UML2012
Exercise set 6
Solutions to be presented in the 27.4.2012 session

Exercise 1:

First some theory: The EM-algorithm is a method to estimate parameters θ when you can only observe a subgroup $\mathbf{x}$ of all the variables $(\mathbf{x}, \mathbf{s})$ in the model. In other words, given the statistical model $p(\mathbf{x}, \mathbf{s}, \theta)$ you want to estimate θ from the iid observations $\mathbf{x}(1) \dots \mathbf{x}(T)$. Note that, however, you do not have observations $\mathbf{s}(1) \dots \mathbf{s}(T)$ available. The s_i in the vector $\mathbf{s}$ are called latent variables.

One solution for such a situation would be to integrate out the latent variables to obtain $p(\mathbf{x}, \theta)$. Then, to maximize the likelihood $\ell(\theta)$,

$$\ell(\theta) = \sum_t \log p(\mathbf{x}(t), \theta) \tag{1}$$

in order to find θ. The EM-algorithm offers an alternative solution.

If you were able to observe both $\mathbf{x}$ and $\mathbf{s}$, your log-likelihood would be $\ell_{xs}(\theta) = \sum_t \log p(\mathbf{x}(t), \mathbf{s}(t), \theta)$ which you would maximize to find θ. We call this likelihood the *full* likelihood because it is based both on $\mathbf{x}(t)$ and the latent variables $\mathbf{s}(t)$. However, we do not know the $\mathbf{s}(t)$, and thus the $\log p(\mathbf{x}(t), \mathbf{s}(t), \theta)$. The idea in the EM-Algorithm is to replace $\log p(\mathbf{x}(t), \mathbf{s}(t), \theta)$ with an estimate, and to maximize then the estimated $\ell_{xs}(\theta)$ with respect to θ.

Assume that you have an initial estimate of θ available, call it θ_0. Then, you can calculate the posterior $p(\mathbf{s}|\mathbf{x}, \theta_0)$

$$p(\mathbf{s}|\mathbf{x}, \theta_0) = \frac{p(\mathbf{x}, \mathbf{s}, \theta_0)}{p(\mathbf{x}, \theta_0)}, \tag{2}$$

and replace $\log p(\mathbf{x}(t), \mathbf{s}(t), \theta)$ by its expected value (expected with respect to $\mathbf{s}(t)$) given θ_0 and the data $\mathbf{x}(t)$, i.e. by

$$\int \left[\log p(\mathbf{x}(t), \mathbf{s}(t), \theta)\right] p(\mathbf{s}(t)|\mathbf{x}(t), \theta_0) d\mathbf{s}(t). \tag{3}$$

The estimated full log-likelihood becomes then

$$J(\theta|\theta_0) = \sum_t J_t(\theta|\theta_0) \tag{4}$$

with

$$J_t(\theta|\theta_0) = \int \left[\log p(\mathbf{x}(t), \mathbf{s}(t), \theta)\right] p(\mathbf{s}(t)|\mathbf{x}(t), \theta_0) d\mathbf{s}(t), \tag{5}$$

which is a function of θ. This is called the E-step in the EM-algorithm. Maximization of $J(\theta|\theta_0)$, which is called the M-step, yields then a new estimate θ_1 for θ. With the new estimate, a new posterior $p(\mathbf{s}|\mathbf{x}, \theta_1)$, and a new estimated log-likelihood $J(\theta|\theta_1)$ is calculated. From here, you obtain the next estimate θ_2 for θ.

The goal of this exercise is to show that the EM-iteration described above leads to estimates θ_k which increase the likelihood $\ell(\theta_k)$ in each iteration. (Recall that $\ell(\theta)$ was the likelihood that would be obtained by integrating out the latent variables $\mathbf{s}$.)

Now to the questions:

1.1 Show that maximization of $J(\theta|\theta_k)$ is the same as maximization of $\tilde{J}(\theta, \theta_k) = \sum_t \tilde{J}_t(\theta|\theta_k)$, where

$$\tilde{J}_t(\theta|\theta_k) = \int \log\left(\frac{p(\mathbf{x}(t), \mathbf{s}(t), \theta)}{p(\mathbf{x}(t), \mathbf{s}(t), \theta_k)}\right) p(\mathbf{s}(t)|\mathbf{x}(t), \theta_k) d\mathbf{s}(t). \tag{6}$$

1.2 Explain why for the EM-algorithm it holds that $\tilde{J}(\theta_{k+1}|\theta_k) \geq 0$.

1.3 Use the fact

$$p(\mathbf{x}, \mathbf{s}, \theta) = p(\mathbf{s}|\mathbf{x}, \theta) p(\mathbf{x}, \theta) \tag{7}$$

to show that

$$\begin{aligned}\tilde{J}(\theta_{k+1}|\theta_k) = \ell(\theta_{k+1}) - \ell(\theta_k) + \\ \sum_t \int \log\left(\frac{p(\mathbf{s}(t)|\mathbf{x}(t), \theta_{k+1})}{p(\mathbf{s}(t)|\mathbf{x}(t), \theta_k)}\right) p(\mathbf{s}(t)|\mathbf{x}(t), \theta_k) d\mathbf{s}(t).\end{aligned} \tag{8}$$

1.4 For this question you can assume it known that

$$D(f, g) = \int \log\left(\frac{f(x)}{g(x)}\right) f(x) dx \tag{9}$$

is ≥ 0 for all functions f, g and $D(f, f) = 0$. $D(f, g)$ is called the Kullback-Leibler (KL) distance between f and g. Using this property of the KL distance, show that $\ell(\theta_{k+1}) \geq \ell(\theta_k)$.

Exercise 2:

We need to calculate the integral in Eq. (11.20) for the E-step of the EM algorithm. The equation (11.20) from the lecture notes is roughly

$$J(\theta|\theta_{k-1}) = \int \log\big(p(X, S; \theta)\big) p(S|X, \theta_{k-1}) dS \tag{10}$$

This form is the general case where the data might not be iid. Notice that the integral is $N_S T$ dimensional integral, where N_S is the dimension of the vector space

where vectors $\mathbf{s}_t$ are, and T is the number of sample points. Here, we derive the expression for the simpler case of iid data that was used in Exercise 1.

2.1 With the notation of Eq. (11.20), assume that

$$p(X, S; \theta) = \prod_t p(\mathbf{x}_t|\mathbf{s}_t; \theta)p(\mathbf{s}_t; \theta) \tag{11}$$

where the T observations are $X = (\mathbf{x}_1 \dots \mathbf{x}_T)$ and the latent variables are $S = (\mathbf{s}_1 \dots \mathbf{s}_T)$. This is the iid assumption. Begin with the definition $p(X; \theta) = \int p(X, S; \theta)dS$, and prove $p(X; \theta) = \prod_t p(\mathbf{x}_t; \theta)$.

2.2 Show that Eq. (11.20) becomes

$$J(\theta|\theta_{k-1}) = \sum_{t=1}^{T} E_t\big(\log p(\mathbf{x}_t, \mathbf{s}_t; \theta)\big) \tag{12}$$

where the expectation E_t is taken with respect to the posterior $p(\mathbf{s}_t|\mathbf{x}_t; \theta_{k-1})$. For continuous data J is thus

$$J(\theta) = \sum_{t=1}^{T} \int \log p(\mathbf{x}_t, \mathbf{s}_t; \theta)p(\mathbf{s}_t|\mathbf{x}_t; \theta_{k-1})d\mathbf{s}_t \tag{13}$$

while for discrete data it is

$$J(\theta) = \sum_{t=1}^{T} \sum_{\mathbf{s}_t} \log p(\mathbf{x}_t, \mathbf{s}_t; \theta)p(\mathbf{s}_t|\mathbf{x}_t; \theta_{k-1}). \tag{14}$$

2.3 Explain why in the lecture notes Eq. (11.16) the numbers $q^*_{t,c}$ are the posteriors $p(\mathbf{s}_t|\mathbf{x}_t; \theta_{k-1})$. Advice: Denote $\mathbf{s}_t = r(t)$, and $\theta = (\mu_c, C_c, \pi_c)_{c=1,2,\dots,k}$.

Exercise 3:

In this exercise, we first set up a statistical model to do clustering and use then the EM-algorithm to estimate the parameters in the model. This material is be treated in Section 11.7.

The data generation process is as follows: For $t = 1$ till T,

(a) Choose randomly a cluster $r(t) \in \{1 \dots C\}$. Here, we assume that the probability for choosing cluster c is $P(r(t) = c) = \pi_c$.

(b) Draw the t-th sample $\mathbf{x}(t) \in \mathbb{R}^n$ from a multivariate normal distribution with mean $\mu_{r(t)}$ and covariance matrix $C_{r(t)}$.

This process generates thus data $(\mathbf{x}(1) \dots \mathbf{x}(T))$, and $(r(1) \dots r(T))$ but only the $\mathbf{x}(t)$ are observed. The latent variable in this statistical model is r. The parameters are π_c, μ_c, and C_c for $c = 1 \dots C$. In what follows, they are together denoted as θ.

3.1 Given iid data $(\mathbf{x}(1)\dots\mathbf{x}(T))$, and $(r(1)\dots r(T))$ set up the (full) log-likelihood $\ell_{xr}(\theta)$, where θ stands for the parameters of the model (compare with Exercise 1 for the idea of the full log-likelihood).

3.2 Show that the posterior $P(r(t)=c|\mathbf{x}(t),\theta)$ is given by

$$P(r(t)=c|\mathbf{x},\theta)=\frac{\frac{\pi_c}{\sqrt{|C_c|}}\exp\left(-0.5\sum_t(\mathbf{x}(t)-\mu_c)^TC_c^{-1}(\mathbf{x}(t)-\mu_c)\right)}{\sum_{k=1}^C\frac{\pi_k}{\sqrt{|C_k|}}\exp\left(-0.5\sum_t(\mathbf{x}(t)-\mu_k)^TC_k^{-1}(\mathbf{x}(t)-\mu_k)\right)} \tag{15}$$

Note that $P(r(t)=c|\mathbf{x}(t),\theta)$ corresponds to $p(\mathbf{s}(t)|\mathbf{x}(t),\theta)$ in Exercise 1. The difference is, however, that r is a discrete random variable while $\mathbf{s}$ is a (collection of) continuous random variable(s). Therefore, what was integration over $\mathbf{s}$ becomes here summation over the values of r.

3.3 Given an estimate θ_k, what is the estimated full log-likelihood $J(\theta|\theta_k)$? (See Equation (4) of Ex. 1)

3.4 Calculate the gradients $\nabla_{\mu_c}J(\theta|\theta_k)$ and $\nabla_{C_c}J(\theta|\theta_k)$. To get the gradients, you may find it useful to check exercise 13 (Maximum Likelihood Estimation for Multivariate Gaussians) (same as Set 2 Ex 3 2012). Set the gradients to zero, to obtain the following EM-update rules for μ_c and C_c, $c=1\dots C$:

$$\mu_c(k+1)=\frac{\sum_{t=1}^T P(r(t)=c|\mathbf{x}(t),\theta_k)\mathbf{x}(t)}{\sum_{t=1}^T P(r(t)=c|\mathbf{x}(t),\theta_k)} \tag{16}$$

$$C_c(k+1)=\frac{\sum_{t=1}^T P(r(t)=c|\mathbf{x}(t),\theta_k)(\mathbf{x}(t)-\mu_c(k+1))(\mathbf{x}(t)-\mu_c(k+1))^T}{\sum_{t=1}^T P(r(t)=c|\mathbf{x}(t),\theta_k)} \tag{17}$$

3.5 The optimization of $J(\theta|\theta_k)$ with respect to the distribution of r, i.e. the weights π_c, is more complicated because it is a constrained optimization problem: $\pi_c\geq 0$ and $\sum_c\pi_c=1$. There is a trick to convert the constrained optimization problem into an unconstrained one: Write π_c as

$$\pi_c=\frac{\exp(\gamma_c)}{\sum_{k=1}^C\exp(\gamma_k)}, \tag{18}$$

where $\gamma_c\in\mathbb{R}$. Verify that the trick works, i.e. that π_c as defined above satisfies the constraints for all γ_i.

3.6 Find the derivative $\nabla_{\gamma_c}J(\theta|\theta_k)$, and set it to zero to find the EM-update rule for π_c:

$$\pi_c(k+1)=\frac{1}{T}\sum_{t=1}^T P(r(t)=c|\mathbf{x}(t),\theta_k) \tag{19}$$