ECE1513
Tutorial 7:
Selected Exercises from Chapter 9

TA: Faye, Pourghasem, fateme.pourghasem@mail.utoronto.ca

October 27, 2023

1 Example 9.3

(⋆) WWW Consider a Gaussian mixture model in which the marginal distribution $p(\mathbf{z})$ for the latent variable is given by (9.10), and the conditional distribution $p(\mathbf{x}|\mathbf{z})$ for the observed variable is given by (9.11). Show that the marginal distribution $p(\mathbf{x})$, obtained by summing $p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$ over all possible values of $\mathbf{z}$, is a Gaussian mixture of the form (9.7).

Solution

From (9.10) and (9.11), we have

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

Exploiting the 1-of-K representation for $\mathbf{z}$, we can re-write the r.h.s. as

$$\sum_{j=1}^{K} \prod_{k=1}^{K} \left(\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)^{I_{kj}} = \sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

where $I_{kj} = 1$ if $k = j$ and 0 otherwise.

2 Example 9.5

(⋆) Consider the directed graph for a Gaussian mixture model shown in Figure 9.6. By making use of the d-separation criterion discussed in Section 8.2, show that the posterior distribution of the latent variables factorizes with respect to the different data points so that

$$p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^{N} p(\mathbf{z}_n|\mathbf{x}_n, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}). \tag{9.80}$$

Solution

Consider any two of the latent variable nodes, which we denote $\mathbf{z}_l$ and $\mathbf{z}_m$. We wish to determine whether these variables are independent, conditioned on the observed data $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and on the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\pi}$. To do this we consider every possible path from $\mathbf{z}_l$ to $\mathbf{z}_m$. The plate denotes that there are N separate copies of the notes $\mathbf{z}_n$ and $\mathbf{x}_n$. Thus the only paths which connect $\mathbf{z}_l$ and $\mathbf{z}_m$ are those which go via one of the parameter nodes $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ or $\boldsymbol{\pi}$. Since we are conditioning on these parameters they represent observed nodes. Furthermore, any path through one of these parameter nodes must be tail-to-tail at the parameter node, and hence all such paths are blocked. Thus $\mathbf{z}_l$ and $\mathbf{z}_m$ are independent, and since this is true for any pair of such nodes it follows that the posterior distribution factorizes over the data set.

3 Example 9.7

(⋆) WWW Verify that maximization of the complete-data log likelihood (9.36) for a Gaussian mixture model leads to the result that the means and covariances of each component are fitted independently to the corresponding group of data points, and the mixing coefficients are given by the fractions of points in each group.

Solution

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left\{ \ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}. \tag{9.36}$$

We begin by calculating the derivative of Eq (9.36) with respect to $\boldsymbol{\mu}_k$:

$$\begin{aligned}
\frac{\partial \ln p}{\partial \boldsymbol{\mu}_k} &= \frac{\partial}{\partial \boldsymbol{\mu}_k} \left\{ \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left[\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right] \right\} \\
&= \frac{\partial}{\partial \boldsymbol{\mu}_k} \left\{ \sum_{n=1}^{N} z_{nk} \left[\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right] \right\} \\
&= \sum_{n=1}^{N} \frac{\partial}{\partial \boldsymbol{\mu}_k} \left\{ z_{nk} \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\} \\
&= \sum_{\mathbf{x}_n \in C_k} \frac{\partial}{\partial \boldsymbol{\mu}_k} \left\{ \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}
\end{aligned}$$

Where we have used $\mathbf{x}_n \in C_k$ to represent the data point $\mathbf{x}_n$ which are assigned to the k-th cluster. Therefore, $\boldsymbol{\mu}_k$ is given by the mean of those $x_n \in C_k$ just as the case of a single Gaussian. It is exactly the same for the covariance. Next, we maximize Eq (9.36) with respect to π_k by enforcing a Lagrange multiplier:

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

We calculate the derivative of L with respect to π_k and set it to 0:

$$\frac{\partial L}{\partial \pi_k} = \sum_{n=1}^{N} \frac{z_{nk}}{\pi_k} + \lambda = 0$$

We multiply both sides by π_k and sum over k making use of the constraint Eq (9.9), yielding $\lambda = -N$. Substituting it back into the expression, we can obtain:

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} z_{nk}$$

Just as required.

4 Example 9.8

(⋆) WWW Show that if we maximize (9.40) with respect to $\boldsymbol{\mu}_k$ while keeping the responsibilities $\gamma(z_{nk})$ fixed, we obtain the closed form solution given by (9.17).

Solution

$$\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_{nk})\left\{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right\}. \quad (9.40)$$

Since $\gamma(z_{nk})$ is fixed, the only dependency of Eq (9.40) on $\boldsymbol{\mu}_k$ occurs in the Gaussian, yielding:

$$\begin{aligned}
\frac{\partial \mathbb{E}_z[\ln p]}{\partial \boldsymbol{\mu}_k} &= \frac{\partial}{\partial \boldsymbol{\mu}_k}\Big\{\sum_{n=1}^{N}\gamma(z_{nk})\ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\Big\} \\
&= \sum_{n=1}^{N}\gamma(z_{nk})\cdot\frac{\partial \ln \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\partial \boldsymbol{\mu}_k} \\
&= \sum_{n=1}^{N}\gamma(z_{nk})\cdot\Big[-\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)\Big]
\end{aligned}$$

Setting the derivative equal to 0, we obtain exactly Eq (9.16), and consequently Eq (9.17) just as required. Note that there is a typo in Eq (9.16), $\boldsymbol{\Sigma}_k$ shoule be $\boldsymbol{\Sigma}_k^{-1}$.

$$0 = -\sum_{n=1}^{N}\underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}}_{\gamma(z_{nk})}\boldsymbol{\Sigma}_k(\mathbf{x}_n - \boldsymbol{\mu}_k) \quad (9.16)$$

where we have made use of the form (2.43) for the Gaussian distribution. Note that the posterior probabilities, or responsibilities, given by (9.13) appear naturally on the right-hand side. Multiplying by $\boldsymbol{\Sigma}_k^{-1}$ (which we assume to be nonsingular) and rearranging we obtain

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n \quad (9.17)$$

5 Example 9.12

(⋆) WWW Consider a mixture distribution of the form

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|k) \tag{9.82}$$

where the elements of $\mathbf{x}$ could be discrete or continuous or a combination of these. Denote the mean and covariance of $p(\mathbf{x}|k)$ by $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, respectively. Show that the mean and covariance of the mixture distribution are given by (9.49) and (9.50).

Solution

First we calculate the mean $\boldsymbol{\mu}_k$:

$$\begin{aligned}
\boldsymbol{\mu}_k &= \int \mathbf{x} p(\mathbf{x}) d\mathbf{x} \\
&= \int \mathbf{x} \sum_{k=1}^{K} \pi_k \, p(\mathbf{x}|k) d\mathbf{x} \\
&= \sum_{k=1}^{K} \pi_k \int \mathbf{x} \, p(\mathbf{x}|k) d\mathbf{x} \\
&= \sum_{k=1}^{K} \pi_k \, \boldsymbol{\mu}_k
\end{aligned}$$

Then we deal with the covariance matrix. For an arbitrary random variable $\mathbf{x}$, according to Eq (2.63) we have:

$$\begin{aligned}
\mathrm{cov}[\mathbf{x}] &= \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T] \\
&= \mathbb{E}[\mathbf{x}\mathbf{x}^T] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{x}]^T
\end{aligned}$$

Since $\mathbb{E}[\mathbf{x}]$ is already obtained, we only need to solve $\mathbb{E}[\mathbf{x}\mathbf{x}^T]$. First we only focus on the k-th component and rearrange the expression above, yielding:

$$\mathbb{E}_k[\mathbf{x}\mathbf{x}^T] = \mathrm{cov}_k[\mathbf{x}] + \mathbb{E}_k[\mathbf{x}]\mathbb{E}_k[\mathbf{x}]^T = \boldsymbol{\Sigma}_k + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T$$

We further use Eq (2.62), yielding:

$$\begin{aligned}
\mathbb{E}[\mathbf{x}\mathbf{x}^T] &= \int \mathbf{x}\mathbf{x}^T \sum_{k=1}^{K} \pi_k \, p(\mathbf{x}|k) \, d\mathbf{x} \\
&= \sum_{k=1}^{K} \pi_k \int \mathbf{x}\mathbf{x}^T \, p(\mathbf{x}|k) \, d\mathbf{x} \\
&= \sum_{k=1}^{K} \pi_k \, \mathbb{E}_k[\mathbf{x}\mathbf{x}^T] \\
&= \sum_{k=1}^{K} \pi_k \, (\boldsymbol{\mu}_k \boldsymbol{\mu}_k^T + \boldsymbol{\Sigma}_k)
\end{aligned}$$

Therefore, we obtain Eq (9.50) just as required.

6 Example 9.19

($\star\star$) Consider a D-dimensional variable $\mathbf{x}$ each of whose components i is itself a multinomial variable of degree M so that $\mathbf{x}$ is a binary vector with components x_{ij} where $i = 1, \ldots, D$ and $j = 1, \ldots, M$, subject to the constraint that $\sum_j x_{ij} = 1$ for all i. Suppose that the distribution of these variables is described by a mixture of the discrete multinomial distributions considered in Section 2.2 so that

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k p(\mathbf{x}|\boldsymbol{\mu}_k) \tag{9.84}$$

where

$$p(\mathbf{x}|\boldsymbol{\mu}_k) = \prod_{i=1}^{D} \prod_{j=1}^{M} \mu_{kij}^{x_{ij}}. \tag{9.85}$$

The parameters μ_{kij} represent the probabilities $p(x_{ij} = 1|\boldsymbol{\mu}_k)$ and must satisfy $0 \leqslant \mu_{kij} \leqslant 1$ together with the constraint $\sum_j \mu_{kij} = 1$ for all values of k and i. Given an observed data set $\{\mathbf{x}_n\}$, where $n = 1, \ldots, N$, derive the E and M step equations of the EM algorithm for optimizing the mixing coefficients π_k and the component parameters μ_{kij} of this distribution by maximum likelihood.

Solution

As usual we introduce a latent variable $\mathbf{z}_n$ corresponding to each observation. The conditional distribution of the observed data set, given the latent variables, is then

$$p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}) = \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\mu}_k)^{z_{nk}}.$$

Similarly, the distribution of the latent variables is given by

$$p(\mathbf{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N} \pi_k^{z_{nk}}.$$

The expected value of the complete-data log likelihood function is given by

$$\sum_{n=1}^{N}\sum_{k=1}^{k} \gamma(z_{nk}) \left\{ \ln \pi_k + \sum_{i=1}^{D}\sum_{j=1}^{M} x_{nij} \ln \mu_{kij} \right\}$$

where as usual we have defined responsibilities given by

$$\gamma(z_{nk}) = \mathbb{E}[z_{nk}] = \frac{\pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k)}{\sum_{j=1}^{M} \pi_j p(\mathbf{x}_n|\boldsymbol{\mu}_j)}.$$

These represent the E-step equations.

To derive the M-step equations we add to the expected complete-data log likelihood function a set of Lagrange multiplier terms given by

$$\lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) + \sum_{k=1}^{K}\sum_{i=1}^{D} \eta_{ki} \left(\sum_{j=1}^{M} \mu_{kij} - 1 \right)$$

to enforce the constraint $\sum_k \pi_k = 1$ as well as the set of constraints

$$\sum_{j=1}^{M} \mu_{kij} = 1$$

for all values of i and k. Maximizing with respect to the mixing coefficients π_k, and eliminating the Lagrange multiplier λ in the usual way, we obtain

$$\pi_k = \frac{N_k}{N}$$

where we have defined

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

Similarly maximizing with respect to the parameters μ_{kij}, and again eliminating the Lagrange multipliers, we obtain

$$\mu_{kij} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_{nij}.$$

This is an intuitively reasonable result which says that the value of μ_{kij} for component k is given by the fraction of those counts assigned to component k which have non-zero values of the corresponding elements i and j.

7 Example 9.20

(⋆) WWW Show that maximization of the expected complete-data log likelihood function (9.62) for the Bayesian linear regression model leads to the M step re-estimation result (9.63) for α.

Solution

$$\mathbb{E}\left[\ln p(\mathbf{t}, \mathbf{w}|\alpha, \beta)\right] = \frac{M}{2}\ln\left(\frac{\alpha}{2\pi}\right) - \frac{\alpha}{2}\mathbb{E}\left[\mathbf{w}^{\mathrm{T}}\mathbf{w}\right] + \frac{N}{2}\ln\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2}\sum_{n=1}^{N}\mathbb{E}\left[(t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n)^2\right]. \tag{9.62}$$

We first calculate the derivative of Eq (9.62) with respect to α and set it to 0:

$$\frac{\partial E[\ln p]}{\partial \alpha} = \frac{M}{2}\frac{1}{2\pi}\frac{2\pi}{\alpha} - \frac{\mathbb{E}[\mathbf{w}^T\mathbf{w}]}{2} = 0$$

We rearrange the equation above, which gives:

$$\alpha = \frac{M}{\mathbb{E}[\mathbf{w}^T\mathbf{w}]} \tag{$*$}$$

Therefore, we now need to calculate the expectation $\mathbb{E}[\mathbf{w}^T\mathbf{w}]$. Notice that the posterior has already been given by Eq (3.49):

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{m}_N, \mathbf{S}_N)$$

To calculate $\mathbb{E}[\mathbf{w}^T\mathbf{w}]$, here we write down an property for a Gaussian random variable: if $\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma})$, we have:

$$\mathbb{E}[\mathbf{x}^T\mathbf{A}\mathbf{x}] = \mathrm{Tr}[\mathbf{A}\boldsymbol{\Sigma}] + \mathbf{m}^T\mathbf{A}\mathbf{m}$$

This property has been shown in Eq(378) in 'the Matrix Cookbook'. Utilizing this property, we can obtain:

$$\mathbb{E}[\mathbf{w}^T\mathbf{w}] = \mathrm{Tr}[\mathbf{S}_N] + \mathbf{m}_N^T\mathbf{m}_N$$

Substituting it back into $(*)$, we obtain what is required.

8 Example 9.25

This follows from the fact that the Kullback-Leibler divergence, $\mathrm{KL}(q\|p)$, is at its minimum, 0, when q and p are identical. This means that

$$\frac{\partial}{\partial \boldsymbol{\theta}}\mathrm{KL}(q\|p) = \mathbf{0},$$

since $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$. Therefore, if we compute the gradient of both sides of (9.70) w.r.t. $\boldsymbol{\theta}$, the contribution from the second term on the r.h.s. will be $\mathbf{0}$, and so the gradient of the first term must equal that of the l.h.s.

Solution

This follows from the fact that the Kullback-Leibler divergence, $\mathrm{KL}(q\|p)$, is at its minimum, 0, when q and p are identical. This means that

$$\frac{\partial}{\partial \boldsymbol{\theta}}\mathrm{KL}(q\|p) = \mathbf{0},$$

since $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$ depends on $\boldsymbol{\theta}$. Therefore, if we compute the gradient of both sides of (9.70) w.r.t. $\boldsymbol{\theta}$, the contribution from the second term on the r.h.s. will be $\mathbf{0}$, and so the gradient of the first term must equal that of the l.h.s.

9 Example 9.26

($\star$) WWW Consider the incremental form of the EM algorithm for a mixture of Gaussians, in which the responsibilities are recomputed only for a specific data point $\mathbf{x}_m$. Starting from the M-step formulae (9.17) and (9.18), derive the results (9.78) and (9.79) for updating the component means.

Solution

From Eq (9.18), we have:

$$N_k^{\text{old}} = \sum_n \gamma^{\text{old}}(z_{nk})$$

If now we just re-evaluate the responsibilities for one data point $\mathbf{x}_m$, we can obtain:

$$\begin{aligned} N_k^{\text{new}} &= \sum_{n \neq m} \gamma^{\text{old}}(z_{nk}) + \gamma^{\text{new}}(z_{mk}) \\ &= \sum_n \gamma^{\text{old}}(z_{nk}) + \gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk}) \\ &= N_k^{\text{old}} + \gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk}) \end{aligned}$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \tag{9.17}$$

Similarly, according to Eq (9.17), we can obtain:

$$
\begin{aligned}
\boldsymbol{\mu}_k^{\text{new}} &= \frac{1}{N_k^{\text{new}}} \sum_{n \neq m} \gamma^{\text{old}}(z_{nk}) \mathbf{x}_n + \frac{\gamma^{\text{new}}(z_{mk}) \mathbf{x}_m}{N_k^{\text{new}}} \\
&= \frac{1}{N_k^{\text{new}}} \sum_{n} \gamma^{\text{old}}(z_{nk}) \mathbf{x}_n + \frac{\gamma^{\text{new}}(z_{mk}) \mathbf{x}_m}{N_k^{\text{new}}} - \frac{\gamma^{\text{old}}(z_{mk}) \mathbf{x}_m}{N_k^{\text{new}}} \\
&= \frac{N_k^{\text{old}}}{N_k^{\text{new}}} \frac{1}{N_k^{\text{old}}} \sum_{n} \gamma^{\text{old}}(z_{nk}) \mathbf{x}_n + \frac{\gamma^{\text{new}}(z_{mk}) \mathbf{x}_m}{N_k^{\text{new}}} - \frac{\gamma^{\text{old}}(z_{mk}) \mathbf{x}_m}{N_k^{\text{new}}} \\
&= \frac{N_k^{\text{old}}}{N_k^{\text{new}}} \boldsymbol{\mu}_k^{\text{old}} + \left[\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk}) \right] \frac{\mathbf{x}_m}{N_k^{\text{new}}} \\
&= \boldsymbol{\mu}_k^{\text{old}} - \frac{N_k^{\text{new}} - N_k^{\text{old}}}{N_k^{\text{new}}} \boldsymbol{\mu}_k^{\text{old}} + \left[\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk}) \right] \frac{\mathbf{x}_m}{N_k^{\text{new}}} \\
&= \boldsymbol{\mu}_k^{\text{old}} - \frac{\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}} \boldsymbol{\mu}_k^{\text{old}} + \left[\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk}) \right] \frac{\mathbf{x}_m}{N_k^{\text{new}}} \\
&= \boldsymbol{\mu}_k^{\text{old}} + \frac{\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}} \cdot \left(\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}} \right)
\end{aligned}
$$

Just as required.