Handed out: April 21 (Wed)

Hints for solution: Exercise class on April 23 (Fr)

Hand in: April 28 (Wed), the latest, @ Room A348

This assignment gives you maximally 5% worth of extra points for the computer assignments and the final exam. Note: Not all exercises have equal weight. This is the last assignment!

Ex. 1 — More on the general form of EM for Mixture of Gaussians (gives 1 %)
In the lecture hand-out, we needed to calculate the integral in Eq. 10.20 for the E-step of the EM algorithm. This form is the general case where the data might not be iid. Here, we derive an expression for the simpler case of iid data.

1. With the notation of Eq. 10.20, assume that

$$p(X, S; \theta) = \prod_t p(\mathbf{x}_t|\mathbf{s}_t; \theta)p(\mathbf{s}_t; \theta) \tag{1}$$

   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. Show that $p(X; \theta) = \prod_t p(\mathbf{x}_t; \theta)$.
2. Show that Eq. 10.20 becomes

$$J(\theta) = \sum_{t=1}^{T} \mathbb{E}_t \log p(\mathbf{x}_t, \mathbf{s}_t; \theta) \tag{2}$$

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

$$J(\theta) = \sum_{t=1}^{T} \int \log p(\mathbf{x}_t, \mathbf{s}_t; \theta)p(\mathbf{s}_t|\mathbf{x}_t; \theta_{k-1})\mathrm{d}\mathbf{s}_t \tag{3}$$

   while for discrete data it is

$$J(\theta) = \sum_{t=1}^{T} \sum_{\mathbf{s}_t} \log p(\mathbf{x}_t, \mathbf{s}_t; \theta)p(\mathbf{s}_t|\mathbf{x}_t; \theta_{k-1}). \tag{4}$$

3. In the lecture hand-out the posterior $p(\mathbf{s}_t|\mathbf{x}_t; \theta_{k-1})$ is given by $q^*_{t,c}$ with $\mathbf{s}_t \equiv r(t)$. Rederive it from the definition of $p(\mathbf{x}(t), r(t))$ in Eq. 10.9.

Ex. 2 — Clustering for binary data (gives 1% for question 1-3 and 3% for 4 to 9)
Assume that the binary random vector $\mathbf{u} \in \{0,1\}^n$ has the distribution

$$p(\mathbf{u};\boldsymbol{\mu}) = \prod_{i=1}^{n} p(u_i;\mu_i) \tag{5}$$

with $\boldsymbol{\mu} = (\mu_1 \dots \mu_n)$, where

$$p(u_i;\mu_i) = \mu_i^{u_i}(1-\mu_i)^{1-u_i}. \tag{6}$$

That is, the elements u_i of the random vector $\mathbf{u} = (u_1, \dots u_n)$ are all independent from each other.

1. Calculate the mean and variance of $\mathbf{u}_i$.
2. Calculate the mean and covariance matrix of $\mathbf{u}$.
3. A binary random vector $\mathbf{x} = (x_1, \dots x_n)$ is said to follow a mixture of multivariate Bernoulli distributions if its distribution $q(\mathbf{x})$ is

$$q(\mathbf{x};\boldsymbol{\mu_c},\pi_c, c = 1 \dots C) = \sum_{c=1}^{C} \pi_c p(\mathbf{x};\boldsymbol{\mu_c}). \tag{7}$$

   where $\sum_{c=1}^{C} \pi_c = 1$. Show that the mean is

$$\mathbb{E}x = \sum_{c=1}^{C} \pi_c \boldsymbol{\mu_c} \tag{8}$$

   and calculate the covariance matrix of $\mathbf{x}$. Are the x_i still uncorrelated?
4. Assume now that you are given a sample $(\mathbf{x}(1), \dots \mathbf{x}(T))$ of size T of the random vector $\mathbf{x}$. What is the log-likelihood for the sample? (Give an expression similar to Eq. 10.13 in case of Gaussian mixtures).
5. We consider now the distribution $q(\mathbf{x})$ in Eq. (7) to be the marginal of the joint distribution $q(\mathbf{x}, r)$ where $r \in \{1, \dots, C\}$ is a hidden variable which denotes a cluster. The joint distribution $q(\mathbf{x}, r)$ is

$$q(\mathbf{x}, r) = p(\mathbf{x};\boldsymbol{\mu_r})\pi_r, \tag{9}$$

   and the conditional distribution $q(\mathbf{x}|r)$ is $p(\mathbf{x};\boldsymbol{\mu_r})$, defined in Eq. (5). What is the log-likelihood $\ell(\boldsymbol{\mu_c}, \pi_c)$ assuming that for each data point $\mathbf{x}(t)$ we also observe the class membership $r(t)$?
6. Calculate the posterior probability that a sample $\mathbf{x}(t)$ belongs to cluster c, i.e. calculate $q(r(t) = c|\mathbf{x}(t))$. Show that it equals

$$q(r(t) = c|\mathbf{x}(t)) = \frac{\pi_c p(\mathbf{x}(t);\boldsymbol{\mu_c})}{\sum_{k=1}^{C} \pi_k p(\mathbf{x}(t);\boldsymbol{\mu_k})}, \tag{10}$$

with

$$p(\mathbf{x};\boldsymbol{\mu_c}) = \prod_{i=1}^{n} p(x_i;\mu_{ci}) \quad (11)$$

$$= \prod_{i=1}^{n} \mu_{ci}^{x_i}(1-\mu_{ci})^{1-x_i} \quad (12)$$

from Eq. (5) and (6). (We derive here a formula that corresponds to Eq. 10.14 in case of a gaussian mixture.)

7. What is the expected log-likelihood $J(\boldsymbol{\mu_c}, \pi_c) = \mathbb{E}(\ell(\boldsymbol{\mu_c}, \pi_c))$, as calculated in Equation (4) of the exercise 1. That is, for each sample t we replace $\log q(\mathbf{x}(t), r(t))$ by its expected value where the expectation is taken with respect to the posterior $q(r(t) = c|\mathbf{x}(t))$? (This corresponds to Eq. 10.21 for the case mixture of Gaussians)
8. Calculate the derivative of $J(\boldsymbol{\mu_c}, \pi_c)$ with respect to $\boldsymbol{\mu_c}$, considering the posteriors $q(r(t) = c|\mathbf{x}(t))$ to be fixed. Show that setting the derivative to zero gives

$$\boldsymbol{\mu_c} = \bar{\mathbf{x}}_c \quad (13)$$

where

$$\bar{\mathbf{x}}_c = \frac{\sum_{t=1}^{T} q(r(t) = c|\mathbf{x}(t))\mathbf{x}(t)}{\sum_{t=1}^{T} q(r(t) = c|\mathbf{x}(t))}. \quad (14)$$

9. Calculate the derivative of $J(\boldsymbol{\mu_c}, \pi_c)$ with respect to π_c under the contraint that $\sum_c \pi_c = 1$, i.e. that π_c gives a proper probability distribution. The technique to include the constraint is to formulate the Lagrangian $\tilde{J}$,

$$\tilde{J}(\pi_c, \lambda) = J + \lambda(1 - \sum_{c=1}^{C} \pi_c). \quad (15)$$

Calculate then the derivatives with respect to π_c and λ. Show that setting this gradient to zero gives

$$\pi_c = \frac{\sum_{t=1}^{T} q(r(t) = c|\mathbf{x}(t))}{T}, \quad (16)$$

This is the final step in the EM algorithm. Hurra! Putting all together, the algorithm is follows (see next page):

Bernoulli-EM:
First initialize the parameters $\boldsymbol{\mu_c}$ and π_c. Then

1. Calculate the posteriors $q(r = c|\mathbf{x})$ in Eq. (10) with the current set of parameters. (E-step)
2. Update the parameters $\boldsymbol{\mu_c}$ and π_c according the formula in Eq (13) and (16). (M-step)

till either the parameters don't change any more, or the objective J does not increase any more.