Machine Learning 2 - Homework 5

Andrii Skliar, 11636785

deadline: May 7, 2018

During the process of solving the homework problems, I have collaborated with the following colleagues:

Davide Belli   Gabriele Bani

NB: credits for the Latex-format go to Iris Verweij, 2nd year MSc AI Student.

Problem 1: Consider a Gaussian mixture model

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

1. Given the expected value of the complete-data log-likelihood (9.40 in Bishops book)

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

Derive update rules for $\boldsymbol{\pi}, \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

Solution: To solve this problem, we need to first write down following Lagrangian:

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

Now, we can derive update rules for $\boldsymbol{\pi}, \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

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

$$\begin{aligned}
\frac{\partial F}{\partial \boldsymbol{\mu}_k} &= \{\text{using equation 86 from matrix cookbook}\} \\
&= \sum_{n=1}^{N} \gamma(z_{nk})(-\frac{1}{2} \cdot -2\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)) \\
&= \sum_{n=1}^{N} \gamma(z_{nk})(\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)) \\
&= \sum_{n=1}^{N} \gamma(z_{nk})(\boldsymbol{\Sigma}_k^{-1}\mathbf{x}_n) - \sum_{n=1}^{N} \gamma(z_{nk})(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k) = 0 \\
\Rightarrow N_k \cdot \boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k &= \sum_{n=1}^{N} \gamma(z_{nk})(\boldsymbol{\Sigma}_k^{-1}\mathbf{x}_n) \\
\boldsymbol{\mu}_k &= (N_k \cdot \boldsymbol{\Sigma}_k^{-1})^{-1} \sum_{n=1}^{N} \gamma(z_{nk})(\boldsymbol{\Sigma}_k^{-1}\mathbf{x}_n) \\
\boldsymbol{\mu}_k &= \frac{1}{N_k} \boldsymbol{\Sigma}_k \boldsymbol{\Sigma}_k^{-1} \sum_{n=1}^{N} \gamma(z_{nk})\mathbf{x}_n \\
\boldsymbol{\mu}_k &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})\mathbf{x}_n
\end{aligned}$$

$$
\begin{aligned}
\frac{\partial F}{\partial \boldsymbol{\Sigma}_k} &= \{\text{using equations 57 and 61 from matrix cookbook} \\
&\quad \text{and the fact that } \boldsymbol{\Sigma} \text{ is symmetric}\} \\
&= \sum_{n=1}^{N} -\frac{1}{2}\gamma(z_{nk})\{\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}\} = 0 \\
-\frac{1}{2}\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{\Sigma}_k^{-1} &= -\frac{1}{2}\sum_{n=1}^{N}\gamma(z_{nk})\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} \\
N_k \cancel{\boldsymbol{\Sigma}_k^{-1}} &= \boldsymbol{\Sigma}_k^{-1} \cdot \sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \cdot \cancel{\boldsymbol{\Sigma}_k^{-1}} \\
\boldsymbol{\Sigma}_k^{-1} &= N_k(\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T)^{-1} \\
\boldsymbol{\Sigma}_k &= \frac{\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T}{N_k}
\end{aligned}
$$

2. Consider a special case of the model above, in which the covariance matrices $\boldsymbol{\Sigma}_k$ of the components are all constrained to have a common value $\boldsymbol{\Sigma}$. Derive EM equations for maximizing the likelihood function under such a model.

Solution:

$$
\begin{aligned}
\frac{\partial F}{\partial \boldsymbol{\Sigma}} &= -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\{\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}\} = 0 \\
(\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk}))\cancel{\boldsymbol{\Sigma}^{-1}} &= \boldsymbol{\Sigma}^{-1}(\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T)\cancel{\boldsymbol{\Sigma}^{-1}} \\
N &= \boldsymbol{\Sigma}^{-1}(\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T) \\
\boldsymbol{\Sigma}^{-1} &= N(\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T)^{-1} \\
\boldsymbol{\Sigma} &= \frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T
\end{aligned}
$$

Problem 2: Suppose we wish to use the EM algorithm to maximize the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$ for a model containing latent variables $\mathbf{z}$ and observed variables $\mathbf{x}$. Show that the E step remains the same as in the maximum likelihood case, where as in the M step, the quantity to be maximized is

$$\sum_{\mathbf{z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})$$

Solution:

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

Due to the fact, that q only appears in $\mathcal{L}(q, \boldsymbol{\theta})$, E-step will stay the same.
M step update rule will be following:

$$\begin{aligned}
\mathcal{L}(q, \boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) - \ln p(\mathbf{X}) &= \mathcal{L}(q, \boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) + const \\
&= \sum_{\mathbf{z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) + \sum_{\mathbf{z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old}) \ln p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\theta}^{old}) + \ln p(\boldsymbol{\theta}) + const \\
&= \sum_{\mathbf{z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{old}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) + const
\end{aligned}$$

Problem 3: Derive the EM algorithm for maximizing the posterior probability $p(\boldsymbol{\mu}, \boldsymbol{\pi}|\mathbf{x}_n{}_{n=1}^N)$ of Mixtures of Bernoulli distribution. (The E step is given in Bishops Book, you only need to do the M step).

Solution: Using equations derived in task 2, we can write equations for M step as following:

$$\begin{aligned}
F &= \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})] + \ln p(\boldsymbol{\theta}) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \{\ln \pi_k + \sum_{i=1}^{D} [x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1 - \mu_{ki})]\} \\
&+ \sum_{k=1}^{K} [\ln Beta(\boldsymbol{\mu}_k | a_k, b_k) + \ln Dir(\pi_k | \alpha_k)] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \{\ln \pi_k + \sum_{i=1}^{D} [x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1 - \mu_{ki})]\} \\
&+ \sum_{k=1}^{K} \Big[\sum_{i=1}^{D} \Big\{ (a_k - 1) \ln \mu_{ki} + (b_k - 1)(1 - \mu_k i) \Big\} - \ln \mathcal{B}(a_k, b_k) \\
&- \ln B(\alpha_k) + (\alpha_k - 1) \ln \pi_k) \Big]
\end{aligned}$$

Now we can perform M step for $\boldsymbol{\mu}$ and $\boldsymbol{\pi}$:

$$\frac{\partial F}{\partial \mu_{ki}} = \sum_{n=1}^{N} \gamma(z_{nk})(\frac{x_{ni}}{\mu_{ki}} - \frac{1-x_{ni}}{1-\mu_{ki}}) + \frac{a_k - 1}{\mu_{ki}} - \frac{b_k - 1}{1-\mu_{ki}} = 0$$

$$\Big[\sum_{n=1}^{N} \gamma(z_{nk})\frac{x_{ni}}{\mu_{ki}}\Big] + \frac{a_k - 1}{\mu_{ki}} = \Big[\sum_{n=1}^{N} \gamma(z_{nk})\frac{1-x_{ni}}{1-\mu_{ni}}\Big] + \frac{b_k - 1}{1-\mu_{ki}}$$

$$\frac{1}{\mu_{ki}}\Big[\sum_{n=1}^{N} \gamma(z_{nk})x_{ni} + a_k - 1\Big] = \frac{1}{1-\mu_{ki}}\Big[\sum_{n=1}^{N} \gamma(z_{nk})(1-x_{ni}) + b_k - 1\Big]$$

$$\frac{1-\mu_{ki}}{\mu_{ki}} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(1-x_{ni}) + b_k - 1}{\sum_{n=1}^{N} \gamma(z_{nk})x_{ni} + a_k - 1}$$

$$\frac{1}{\mu_{ki}} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(1-x_{ni}) + b_k - 1 + \sum_{n=1}^{N} \gamma(z_{nk})x_{ni} + a_k - 1}{\sum_{n=1}^{N} \gamma(z_{nk})x_{ni} + a_k - 1}$$

$$\frac{1}{\mu_{ki}} = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) + b_k + a_k - 2}{\sum_{n=1}^{N} \gamma(z_{nk})x_{ni} + a_k - 1}$$

$$\mu_{ki} = \frac{\sum_{n=1}^{N} \gamma(z_{nk})x_{ni} + a_k - 1}{N_k + b_k + a_k - 2}$$

To perform M step for $\boldsymbol{\pi}$, we also need to write proper Lagrangian, which in that case will be:

$$L = F + \lambda(\sum_{k=1}^{K} \pi_k - 1)$$

Therefore, M step update will look as following:

$$\frac{\partial L}{\partial \pi_k} = (\sum_{n=1}^{N} \gamma(z_{nk}\frac{1}{\pi_k}) + \frac{\alpha_k - 1}{\pi_k} - \lambda \qquad\qquad = 0$$

$$\frac{N_k}{\pi_k} + \frac{\alpha_k - 1}{\pi_k} - \lambda = 0$$

$$\pi_k = \frac{N_k + \alpha_k - 1}{\lambda}$$

$$\frac{N_k + \alpha_k - 1}{\pi_k} - \lambda = 0$$

$$N_k + \alpha_k - 1 - \lambda\pi_k = 0$$

$$\sum_{k=1}^{K}\{N_k + \alpha_k - 1 - \lambda\pi_k\} = 0$$

$$N + \sum_{k=1}^{K} \alpha_k - K - \lambda = 0$$

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

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