Machine Learning 2 **Due: May 03, 2016**

Homework 4

Instructor: Ke Tran Email: m.k.tran@uva.nl

Student: Andrea Jemmett UvA ID: 11162929
Collaborators: N/A Email: andrea.jemmett@student.uva.nl

You are allowed to discuss with your colleagues but you should write the answers in *your own words*. If you discuss with others, write down the name of your collaborators on top of the first page. No points will be deducted for collaborations. If we find similarities in solutions beyond the listed collaborations we will consider it as cheating.
We will not accept any late submissions under any circumstances. The solutions to the previous homework will be handed out in the class at the beginning of the next homework session. After this point, late submissions will be automatically graded zero.

Problem 1. Consider a Gaussian mixture model

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

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

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

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

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.

Problem 2. Suppose we wish to use the EM algorithm to maximize the posterior distribution

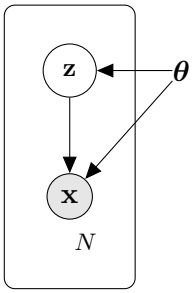

Figure 1: A simple generative model.

$p(\boldsymbol{\theta}|\mathbf{X})$ for a model (Figure 1) 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

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

Problem 3.

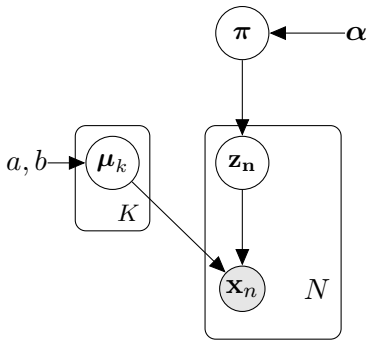

Figure 2: Mixtures of Bernoulli distribution

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

Derive the EM algorithm for maximizing the posterior probability $p(\boldsymbol{\mu}, \boldsymbol{\pi}|\{\mathbf{x}_n\}_{n=1}^N)$. (The E step is given in Bishop's Book, you only need to do the M step)

Solution: The M step consists in the maximization of the expectation of the complete log-likelihood. In the case of the model in Figure 2 it is given by

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\pi})] = \ln \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} + \sum_{k=1}^K \Bigg[& D \ln \frac{\Gamma(a_k + b_k)}{\Gamma(a_k)\Gamma(b_k)} + (\alpha_k - 1) \ln \pi_k \\
& + (a_k - 1) \sum_{i=1}^D \ln \mu_{ki} + (b_k - 1) \sum_{i=1}^D (1 - x_{ni}) \ln(1 - \mu_{ki}) \Bigg] \\
& + \sum_{n=1}^N \sum_{k=1}^K \gamma(z_{nk}) \left[\ln \pi_k + \sum_{i=1}^D x_{ni} \ln \mu_{ki} + \sum_{i=1}^D (1 - x_{ni})(1 - \mu_{ki}) \right]
\end{aligned} \tag{1}$$

In the M step we need to maximize the expectation of the complete data log-likelihood w.r.t. the

parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\pi}$. To do this we evaluate the partial derivatives w.r.t. μ_{ki}

$$\begin{aligned}
\frac{\partial}{\partial \mu_{ki}}\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X},\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\pi})] &= \frac{a_k-1}{\mu_{ki}} - \frac{b_k-1}{1-\mu_{ki}} + \sum_{n=1}^{N}\sum_{k=1}^{K}\{\gamma(z_{nk}) \\
&\quad \times \frac{\partial}{\partial \mu_{ki}}\left[\ln \pi_k + \sum_{i=1}^{D} x_{ni}\ln \mu_{ki} + \sum_{i=1}^{D}(1-x_{ni})\ln(1-\mu_{ki})\right]\Bigg\} && (2)\\
&= \frac{a_k-1}{\mu_{ki}} - \frac{b_k-1}{1-\mu_{ki}} + \sum_{n=1}^{N}\gamma(z_{nk})\left[\frac{x_{ni}}{\mu_{ki}} - \frac{1-x_{ni}}{1-\mu_{ki}}\right] && (3)\\
&= \frac{a_k-1}{\mu_{ki}} - \frac{b_k-1}{1-\mu_{ki}} + \frac{\sum_{n=1}^{N}\gamma(z_{nk})x_{ni}}{\mu_{ki}} - \frac{\sum_{n=1}^{N}(1-x_{ni})\gamma(z_{nk})}{1-\mu_{ki}} && (4)\\
&= \frac{1}{\mu_{ki}(1-\mu_{ki})}\Bigg[(1-\mu_{ki})(a_k-1) - \mu_{ki}(b_k-1) + (1-\mu_{ki})\sum_{n=1}^{N}\gamma(z_{nk})x_{ni} \\
&\quad -\mu_{ki}\sum_{n=1}^{N}\gamma(z_{nk})(1-x_{ni})\Bigg] && (5)\\
&= \frac{1}{\mu_{ki}(1-\mu_{ki})}\Bigg[a_k - 1 - \mu_{ki}a_k + \mu_{ki} - \mu_{ki}b_k + \mu_{ki} + \sum_{n=1}^{N}\gamma(z_{nk})x_{ni} \\
&\quad -\mu_{ki}\sum_{n=1}^{N}\gamma(z_{nk})x_{ni} - \mu_{ki}\sum_{n=1}^{N}\gamma(z_{nk}) + \mu_{ki}\sum_{n=1}^{N}\gamma(z_{nk})x_{ni}\Bigg] && (6)
\end{aligned}$$

and by setting this derivative to zero and solving for μ_{ki}

$$0 = a_k - 1 - \mu_{ki}a_k + \mu_{ki} - \mu_{ki}b_k + \mu_{ki} + \sum_{n=1}^{N}\gamma(z_{nk})x_{ni} - \mu_{ki}\sum_{n=1}^{N}\gamma(z_{nk}) \quad (7)$$

$$1 - a_k - \sum_{n=1}^{N}\gamma(z_{nk})x_{ni} = \mu_{ki}\left[2 - a_k - b_k - \sum_{n=1}^{N}\gamma(z_{nk})\right] \quad (8)$$

$$\mu_{ki} = \frac{1 - a_k - \sum_{n=1}^{N}\gamma(z_{nk})x_{ni}}{2 - a_k - b_k - \sum_{n=1}^{N}\gamma(z_{nk})} = \frac{\sum_{n=1}^{N}\gamma(z_{nk})x_{ni} + a_k - 1}{\sum_{n=1}^{N}\gamma(z_{nk}) + a_k + b_k - 2} \quad (9)$$

which rewritten in vector form gives

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n + a_k - 1}{\sum_{n=1}^{N}\gamma(z_{nk}) + a_k + b_k - 2} \quad (10)$$

Next we need to maximize the expectation of the log-likelihood w.r.t. π_k using Lagrange multipliers to enforce the constraint $\sum_{k=1}^{K}\pi_k = 1$

$$\begin{aligned}
\frac{\partial}{\partial \pi_k}\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X},\mathbf{Z},\boldsymbol{\mu},\boldsymbol{\pi})] &= \frac{\alpha_k - 1}{\pi_k} + \frac{\sum_{n=1}^{N}\gamma(z_{nk})}{\pi_k} + \lambda && (11)\\
&= \frac{\alpha_k - 1 + \sum_{n=1}^{N}\gamma(z_{nk}) + \lambda\pi_k}{\pi_k} && (12)
\end{aligned}$$

and by setting this derivative to zero and solving for π_k

$$0 = \alpha_k - 1 + \sum_{n=1}^{N} \gamma(z_{nk}) + \lambda \pi_k \tag{13}$$

$$\pi_k = \frac{1 - \alpha_k - \sum_{n=1}^{N} \gamma(z_{nk})}{\lambda} \tag{14}$$

Using the constraint on $\boldsymbol{\pi}$ we can find the value of λ

$$\sum_{k=1}^{K} \pi_k = 1 \tag{15}$$

$$\frac{1}{\lambda} \sum_{k=1}^{K} \left[1 - \alpha_k - \sum_{n=1}^{N} \gamma(z_{nk}) \right] = 1 \tag{16}$$

$$\lambda = K - \sum_{k=1}^{K} \alpha_k - \sum_{k=1}^{K} \sum_{n=1}^{N} \gamma(z_{nk}) \tag{17}$$

$$\lambda = K - \sum_{k=1}^{K} \alpha_k - N \tag{18}$$

substituting back we get

$$\pi_k = \frac{1 - \alpha_k - \sum_{n=1}^{N} \gamma(z_{nk})}{K - \sum_{k=1}^{K} \alpha_k - N} \tag{19}$$

■

Problem 4⋆. Bishop 10.38 Verify the results of the calculation of mean, variance of $q^{\backslash n}(\boldsymbol{\theta})$ and normalizing constant Z_n for the expectation propagation algorithm applied to the cluster problem

$$\mathbf{m}^{\backslash n} = \mathbf{m} + v^{\backslash n} v_n^{-1} (\mathbf{m} - \mathbf{m}_n) \tag{20}$$

$$(v^{\backslash n})^{-1} = v^{-1} - v_n^{-1} \tag{21}$$

$$Z_n = (1 - w) \mathcal{N}\left(\mathbf{x}_n | \mathbf{m}^{\backslash n}, (v^{\backslash n} + 1) \mathbf{I} \right) + w \mathcal{N}\left(\mathbf{x}_n | \mathbf{0}, a \mathbf{I} \right) \tag{22}$$

Hint: see 10.38 in the book for more hint. You need equation (2.115) in Bishop to solve this homework.

Solution: The first step is to use the division formula (eq. 10.205 in Bishop) to derive $q^{\backslash n}(\boldsymbol{\theta})$

$$
\begin{aligned}
q^{\backslash n}(\boldsymbol{\theta}) &= \frac{q(\boldsymbol{\theta})}{\widetilde{f}_n(\boldsymbol{\theta})} && (23)\\
&= \frac{\mathcal{N}(\boldsymbol{\theta}|\mathbf{m}, v\mathbf{I})}{s_n\mathcal{N}(\boldsymbol{\theta}|\mathbf{m}_n, v_n\mathbf{I})} && (24)\\
&\propto \exp\left\{\frac{1}{2v_n}(\boldsymbol{\theta}-\mathbf{m}_n)^\mathsf{T}(\boldsymbol{\theta}-\mathbf{m}_n) - \frac{1}{2v}(\boldsymbol{\theta}-\mathbf{m})^\mathsf{T}(\boldsymbol{\theta}-\mathbf{m})\right\} && (25)\\
&= \exp\left\{\frac{1}{2v_n}(\boldsymbol{\theta}^\mathsf{T}\boldsymbol{\theta} - 2\boldsymbol{\theta}^\mathsf{T}\mathbf{m}_n + \mathbf{m}_n^\mathsf{T}\mathbf{m}_n) - \frac{1}{2v}(\boldsymbol{\theta}^\mathsf{T}\boldsymbol{\theta} - 2\boldsymbol{\theta}^\mathsf{T}\mathbf{m} + \mathbf{m}^\mathsf{T}\mathbf{m})\right\} && (26)\\
&= \exp\left\{\frac{1}{2}\left(\frac{1}{v_n} - \frac{1}{v}\right)\boldsymbol{\theta}^\mathsf{T}\boldsymbol{\theta} - \boldsymbol{\theta}^\mathsf{T}\left(\frac{\mathbf{m}_n}{v_n} - \frac{\mathbf{m}}{v}\right) + \frac{1}{2v_n}\mathbf{m}_n^\mathsf{T}\mathbf{m}_n - \frac{1}{2v}\mathbf{m}^\mathsf{T}\mathbf{m}\right\} && (27)\\
&= \exp\left\{-\frac{1}{2}(v^{-1} - v_n^{-1})\boldsymbol{\theta}^\mathsf{T}\boldsymbol{\theta} + \boldsymbol{\theta}^\mathsf{T}(v^{-1}\mathbf{m} - v^{-1}\mathbf{m}_n) + \text{const}\right\} && (28)\\
&\propto \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}^{\backslash n}, v^{\backslash n}\mathbf{I}) && (29)
\end{aligned}
$$

which is a Gaussian distribution with mean and variance given by

$$(v^{\backslash n})^{-1} = v^{-1} - v_n^{-1} \tag{30}$$

$$
\begin{aligned}
(v^{\backslash n})^{-1}\mathbf{m}^{\backslash n} &= v^{-1}\mathbf{m} - v_n^{-1}\mathbf{m}_n && (31)\\
\mathbf{m}^{\backslash n} &= (v^{-1}\mathbf{m} - v_n^{-1}\mathbf{m}_n)v^{\backslash n} && (32)
\end{aligned}
$$

To evaluate Z_n we can use eq. 10.206 in Bishop

$$
\begin{aligned}
Z_n &= \int q^{\backslash n}(\boldsymbol{\theta}) f_n(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} && (33)\\
&= \int \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}^{\backslash n}, v^{\backslash n}\mathbf{I})\Big[(1-w)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\theta}, \mathbf{I}) + w\mathcal{N}(\mathbf{x}_n|\mathbf{0}, a\mathbf{I})\Big]\mathrm{d}\boldsymbol{\theta} && (34)
\end{aligned}
$$

We note that $q^{\backslash n}(\boldsymbol{\theta})$ is a Gaussian distribution over $\boldsymbol{\theta}$ and $f_n(\boldsymbol{\theta})$ is a Gaussian distribution over $\mathbf{x}_n$ conditioned on $\boldsymbol{\theta}$. We can then interpret Z_n in the following way

$$
\begin{aligned}
Z_n &= \int p(\boldsymbol{\theta})p(\mathbf{x}_n|\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} && (35)\\
&= \int p(\boldsymbol{\theta}, \mathbf{x}_n)\mathrm{d}\boldsymbol{\theta} && (36)\\
&= p(\mathbf{x}_n) && (37)
\end{aligned}
$$

so that we can apply the result 2.115 in Bishop to obtain

$$
\begin{aligned}
Z_n &= p(\mathbf{x}_n) && (38)\\
&= (1-w)\mathcal{N}\left(\mathbf{x}_n|\mathbf{m}^{\backslash n}, \mathbf{I} + v^{\backslash n}\mathbf{I}\right) + w\mathcal{N}(\mathbf{x}_n|\mathbf{0}, a\mathbf{I}) && (39)\\
&= (1-w)\mathcal{N}\left(\mathbf{x}_n|\mathbf{m}^{\backslash n}, (1 + v^{\backslash n})\mathbf{I}\right) + w\mathcal{N}(\mathbf{x}_n|\mathbf{0}, a\mathbf{I}) && (40)
\end{aligned}
$$

■