Philipp Lintl
12152498
philipp.lintl@student.uva.nl

Homework Assignment 5
Machine Learning 2, 19/20

2019-10-07

I worked on it alone, Group E.

1 Problem 1

1. In order to obtain the update rules for the asked parameters, we derive the expected value of the complete data log-likelihood with regard to the respective parameters and set it zero. In the same time, the term $\gamma(z_{nk})$ is kept fixed and treated as a constant. In order to solve the entire problem, we have to draw on Lagrangian Multipliers with the following form:

$$
\begin{aligned}
F &= \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma\left(z_{nk}\right)\left\{\ln\pi_k + \ln N\left(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k\right)\right\} + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma\left(z_{nk}\right)\left\{\ln\pi_k - \frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| - \frac{1}{2}\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^T\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)\right\} + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right) + const.,
\end{aligned}
$$

where the term on the right is the introduced constraint which is necessary for Lagrangian optimization with constraints. It says: $\sum_{k=1}^{K}\pi_k = 1$.
Now, we can start deriving this full Lagrangian with respect to the individual parameters to end up with their update rules:

$$
\begin{aligned}
\frac{\partial F}{\partial\pi_k} &= \sum_{n=1}^{N}\frac{\gamma\left(z_{nk}\right)}{\pi_k} + \lambda = 0 \\
N_k &= -\lambda\pi_k \\
\sum_{k=1}^{K}N_k &= -\sum_{k=1}^{K}\lambda\pi_k \\
N &= -\lambda \\
\pi_k &= \frac{N_k}{N},
\end{aligned}
$$

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

For μ: At first, we can simplify the F term for the terms being affected by the derivative regarding μ and also multiply the quadratic term:

$$
F_\mu = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma\left(z_{nk}\right)\left[-\frac{1}{2}\left(\mathbf{x}_n^T\boldsymbol{\Sigma}_k^{-1}\mathbf{x}_n - 2\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\mathbf{x}_n + \boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k\right)\right] + \text{ const },
$$

which is then again derived for the respective parameter as such:

$$\frac{\partial F_\mu}{\partial \boldsymbol{\mu}_k} = -\frac{1}{2}\sum_{n=1}^{N}\gamma\left(z_{nk}\right)\left(-2\boldsymbol{\Sigma}_k^{-1}\mathbf{x}_n + 2\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k\right) = 0$$

$$\sum_{n=1}^{N}\gamma\left(z_{nk}\right)\boldsymbol{\Sigma}_k^{-1}\mathbf{x}_n = \sum_{n=1}^{N}\gamma\left(z_{nk}\right)\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k$$

$$\sum_{n=1}^{N}\gamma\left(z_{nk}\right)\boldsymbol{\Sigma}_k^{-1}\mathbf{x}_n = N_k\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma\left(z_{nk}\right)\mathbf{x}_n$$

Similarly for the Covariance matrix $\boldsymbol{\Sigma}_k$ a form of the Lagrangian with only relevant terms remaining:

$$F_{\Sigma_k} = -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma\left(z_{nk}\right)\left[\ln\left(|\boldsymbol{\Sigma}_k|\right) + \left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^T\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)\right] + \text{ const}$$

Then the derivation with setting zero asrives as:

$$\frac{\partial F_{\Sigma_k}}{\partial \boldsymbol{\Sigma}_k} = -\frac{1}{2}\sum_{n=1}^{N}\gamma\left(z_{nk}\right)\left[\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^T\boldsymbol{\Sigma}_k^{-1}\right] = 0$$

$$\sum_{n=1}^{N}\gamma\left(z_{nk}\right)\boldsymbol{\Sigma}_k^{-1} = \sum_{n=1}^{N}\gamma\left(z_{nk}\right)\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^T\boldsymbol{\Sigma}_k^{-1}$$

$$\boldsymbol{\Sigma}_k N_k\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_k = \sum_{n=1}^{N}\gamma\left(z_{nk}\right)\boldsymbol{\Sigma}_k\boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Sigma}_k$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma\left(z_{nk}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^T$$

I used the following formulas in this order: matrix cookbook equation (57): $\frac{\partial \ln|\det(\mathbf{X})|}{\partial \mathbf{X}} = \left(\mathbf{X}^{-1}\right)^T = \left(\mathbf{X}^T\right)^{-1}$. Moreover, matrix cookbook equation (61): $\frac{\partial \mathbf{a}^T\mathbf{X}^{-1}\mathbf{b}}{\partial \mathbf{X}} = -\mathbf{X}^{-T}\mathbf{a}\mathbf{b}^T\mathbf{X}^{-T}$ and $\boldsymbol{\Sigma}_\mathbf{k}\boldsymbol{\Sigma}_\mathbf{k}{}^{-1} = I$. In the third row, $\boldsymbol{\Sigma}_\mathbf{k}$ is multiplied from the left and the right to eliminate the inverse of this term.

2. Supposing, the covariance matrices are constrained to have the same value, the update rules of μ and π do not change, as they are independent of $\boldsymbol{\Sigma}$. Again, I give the respective Lagrangian with only the relevant terms remaining, now $\boldsymbol{\Sigma}$ instead of $\boldsymbol{\Sigma}_\mathbf{k}$:

$$F_{\boldsymbol{\Sigma}} = -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma\left(z_{nk}\right)\left[\ln(|\boldsymbol{\Sigma}|) + \left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)\right] + \text{ const}$$

Analogously applying the same derivation formulas as before, except multiplying $\boldsymbol{\Sigma}$ with no index k from the left and the right:

$$\frac{F_{\boldsymbol{\Sigma}_\mathbf{k}}}{\partial \boldsymbol{\Sigma}} = -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma\left(z_{nk}\right)\left[\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^T\boldsymbol{\Sigma}^{-1}\right] = 0$$

$$\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma\left(z_{nk}\right)\boldsymbol{\Sigma}^{-1} = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma\left(z_{nk}\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^T\boldsymbol{\Sigma}^{-1}$$

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma\left(z_{nk}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^T,$$

with in the last row: $\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma\left(z_{nk}\right) = \sum_{k=1}^{K}N_k = N$

2 Problem 2

Given the graphical model with its dependencies, the posterior distribution can be written as:

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

In the E-Step of the EM Algorithm, the posterior is maximized with respect to the latent variables, while at the same time the parameters θ are kept fixed:

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

As the prior is independent of the latent parameters, which are being maximized, it can be dropped. That leaves the maximization of the log-likelihood, which is the same as for Maximum Likelihood:

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

In the M-Step, the latent parameters are fixed, while at the same time maximizing the posterior with respect to the parameters θ. This time, the prior remains:

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

However, as the latent parameters are not given, maximizing over the full data log-likelihood is not possible. Instead, the only derivable info on the latent parameters is with respect to the posterior distribution: $p(\mathbf{Z}|\mathbf{X}, \theta)$. Therefore, the the left term in the M-step formula can has to approximated with the expected value under the posterior distribution of the latent variables, as we can derive this information. With that, we end up as such:

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

3 Problem 3

In the M-Step, update rules for π and μ are sought. At first, we formulate the log-posterior as such:

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

Now, the first distribution in the second row is the Beta distribution of parameter $\mu_{\mathbf{k}}$ and the second is the Dirichlet of π. Plugging those distributions in, we arrive at:

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

whereas const entails the Beta function $B(a_k, b_k)$ term of the beta distribution and $B(a_k)$ of the Dirichlet distribution. As we derive for μ and π, they do not matter and can be summarized as a constant.

As seen in the first Problem, for Lagrange Multipliers, we need to introduce a constraint. Again, with $\sum_{k=1}^{K} \pi_k = 1$:

$$F = \ln p \left(\boldsymbol{\mu}, \boldsymbol{\pi} | \{x_n\}_{n=1}^{N} \right) - \lambda \left(\sum_{k=1}^{K} \boldsymbol{\pi}_k - 1 \right)$$

- So at first deriving for π_k affects two terms of the full posterior:

$$\frac{\partial F}{\partial \boldsymbol{\pi}_k} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})}{\pi_k} + \frac{\alpha_k - 1}{\pi_k} - \lambda = 0$$

  Again, we can use $\sum_{n=1}^{N} \gamma(z_{nk}) = N_k$ to get:

$$\Rightarrow N_k + \alpha_k - 1 = \lambda \pi_k$$

  The Lagrange Multiplier therefore is:

$$\sum_{k=1}^{K} (N_k + \alpha_k - 1) = \lambda$$
$$N + \sum_{k=1}^{K} \alpha_k - K = \lambda$$

  So, finally the update rule for $\pi_{\mathbf{k}}$ is:

$$\pi_k = \frac{N_k + \alpha_k - 1}{N + \sum_{k=1}^{K} \alpha_k - K}$$

- The second update rule is derived for $\mu_{\mathbf{k}}$:

$$\frac{\partial \ln p \left(\boldsymbol{\mu}, \boldsymbol{\pi} | \{x_n\}_{n=1}^{N} \right)}{\partial \boldsymbol{\mu}_k} = \sum_{n=1}^{N} \gamma(z_{nk}) \left(\sum_{i=1}^{N} \frac{x_{ni}}{\mu_{ki}} - \frac{1 - x_{ni}}{1 - \mu_{ki}} \right) + \frac{a_k - 1}{\boldsymbol{\mu}_k} - \frac{b_k - 1}{1 - \boldsymbol{\mu}_k} = 0$$

  Rearranging yields:

$$\sum_{n=1}^{N} \gamma(z_{nk}) \left(\frac{\mathbf{x}_n}{\boldsymbol{\mu}_k} \right) + \frac{a_k - 1}{\boldsymbol{\mu}_k} = \sum_{n=1}^{N} \gamma(z_{nk}) \left(\frac{1 - \mathbf{x}_n}{1 - \boldsymbol{\mu}_k} \right) + \frac{b_k - 1}{1 - \boldsymbol{\mu}_k}$$

  Now, to get rid of the denominators, we multiply both $\mu_{\mathbf{k}}$ and $(1 - \mu_{\mathbf{k}})$ on both sides:

$$(1 - \boldsymbol{\mu}_k) \left(\sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n + a_k - 1 \right) = \boldsymbol{\mu}_k \left(\sum_{n=1}^{N} \gamma(z_{nk}) (1 - \mathbf{x}_n) + b_k - 1 \right)$$

Now, the term $\mu_{\mathbf{k}} \sum_{n=1}^{N} \gamma(z_{nk})$ cancels out

$$\rightarrow \sum_{n=1}^{N} \gamma\left(z_{nk}\right) \mathbf{x}_n + a_k - 1 = \boldsymbol{\mu}_k \left(\sum_{n=1}^{N} \gamma\left(z_{nk}\right) + b_k - 1 + a_k - 1 \right)$$

Therefore, the M-Step update for $\mu_{\mathbf{k}}$ is:

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \gamma\left(z_{nk}\right) \mathbf{x}_n + a_k - 1}{N_k + b_k + a_k - 2}$$