Statistical Machine Learning 2018

Exercises and answers, week 12

7 December 2018

TUTORIAL

Exercise 1

A Gaussian mixture model (Bishop, §9.2) can be written in terms of discrete *latent* variables. The latent variable $\mathbf{z}$ determines/indicates from which Gaussian the observed variable $\mathbf{x}$ is sampled.
Assume that the latent variable $\mathbf{z}$ can take on K different values, each corresponding to a different Gaussian distribution for $\mathbf{x}$.

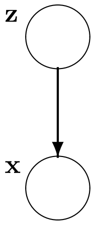

Figure 1: mixture model

Let π_k denote the probability that $\mathbf{z}$ takes on value k. Using the familiar '1-of-K' representation, with $\mathbf{z}$ a K-dimensional binary vector containing a single 1 and zeros for the rest, using shorthand notation $\mathbf{z}[k] \equiv z_k$ for the k-th element of $\mathbf{z}$, this can be written as

$$p(z_k = 1) = \pi_k \tag{1}$$

1. Verify that this latent variable model corresponds to the Gaussian mixture distribution

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{2}$$

ANSWER: Technically, you should find an explicit expression for the joint distribution $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$ and then marginalizing this over all possible states of $\mathbf{z}$ to obtain the expression for $p(\mathbf{x})$.
Doing so, using the '1-of-K' representation introduced with the multinomial distribution (eq.2.26), we can write

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \tag{3}$$

Likewise for the conditional distribution $p(\mathbf{x}|\mathbf{z})$ we can write

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k} \tag{4}$$

Putting the two together and summing over $\mathbf{z}$ then gives

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}|\mathbf{z})p(\mathbf{z}) = \sum_{\mathbf{z}} \prod_{k=1}^{K} (\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))^{z_k} \tag{5}$$

The k-th element of $\mathbf{z}$ ensures that for each $\mathbf{z}$ only a single (Gaussian) component contributes to the sum, resulting in (2).

Suppose, for a certain application, we have such a mixture model, consisting of two, a priori equally likely, Gaussian components, with means $\boldsymbol{\mu}_1 = (1, \frac{1}{2})$ and $\boldsymbol{\mu}_2 = (2, 1)$, and with both components having identical isotropic covariance matrices with unit variance $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}_2$. We also have a data set of no less than four data points.
This situation is depicted in Figure 2.

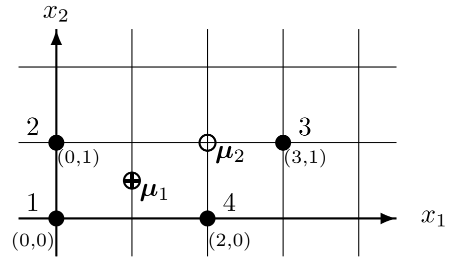

Figure 2: four data points, two classes ...

The *responsibility* γ_{nk} that component k takes for explaining an observation $\mathbf{x}_n$ is defined as

$$\gamma_{nk} \equiv p(z_k = 1|\mathbf{x}_n) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{6}$$

2. Calculate the responsibilities γ_{11} and γ_{12} for point $1 = (0, 0)$. Verify they add to 1. Does this hold in general?

   ANSWER: Filling in from the probability distributions for both components we have

$$\begin{aligned} \mathcal{N}_1(\mathbf{x} = (0,0)|\boldsymbol{\mu}_1, \mathbf{I}_2) &= (2\pi)^{-1} \exp(-5/8) \approx 0.08519 \\ \mathcal{N}_2(\mathbf{x} = (0,0)|\boldsymbol{\mu}_2, \mathbf{I}_2) &= (2\pi)^{-1} \exp(-5/2) \approx 0.01306 \end{aligned}$$

   With $\pi_1 = \pi_2 = 0.5$ the priors cancel each other, and so

$$\begin{aligned} \gamma_{11} &= \frac{0.08519}{(0.08519 + 0.01306)} \approx 0.867 \\ \gamma_{12} &= \frac{0.01306}{(0.08519 + 0.01306)} \approx 0.133 \end{aligned}$$

   Together they add to one, which should be obvious from the definition (6).

   Notice that Bishop uses the slightly obscure notation $\gamma(z_{nk})$ instead of γ_{nk} to denote the responsibility component k takes in 'explaining' datapoint x_n.

Exercise 2

EM-algorithm for Gaussian mixtures (Bishop, §9.2.2). Setting the derivatives of the log-likelihood function $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with respect to the means $\boldsymbol{\mu}_k$ of the Gaussian components to zero, it can be shown that in a maximum the means must satisfy

$$\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0 \tag{7}$$

1. Show that from this it follows that in a maximum

$$\begin{aligned} \boldsymbol{\mu}_k &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \mathbf{x}_n & (8) \\ N_k &= \sum_{n=1}^{N} \gamma_{nk} & (9) \end{aligned}$$

ANSWER: Recognising that the first term in (7) is simply the responsibility (6) of component k for explaining data point $\mathbf{x}_n$

$$\frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \equiv \gamma_{nk}$$

we can multiply through by $\boldsymbol{\Sigma}_k$ to get

$$\sum_{n=1}^{N} \gamma_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0 \tag{10}$$

Rearranging terms then gives the expressions (8) and (9).

Similarly, in a maximum we find for the covariance matrices $\boldsymbol{\Sigma}_k$ and mixing coefficients π_k

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k}\sum_{n=1}^{N} \gamma_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \tag{11}$$

$$\pi_k = \frac{N_k}{N} \tag{12}$$

Consider again the data points in Figure 1. Assume that at a given stage in the EM algorithm the responsibilites γ_{nk} are given as $\gamma_{11} = \gamma_{21} = \gamma_{32} = \gamma_{42} = 0.9$. (In words: points 1 and 2 mostly belong to the first component, points 3 and 4 mostly to the second).

2. Describe the E and M steps of the algorithm. Compute one full cycle of the EM-algorithm for this situation.

ANSWER: Responsibilities are given, so now first the M-step, and then an E-step ...
First write

$$\mathbf{X} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \\ 3 & 1 \\ 2 & 0 \end{bmatrix} \qquad \gamma_{nk} = \begin{bmatrix} 0.9 & 0.1 \\ 0.9 & 0.1 \\ 0.1 & 0.9 \\ 0.1 & 0.9 \end{bmatrix}$$

then from (9) we have $N_1 = N_2 = 2$, from which we compute $\pi_1 = \pi_2 = 0.5$ as well as $\boldsymbol{\mu} = \frac{1}{2}\gamma_{nk}^T\mathbf{X}$, resulting in

$$\begin{aligned} \boldsymbol{\mu}_1 &= [0.25, 0.5]^T \\ \boldsymbol{\mu}_2 &= [2.25, 0.5]^T \end{aligned}$$

Using this in (11) we can calculate the covariance matrices $\boldsymbol{\Sigma}_k$ to be

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.5875 & 0.025 \\ 0.025 & 0.25 \end{bmatrix} \qquad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.7875 & 0.225 \\ 0.225 & 0.25 \end{bmatrix}$$

which completes the M-step. Recalculating the responsibilities from (6) we find

$$\gamma_{nk} = \begin{bmatrix} 0.9391 & 0.0609 \\ 0.9979 & 0.0021 \\ 0.0021 & 0.9979 \\ 0.0609 & 0.9391 \end{bmatrix}$$

which completes the E-step. A recalculation of the $\boldsymbol{\mu}$ shows that now

$$\begin{aligned} \boldsymbol{\mu}_1 &= [0.0641, 0.5]^T \\ \boldsymbol{\mu}_2 &= [2.4359, 0.5]^T \end{aligned}$$

In other words: the cluster centers have moved significantly towards the centers of the clusters {1,2} and {3,4}.

3. What will probably be the final outcome after convergence for this situation (sketch)?

   ANSWER: This will result in another type of degenerate solution (not infinite), now for the covariance matrix. The means of the two clusters will fall exactly halfway between points 1 and 2 and between points 3 and 4. The covariance, however, will become 1-dimensional: it will shrink to zero in the direction perpendicular to the line through {1,2} and the line through {3,4}. So

$$\begin{aligned} \boldsymbol{\mu}_1 &= [0.0, 0.5]^T \\ \boldsymbol{\mu}_2 &= [2.5, 0.5]^T \end{aligned}$$

   with covariances

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 0 & 0 \\ 0 & 0.125 \end{bmatrix} \qquad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.125 & 0.125 \\ 0.125 & 0.125 \end{bmatrix}$$

   (see Matlab).

Exercise 3

It should come as no surprise that the EM algorithm does not only apply to Gaussians, but to many other mixtures of distributions as well. As an example we will now look at a Bernoulli mixture model and the way it can be applied to handwritten digit recognition.
Consider a set of 800 digital images of handwritten examples of the numbers ’2’, ’3’ and ’4’, see Figure 9.10. Each image consists of 20x20 binary pixels that are each either black (pixelvalue = 1) or white (pixelvalue = 0). The labels belonging to the images are unknown.
We model the handwritten digits with a Bernoulli mixture model:

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \prod_{i=1}^{D} \mu_{ki}^{x_i} (1 - \mu_{ki})^{(1-x_i)} \tag{13}$$

where $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$ and $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_K\}$.

1. Think of how you would model the handwritten digit images. Verify the Bernoulli mixture model (13) also does the job: interpret the parameters and put numbers to the constants.

   ANSWER: In this model numbers ’2’, ’3’ and ’4’ represent the three classes or components k in the mixture, so $K = 3$. The parameter vector $\boldsymbol{\pi}$ represents the probability on each class. The index i runs over all the pixels in the image, so $D = 400$. The x_i represent the value of pixel i (zero or one), whereas the μ_{ki} correspond to the probability that pixel i is equal to one in class k. The vector $\mathbf{x}$ corresponds to a single data point (image) with length 400. In total there are 800 data points, so $N = 800$ in subsequent eqs. below.

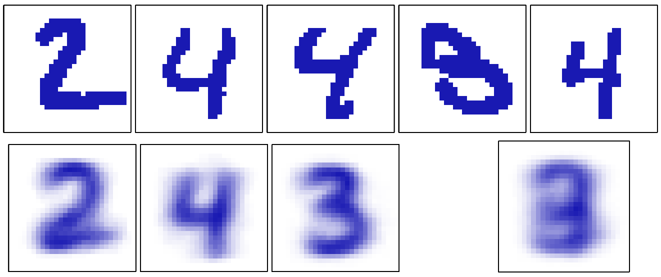

Figure 9.10 - Bernoulli mixture model for handwritten digit recognition.

2. Compute the complete-data log likelihood function $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\pi})$.
Hint: write the likelihood (13) in terms of a latent variable $\mathbf{z}$ using a 1-of-K coding scheme.

   ANSWER: Introducing an explicit latent variable $\mathbf{z}$ to represent the component of the mixture we decompose (13) into

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{\mathbf{z}} p(\mathbf{x}|\mathbf{z}, \boldsymbol{\mu}) p(\mathbf{z}|\boldsymbol{\pi})$$

   with $p(\mathbf{z}|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_k}$ and $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\mu}) = \prod_{k=1}^{K} p(\mathbf{x}|\boldsymbol{\mu}_k)^{z_k}$. For a data point $\{\mathbf{x}, \mathbf{z}\}$, the likelihood becomes

$$\begin{aligned} p(\mathbf{x}, \mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\pi}) &= p(\mathbf{x}|\mathbf{z}, \boldsymbol{\mu}) p(\mathbf{z}|\boldsymbol{\pi}) \\ &= \prod_{k=1}^{K} [\pi_k p(\mathbf{x}|\boldsymbol{\mu}_k)]^{z_k} \\ &= \prod_{k=1}^{K} [\pi_k \prod_{i=1}^{D} \mu_{ki}^{x_i} (1-\mu_{ki})^{(1-x_i)}]^{z_k} \end{aligned}$$

   Multiplying for all N points in the data set and taking the log then gives

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left\{ \ln \pi_k + \sum_{i=1}^{D} [x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1 - \mu_{ki})] \right\} \quad (14)$$

3. Show that the expectation of the complete-data log likelihood w.r.t. the posterior distribution of $\mathbf{Z}$ is given by

$$\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\pi})] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \left\{ \ln \pi_k + \sum_{i=1}^{D} [x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1 - \mu_{ki})] \right\} \quad (15)$$

   ANSWER: Using Bayes' theorem we can use expressions for $p(\mathbf{x}|\mathbf{z})$ and $p(\mathbf{z})$ to obtain the posterior distribution of the latent variables in the form

$$p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\pi}) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} [\pi_k p(\mathbf{x}|\boldsymbol{\mu}_k)]^{z_{nk}} \quad (16)$$

   Now, from all the terms on the r.h.s. of (14), only the indicator variable z_{nk} actually depends on $\mathbf{z}$. Therefore, in computing the expectation w.r.t. this posterior distribution we only need to consider $\mathbb{E}[z_{nk}]$. For this we find

$$\begin{aligned} \mathbb{E}[z_{nk}] &= \frac{\sum_{z_n} z_{nk} \prod_{k'} [\pi_{k'} p(\mathbf{x}_n|\boldsymbol{\mu}_{k'})]^{z_{nk'}}}{\sum_{z_n} \prod_j [\pi_j p(\mathbf{x}_n|\boldsymbol{\mu}_j)]^{z_{nj}}} \\ &= \frac{\pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k)}{\sum_{j=1}^{K} \pi_j p(\mathbf{x}_n|\boldsymbol{\mu}_j)} \\ &= \gamma_{nk} \end{aligned}$$

   the familiar expression for the responsibility that component k takes on for data point $\mathbf{x}_n$. Substituting this result in (14) then gives the desired result.

4. In the M-step of the EM algorithm the expected complete-data log likelihood is maximized w.r.t. the parameters. Find an expression for the value of $\boldsymbol{\mu}_k$ that maximizes eq.(15).

ANSWER: Taking the derivative of (15) w.r.t. component μ_{ki} gives

$$\begin{aligned}
\frac{\partial}{\partial \mu_{ki}} \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\pi})] &= \sum_{n=1}^{N} \gamma_{nk} \left(\frac{x_{ni}}{\mu_{ki}} - \frac{(1 - x_{ni})}{(1 - \mu_{ki})} \right) \\
&= \frac{\sum_n \gamma_{nk} x_{ni} - \sum_n \gamma_{nk} \mu_{ki}}{\mu_{ki}(1 - \mu_{ki})}
\end{aligned}$$

Setting equal to zero results in

$$\mu_{ki} = \frac{\sum_n \gamma_{nk} x_{ni}}{\sum_n \gamma_{nk}} \equiv \bar{x}_{ki} \tag{17}$$

Combining the components into a single vector then gives $\boldsymbol{\mu}_k = \bar{\mathbf{x}}_k$ (eq.9.59).