Statistical Machine Learning 2018

Exercises, 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 $\mathbf{x}$.

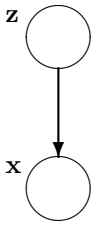

Figure 1: mixture model

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

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

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.

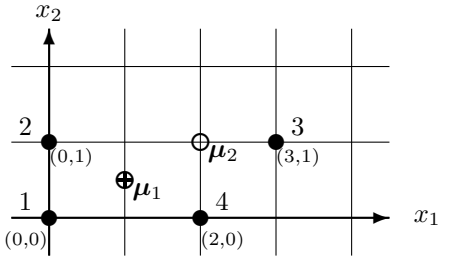

Figure 2: four data points, two classes ...

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

$$\gamma_{nk} \equiv p(z_k = 1|\mathbf{x}_n) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{3}$$

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

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

$$\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0 \tag{4}$$

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

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} \mathbf{x}_n \tag{5}$$

$$N_k = \sum_{n=1}^{N} \gamma_{nk} \tag{6}$$

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

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \tag{7}$$

$$\pi_k = \frac{N_k}{N} \tag{8}$$

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.

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

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:

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \prod_{i=1}^{D} \mu_{ki}^{x_i} (1 - \mu_{ki})^{(1-x_i)} \tag{9}$$

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 (9) also does the job: interpret the parameters and put numbers to the constants.

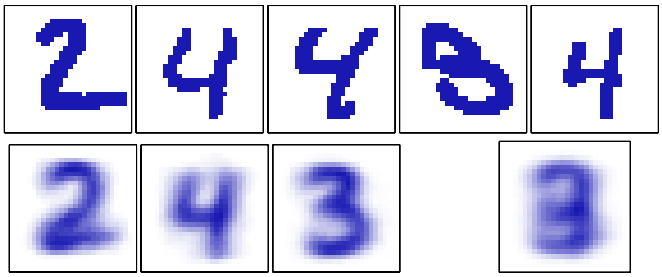

Figure 9.10 - Bernoulli mixture model for handwritten digit recognition.

2. Compute the complete-data log likelihood function $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\pi})$.
   Hint: write the likelihood (9) in terms of a latent variable $\mathbf{z}$ using a 1-of-K coding scheme.

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

$$\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\pi})] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \left\{ \ln \pi_k + \sum_{i=1}^{D} [x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1 - \mu_{ki})] \right\} \quad (10)$$

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.(10).