CS-E4820 Machine Learning: Advanced Probabilistic Methods
Pekka Marttinen, Paul Blomstedt, Homayun Afrabandpey, Reza Ashrafi, Betül Güvenç, Tianyu Cui, Pedram Daee, Marko Järvenpää, Santosh Hiremath (Spring 2019)
Exercise problems, round 5, due on Tuesday, 5th March 2019, at 23:55
Please return your solutions in MyCourses as a single PDF file.

Problem 1. *"EM for missing observations."*

Suppose random variables X_i follow a bivariate normal distribution $X_i \sim \mathcal{N}_2(0, \Sigma)$, where

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Suppose further that we have observations on $X_1 = (X_{11}, X_{12})^T$, $X_2 = (X_{21}, X_{22})^T$ and $X_3 = (X_{31}, X_{32})^T$, such that X_1 and X_3 are fully observed, and from X_2 we have observed only the second coordinate. Thus, our data matrix can be written as

$$\begin{bmatrix} x_{11} & x_{12} \\ ? & x_{22} \\ x_{31} & x_{32} \end{bmatrix},$$

where the rows correspond to the transposed observations $\mathbf{x}_1^T, \mathbf{x}_2^T, \mathbf{x}_3^T$. Suppose we want to learn the unknown parameter ρ using the EM-algorithm. Denote the missing observation by Z and derive the E-step of the algorithm, i.e., (1) write the complete data log-likelihood $\ell(\rho)$, (2) compute the posterior distribution of the missing observation, given the observed variables and current estimate for ρ, and (3) evaluate the expectation of $\ell(\rho)$ with respect to the posterior distribution of the missing observations.

Hints:

1. In general, for $X \sim \mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $X = (X_1, X_2)^T$, $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$, we have

   $$X_1 | X_2 = x_2 \sim \mathcal{N}\left(\mu_1 + \frac{\sigma_1}{\sigma_2}\rho(x_2 - \mu_2), (1 - \rho^2)\sigma_1^2\right),$$

   with ρ being the correlation coefficient.

2. For evaluating the expectation of $\ell(\rho)$, you can make use of the following two rules:

   - $\mathbf{x}_2^T \boldsymbol{\Sigma}^{-1} \mathbf{x}_2 = \text{trace}(\boldsymbol{\Sigma}^{-1} \mathbf{x}_2 \mathbf{x}_2^T)$.
   - if $X \sim \mathcal{N}(\mu, \sigma^2)$ then $\langle X^2 \rangle = \mu^2 + \sigma^2$.

Problem 2. *"Extension of the "simple example" from the lecture."*

Suppose that we have N independent observations $\mathbf{x} = (x_1, \ldots, x_N)$ from a two-component mixture of univariate Gaussian distributions with unknown mixing coefficients and unknown mean of the second component:

$$p(x_n | \theta, \tau) = (1 - \tau)\mathcal{N}(x_n | 0, 1) + \tau \mathcal{N}(x_n | \theta, 1).$$

(a) Write down the complete data log-likelihood and derive the EM-algorithm for learning the maximum likelihood estimates for θ and τ.

(b) Implement the EM-algorithm. Simulate some data from the model ($N = 100$ samples) with the true values of parameters $\theta = 3$ and $\tau = 0.5$. Run your EM algorithm to see whether the learned parameters converge close to the true values (by e.g. just listing the estimates from a few iterations or plotting them).

Instructions: Use `ex5_2_template.py` as a starting point. As a further aid, you can use `ex5_2_simple_em.py`, which is a Python implementation of `simple_example.pdf` from the lecture material. Inlclude your implementation of the E and M steps in your solutions.

Problem 3. *"Probabilistic progamming with Edward."*

Read the first four pages of the article: Dustin Tran, Alp Kucukelbir, Adji B. Dieng, Maja Rudolph, Dawen Liang, and David M. Blei (2016). *Edward: A library for probabilistic modeling, inference, and criticism*, https://arxiv.org/abs/1610.09787. Get familiar with Edward using the notebook `getting_started.ipynb`, and answer the following questions (you can also refer to Edward's documentation in http://edwardlib.org/ or any other source you can find):

1. What is a probabilistic programming language?
2. Why is probabilistic programming needed?
3. What is Edward?
4. What are the three steps in the iterative process for probabilistic modeling that Edward is built around?
5. Identify parts of the example probabilistic program (on p. 3–4) that correspond to the three steps mentioned in the previous question.