# Machine Learning 2 — Homework 5

Maurice Frank
11650656
maurice.frank@posteo.de

October 7, 2019

## Problem 1.

We have $p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_l \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

$$\mathbb{E}_{\text{posterior}}[\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}\{\ln \pi_k + \ln \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\})$$

**1.**

For the update of $\boldsymbol{\pi}$ we have to formulate the Lagrangian:

$$\begin{aligned}
\mathcal{L} &= \mathbb{E}_{\text{posterior}}[\ln p(\boldsymbol{X}, \boldsymbol{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] + \lambda \cdot (\sum_{k=1}^{K} \pi_k - 1) \\
\frac{\partial}{\partial \pi_k} \mathcal{L} &= \sum_{n=1}^{N} \frac{\partial}{\partial \pi_k} \gamma(z_{nk}) \ln \pi_k + \lambda \cdot \frac{\partial}{\partial \pi_k} (\pi_k - 1) \\
&= \sum_{n=1}^{N} \frac{\gamma(z_{nk})}{\pi_k} + \lambda \\
&= \frac{1}{\pi_k} \cdot (\sum_{n=1}^{N} \gamma(z_{nk})) + \lambda \\
&\equiv 0 \\
&\iff \\
\pi_k &= -\sum_{n=1}^{N} \frac{\gamma(z_{nk})}{\lambda} \\
&= -\frac{N_k}{\lambda} \\
\frac{\partial \mathcal{L}}{\partial \lambda} &= \sum_{k=1}^{K} \pi_k - 1 \\
&\equiv 0 \quad \implies \\
\sum_{k=1}^{K} \pi_k &= 1 \quad \implies \\
-\sum_{k=1}^{K} \frac{N_k}{\lambda} &= 1 \quad \implies \\
\lambda &= -\sum_{k=1}^{K} N_k = N \\
&\implies \\
\pi_k &= \frac{N_k}{N}
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\mu}_k}\mathbb{E}_{\text{posterior}}[\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi})] &= \sum_{n=1}^{N}\gamma(z_{nk})\frac{\partial}{\partial \boldsymbol{\mu}_k}\ln\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)) \\
&= \sum_{n=1}^{N}\gamma(z_{nk})\frac{\frac{1}{2}(\boldsymbol{x}_n^T\boldsymbol{\Sigma}_k^{-1}+\boldsymbol{\Sigma}_k^{-1}\boldsymbol{x}_n-2\cdot\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k)\cdot\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k))}{\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k))} \\
&= \sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{\Sigma}_k^{-1}\cdot(\boldsymbol{x}_n-\boldsymbol{\mu}_k) \\
&\equiv 0 \\
&\Longrightarrow \\
\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k &= \sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{\Sigma}_k^{-1}\boldsymbol{x}_n \\
&\Longleftrightarrow \\
\mu_k &= \frac{1}{\sum_{n=1}^{N}\gamma(z_{nk})}\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{x}_n \\
&= \frac{\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{x}_n}{\sum_{n=1}^{N}\gamma(z_{nk})} \\
&= \frac{\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{x}_n}{N_k}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \mathbb{E}_{\text{posterior}}[\ln p(\boldsymbol{X}, \boldsymbol{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] &= \sum_{n=1}^{N} \gamma(z_{nk}) \cdot \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \log \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\
&= \sum_{n=1}^{N} \gamma(z_{nk}) \cdot \left( -\frac{1}{2} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \ln |\boldsymbol{\Sigma}_k| - \frac{1}{2} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k) \right) \\
&= \sum_{n=1}^{N} \gamma(z_{nk}) \cdot -\frac{1}{2} \left( \frac{|\boldsymbol{\Sigma}_k| \boldsymbol{\Sigma}_k^{-1}}{|\boldsymbol{\Sigma}_k|} - \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} \right) \\
&= \sum_{n=1}^{N} \gamma(z_{nk}) \cdot \frac{1}{2} \left( \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1} \right) \\
&\equiv 0 \\
&\Longrightarrow \\
\frac{N_k}{2} \boldsymbol{\Sigma}_k^{-1} &= \frac{1}{2} \sum_{n=1}^{N} \gamma(z_{nk}) \cdot \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} \\
&\Longrightarrow \\
\boldsymbol{\Sigma}_k &= \sum_{n=1}^{N} \frac{\gamma(z_{nk})}{N_k} \cdot (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T
\end{aligned}
$$

## 2.

The updates given for $\boldsymbol{\pi}_k$ and $\boldsymbol{\mu}_k$ do not depend on the covariance. Therefore these two update rules do not change.

For the update of the common $\boldsymbol{\Sigma}$ we have then:

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\Sigma}} \mathbb{E}_{\text{posterior}}[\ln p(\boldsymbol{X}, \boldsymbol{Z} | \boldsymbol{\mu}, \Sigma, \boldsymbol{\pi})] &= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \cdot \frac{\partial}{\partial \boldsymbol{\Sigma}} \log \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \cdot \frac{1}{2} \left( \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1} \right) \\
&\equiv 0 \\
&\Longrightarrow \\
\boldsymbol{\Sigma} &= \sum_{n=1}^{N} \sum_{k=1}^{K} y \frac{\gamma(z_{nk})}{N_k} \cdot (\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T
\end{aligned}
$$

## Problem 2.

$$\begin{aligned}
\ln p(\boldsymbol{\theta}|\boldsymbol{X}) &= \ln p(\boldsymbol{\theta}, \boldsymbol{X}) - \ln p(\boldsymbol{X}) \\
&= \ln p(\boldsymbol{X}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) - \ln p(\boldsymbol{X}) \\
&= \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}[q||p] + \ln p(\boldsymbol{\theta}) - \ln p(\boldsymbol{X}) \\
&\geqslant \mathcal{L}(q, \boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) - \ln p(\boldsymbol{X})
\end{aligned}$$

For the E-Step we maximize the lower bound w.r.t. $q$. As only $\mathcal{L}(q, \boldsymbol{\theta})$ is dependent on $q$ we have the same situation as in the E-step for the ML estimate.

For the M-step we maximize the lower bound w.r.t. $\boldsymbol{\theta}$:

$$\begin{aligned}
\arg\max_{\boldsymbol{\theta}} \mathcal{L}(q, \boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) - \ln p(\boldsymbol{X}) &= \arg\max_{\boldsymbol{\theta}} \mathcal{L}(q, \boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_{z} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}}) \cdot \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}})} + \ln p(\boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_{z} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}}) \cdot \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \\
&\quad - \sum_{z} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}}) \cdot \ln p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}}) + \ln p(\boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_{z} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) + H[p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}})] + \ln p(\boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \sum_{z} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})
\end{aligned}$$

The entropy of $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}})$ drops out as it does not depend on $\boldsymbol{\theta}$.

## Problem 3.

We have:

$$\begin{aligned}
\pi|\boldsymbol{\alpha} &\sim \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \\
\mathbf{z}_n|\boldsymbol{\pi} &\sim \mathrm{Mult}(\boldsymbol{z}_n|\boldsymbol{\pi}) \\
\mu_k|a_k, b_k &\sim \mathrm{Beta}(\boldsymbol{\mu}|a_k, b_k) \\
\mathbf{x}_n|\boldsymbol{z}_n, \boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\} &\sim \prod_{k=1}^{K} (\mathrm{Bern}(\boldsymbol{x}_n|\boldsymbol{\mu}_k))^{z_{nk}}
\end{aligned}$$

Further lets write some stuff that we gonna use later:

$$\ln p(\boldsymbol{\mu}) = \sum_{k=1}^{K}\sum_{j}^{D}(a_k - 1)\ln \mu_{kj} + (b_k - 1)\ln(1 - \mu_{kj}) - \ln \mathrm{B}(a_k, b_k)$$

$$\ln p(\boldsymbol{\pi}) = \sum_{k=1}^{K}(\alpha_k - 1)\ln \pi_k - \ln \mathrm{B}(\alpha_k)$$

$$\mathbb{E}_{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\mu}^{\text{old}},\boldsymbol{\pi}^{\text{old}})}[ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\mu},\boldsymbol{\pi})] = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\left\{\ln \pi_k + \sum_{i=1}^{D}[x_{ni}\ln \mu_{ki} + (1 - x_{ni})\ln(1 - \mu_{ki})]\right\}$$

with the last being already the result from the E-step as described in the book.

To calculate the updates for $\boldsymbol{\mu}$ and $\boldsymbol{\pi}$ in the M-step we can use the result from Problem 2:

$$\begin{aligned}
\arg\max_{\boldsymbol{\mu},\boldsymbol{\pi}} \mathcal{L}(q,\boldsymbol{\mu},\boldsymbol{\pi}) + \ln p(\boldsymbol{\theta}) + \ln p(\boldsymbol{\pi}) &= \arg\max_{\boldsymbol{\mu},\boldsymbol{\pi}} \sum_{z} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\mu}^{\text{old}},\boldsymbol{\pi}^{\text{old}})\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\mu},\boldsymbol{\pi}) + \ln p(\boldsymbol{\mu}) + \ln p(\boldsymbol{\pi}) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\left\{\ln \pi_k + \sum_{i=1}^{D}[x_{ni}\ln \mu_{ki} + (1 - x_{ni})\ln(1 - \mu_{ki})]\right\} \\
&\quad + \sum_{k=1}^{K}\sum_{j}^{D}(a_k - 1)\ln \mu_{kj} + (b_k - 1)\ln(1 - \mu_{kj}) - \ln \mathrm{B}(a_k, b_k) \\
&\quad + \sum_{k=1}^{K}(\alpha_k - 1)\ln \pi_k - \ln \mathrm{B}(\alpha_k) \\
&\quad + \lambda \cdot (\sum_{k=1}^{K}\pi_k - 1)
\end{aligned}$$

we already added the lagrange Multiplier for the condition of the Dirichlet distribution of $\pi$.

Now we maximize for the two variables.

First for $\boldsymbol{\mu}$:

$$
\begin{aligned}
\frac{\partial}{\partial \mu_{ki}} \mathcal{L}(q, \boldsymbol{\mu}, \boldsymbol{\pi}) + \ln p(\boldsymbol{\theta}) + \ln p(\boldsymbol{\pi}) &= \sum_{n=1}^{N} \gamma(z_{nk}) \left\{ x_{ni} \frac{1}{\mu_{ki}} - (1 - x_{ni}) \frac{1}{1 - \mu_{ki}} \right\} \\
&\quad + (a_k - 1) \frac{1}{\mu_{ki}} - (b_k - 1) \frac{1}{1 - \mu_{ki}} \\
&= \frac{1}{\mu_{ki}} \left( \sum_{n=1}^{N} \gamma(z_{nk}) x_{ni} + (a_k - 1) \right) \\
&\quad - \frac{1}{1 - \mu_{ki}} \left( \sum_{n=1}^{N} \gamma(z_{nk}) (1 - x_{ni}) + (b_k - 1) \right) \\
&\equiv 0 \\
&\Longrightarrow \\
\left( \sum_{n=1}^{N} \gamma(z_{nk}) x_{ni} + (a_k - 1) \right) &= \mu_{ki} \left[ \left( \sum_{n=1}^{N} \gamma(z_{nk}) (1 - x_{ni}) + (b_k - 1) \right) + \left( \sum_{n=1}^{N} \gamma(z_{nk}) x_{ni} + (a_k - 1) \right) \right] \\
\mu_{ki} &= \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_{ni} + (a_k - 1)}{\sum_{n=1}^{N} \gamma(z_{nk}) + b_k + a_k - 2} \\
&= \frac{\sum_{n=1}^{N} \gamma(z_{nk}) x_{ni} + (a_k - 1)}{N_k + b_k + a_k - 2}
\end{aligned}
$$

And lastly for $\boldsymbol{\pi}$:

$$
\begin{aligned}
\frac{\partial}{\partial \pi_k} \mathcal{L}(q, \boldsymbol{\mu}, \boldsymbol{\pi}) + \ln p(\boldsymbol{\theta}) + \ln p(\boldsymbol{\pi}) &= \sum_{n=1}^{N} \gamma(z_{nk}) \frac{1}{\pi_k} \\
&\quad + (\alpha_k - 1) \frac{1}{\pi_k} \\
&\quad + \lambda \\
&= \frac{1}{\pi_k} \cdot (N_k + \alpha_k - 1) + \lambda \\
&\equiv 0 \\
&\Longrightarrow \\
\pi_k &= -\frac{N_k + \alpha_k - 1}{\lambda} \\
&\text{and:} \\
\lambda &= \lambda \sum_{k=1}^{K} \pi_k \\
&= -(\sum_{k=1}^{K} N_k + \sum_{k=1}^{K} \alpha_k - \sum_{k=1}^{K} 1) \\
&= -(N + \sum_{k=1}^{K} \alpha_k - K) \\
&\Longrightarrow \\
\pi_k &= \frac{N_k + \alpha_k - 1}{N + \sum_{k=1}^{K} \alpha_k - K}
\end{aligned}
$$