{\mu}$. For this, we use the log-posterior:

$$
\begin{aligned}
\ln p(\boldsymbol{\mu}, \boldsymbol{\pi} | \{x_n\}_{n=1}^N) \propto & \sum_{n=1}^N \sum_{k=1}^K \gamma(z_{nk}) \left(\ln \boldsymbol{\pi}_k + \sum_{i=1}^D [x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1 - \mu_{ki})] \right) \\
& + \sum_{k=1}^K \ln p(\boldsymbol{\mu}_k | a_k, b_k) + \ln p(\pi | \boldsymbol{\alpha}) \\
= & \sum_{n=1}^N \sum_{k=1}^K \gamma(z_{nk}) \left(\ln \boldsymbol{\pi}_k + \sum_{i=1}^D [x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1 - \mu_{ki})] \right) \\
& + \sum_{k=1}^K (a_k - 1) \ln \boldsymbol{\mu}_k + (b_k - 1) \ln(1 - \boldsymbol{\mu}_k) + \sum_{k=1}^K (\alpha_k - 1) \ln \boldsymbol{\pi}_k + \text{const}
\end{aligned}
$$

For the update rule for $\boldsymbol{\pi}$, we need to introduce a Lagrangian Multiplier in order to fulfill the constraint $\sum_{k=1}^K \boldsymbol{\pi}_k = 1$:

$$
\begin{aligned}
\frac{\partial \ln p(\boldsymbol{\mu}, \boldsymbol{\pi} | \{x_n\}_{n=1}^N) - \lambda(\sum_{k=1}^K \boldsymbol{\pi}_k - 1)}{\partial \boldsymbol{\pi}_k} &= \frac{\sum_{n=1}^N \gamma(z_{nk})}{\boldsymbol{\pi}_k} + \frac{\alpha_k - 1}{\boldsymbol{\pi}_k} - \lambda = 0 \\
N_k + \alpha_k - 1 &= \lambda \pi_k \\
\sum_{k=1}^K (N_k + \alpha_k - 1) &= \lambda \\
N + \sum_{k=1}^K \alpha_k - K &= \lambda \\
\boldsymbol{\pi}_k &= \frac{N_k + \alpha_k - 1}{N + \sum_{k=1}^K \alpha_k - K}
\end{aligned}
$$

For the update rule of $\boldsymbol{\mu}_k$ we get:

$$
\begin{aligned}
\frac{\partial \ln p(\boldsymbol{\mu}, \boldsymbol{\pi} | \{x_n\}_{n=1}^N)}{\partial \boldsymbol{\mu}_k} &= \sum_{n=1}^N \gamma(z_{nk}) \left(\sum_{i=1}^D \frac{x_{ni}}{\mu_{ki}} - \frac{1 - x_{ni}}{1 - \mu_{ki}} \right) + \frac{a_k - 1}{\boldsymbol{\mu}_k} - \frac{b_k - 1}{1 - \boldsymbol{\mu}_k} = 0 \\
\sum_{n=1}^N \gamma(z_{nk}) \left(\frac{\mathbf{x}_n}{\boldsymbol{\mu}_k} \right) + \frac{a_k - 1}{\boldsymbol{\mu}_k} &= \sum_{n=1}^N \gamma(z_{nk}) \left(\frac{1 - \mathbf{x}_n}{1 - \boldsymbol{\mu}_k} \right) + \frac{b_k - 1}{1 - \boldsymbol{\mu}_k} \\
(1 - \boldsymbol{\mu}_k) \left(\sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n + a_k - 1 \right) &= \boldsymbol{\mu}_k \left(\sum_{n=1}^N \gamma(z_{nk})(1 - \mathbf{x}_n) + b_k - 1 \right) \\
\sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n + a_k - 1 &= \boldsymbol{\mu}_k \left(\sum_{n=1}^N \gamma(z_{nk}) + b_k - 1 + a_k - 1 \right) \\
\boldsymbol{\mu}_k &= \frac{\sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n + a_k - 1}{N_k + b_k + a_k - 2}
\end{aligned}
$$