Exercises
Introduction to Machine Learning
FS 2020

Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause
Web: `https://las.inf.ethz.ch/teaching/introml-s20`
For questions, please refer to Piazza.

Series 7, May 30th, 2020 (Mixture Models, EM Algorithm)

Problem 1 (Mixture Models and Expectation-Maximization Algorithm):

Consider a one-dimensional Gaussian Mixture Model with 2 clusters and parameters $(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, w_1, w_2)$. Here (w_1, w_2) are the mixing weights, and $(\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2)$ are the centers and variances of the clusters. We are given a dataset $\mathcal{D} = \{x_1, x_2, x_3\} \subset \mathbb{R}$, and apply the EM-algorithm to find the parameters of the Gaussian mixture model.

1. What is the complete log-likelihood that is being optimized, for this problem?

   (a) ln $f(\mathcal{D}|(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, w_1, w_2)) = ln\{w_1\mathcal{N}(x_1; \mu_1, \sigma_1) + w_2\mathcal{N}(x_1; \mu_2, \sigma_2)\} + ln\{w_1\mathcal{N}(x_2; \mu_1, \sigma_1) + w_2\mathcal{N}(x_2; \mu_2, \sigma_2)\} + ln\{w_1\mathcal{N}(x_3; \mu_1, \sigma_1) + w_2\mathcal{N}(x_3; \mu_2, \sigma_2)\}$

   (b) ln $f(\mathcal{D}|(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, w_1, w_2)) = ln\{w_1\mathcal{N}(x_1; \mu_1, \sigma_1) - w_2\mathcal{N}(x_1; \mu_2, \sigma_2)\} + ln\{w_1\mathcal{N}(x_2; \mu_1, \sigma_1) - w_2\mathcal{N}(x_2; \mu_2, \sigma_2)\} + ln\{w_1\mathcal{N}(x_3; \mu_1, \sigma_1) - w_2\mathcal{N}(x_3; \mu_2, \sigma_2)\}$

   (c) ln $f(\mathcal{D}|(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, w_1, w_2)) = ln\{\frac{w_1}{w_1+w_2}\mathcal{N}(x_1; \mu_1, \sigma_1) + \frac{w_2}{w_1+w_2}\mathcal{N}(x_1; \mu_2, \sigma_2)\} + ln\{\frac{w_1}{w_1+w_2}\mathcal{N}(x_2; \mu_1, \sigma_1) + \frac{w_2}{w_1+w_2}\mathcal{N}(x_2; \mu_2, \sigma_2)\} + ln\{\frac{w_1}{w_1+w_2}\mathcal{N}(x_3; \mu_1, \sigma_1) + \frac{w_2}{w_1+w_2}\mathcal{N}(x_3; \mu_2, \sigma_2)\}$

   (d) ln $f(\mathcal{D}|(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, w_1, w_2)) = ln\{\frac{w_1}{w_1+w_2}\mathcal{N}(x_1; \mu_1, \sigma_1) - \frac{w_2}{w_1+w_2}\mathcal{N}(x_1; \mu_2, \sigma_2)\} + ln\{\frac{w_1}{w_1+w_2}\mathcal{N}(x_2; \mu_1, \sigma_1) - \frac{w_2}{w_1+w_2}\mathcal{N}(x_2; \mu_2, \sigma_2)\} + ln\{\frac{w_1}{w_1+w_2}\mathcal{N}(x_3; \mu_1, \sigma_1) - \frac{w_2}{w_1+w_2}\mathcal{N}(x_3; \mu_2, \sigma_2)\}$

   Solution:
   The correct answers are (a) and (c).
   ln $f(\mathcal{D}|(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2, w_1, w_2)) = ln\{w_1\mathcal{N}(x_1; \mu_1, \sigma_1) + w_2\mathcal{N}(x_1; \mu_2, \sigma_2)\} + ln\{w_1\mathcal{N}(x_2; \mu_1, \sigma_1) + w_2\mathcal{N}(x_2; \mu_2, \sigma_2)\} + ln\{w_1\mathcal{N}(x_3; \mu_1, \sigma_1) + w_2\mathcal{N}(x_3; \mu_2, \sigma_2)\}$
   Since $w_1 + w_2 = 1$, even (c) is a correct solution.

   Assume that the dataset $\mathcal{D}$ consists of the following three points, $x_1 = 1, x_2 = 10, x_3 = 20$. At some step in the EM-algorithm, we compute the expectation step which results in the following matrix: $R = \begin{pmatrix} 1 & 0 \\ 0.4 & 0.6 \\ 0 & 1 \end{pmatrix}$.
   where r_{ic} denotes the probability of x_i belonging to cluster c.

   Given the above R for the expectation step, write the result of the maximization step for the mixing weights w_1, w_2. Round your answer to two decimal points.

2. $w_1 =$ **Solution:**
   $w_1 = 0.47$

3. $w_2 =$ **Solution:**
   $w_2 = 0.53$

$$w_1 = \frac{1}{3}(1 + 0.4 + 0) = \frac{1.4}{3}$$
$$w_2 = \frac{1}{3}(0 + 0.6 + 1) = \frac{1.6}{3}$$

Given the above R for the expectation step, write the result of the maximization step for the centers μ_1, μ_2. Round your answer to two decimal points.

4. $\mu_1 =$ **Solution:**
   $\mu_1 = 3.57$

5. $\mu_2 =$ **Solution:**
   $\mu_2 = 16.25$

$$\mu_k = \frac{1}{N_k}\Sigma_{n=1}^{N}\gamma_k(x_n)x_n$$

where $N_k = \Sigma_{n=1}^{N}\gamma_k(x_n)$.
For this example,

$$\mu_1 = \frac{1}{1.4}(1\cdot 1 + 0.4\cdot 10 + 0\cdot 20) = \frac{5}{1.4}$$
$$\mu_2 = \frac{1}{1.6}(0\cdot 1 + 0.6\cdot 10 + 1\cdot 20) = \frac{26}{1.6}$$

Given the above R for the expectation step, write the result of the maximization step for the variance values σ_1^2, σ_2^2 . Round your answer to two decimal points.

6. $\sigma_1^2 =$ **Solution:**
   $\sigma_1^2 = 16.53$

7. $\sigma_2^2 =$ **Solution:**
   $\sigma_2^2 = 23.44$

$$\sigma_k^2 = \frac{1}{N_k}\Sigma_{n=1}^{N}\gamma_k(x_n)(x_n - \mu_k)^2$$

where $N_k = \Sigma_{n=1}^{N}\gamma_k(x_n)$.
For this example,

$$\mu_1 = \frac{1}{1.4}(1\cdot(1-\frac{5}{1.4})^2 + 0.4\cdot(10-\frac{5}{1.4})^2 + 0\cdot(20-\frac{5}{1.4})^2)$$
$$\mu_2 = \frac{1}{1.6}(0\cdot(1-\frac{26}{1.6})^2 + 0.6\cdot(10-\frac{26}{1.6})^2 + 1\cdot(20-\frac{26}{1.6})^2)$$

The previous two questions are doing soft-EM. Calculate the maximization step of $\hat{\mu}_1, \hat{\mu}_2$ for hard-EM.

8. $\hat{\mu}_1 =$ **Solution:**
   $\hat{\mu}_1 = 1$

9. $\hat{\mu}_2 =$ **Solution:**
   $\hat{\mu}_2 = 15$

$$\hat{\mu}_1 = \frac{1}{1}(1) = 1$$
$$\hat{\mu}_2 = \frac{1}{2}(10+20) = 15$$

Problem 2 (Mixture Models and Maximum a Posteriori estimation):

We are given a dataset $\mathcal{D} = \{\mathbf{x_1}, ..., \mathbf{x_n}\} \subset \mathbb{R}^d$. Consider a mixture of K multivariate Bernoulli distributions with parameters $\mu = (\mathbf{\mu_1}, \mathbf{\mu_2}, ..., \mathbf{\mu_K})$, where $\mathbf{\mu_k} = \{\mu_{k1}, ...\mu_{kd}\}$. You will use EM algorithm to compute MLE and MAP estimates.

10. What is the M step for $\mu_k i$ using MLE? Select the correct answer. Here, r_{nk} is the responsibility of the data point $\mathbf{x_n}$ belonging cluster center $\mathbf{\mu_k}$, as computed in the E step.

(a) $\mu_{ki} = \frac{\Sigma_{n=1}^{N} r_{nk} x_{ni}}{\Sigma_{n=1}^{N} r_{nk}}$

(b) $\mathbb{E}[log(p(x, z|\pi, \mu))] = \Sigma_{n=1}^{N} \Sigma_{k=1}^{K} r_{nk} (log\pi_k + \Sigma_{i=1}^{d} (x_{ni} log\mu_{ki}$

(c) $\mu_{ki} = \frac{\Sigma_{n=1}^{N} x_{ni}}{N}$

(d) $\mathbb{E}[log(p(x, z|\pi, \mu))] = \Sigma_{n=1}^{N} \Sigma_{k=1}^{K} r_{nk} (\Sigma_{i=1}^{d} (x_{ni} log\mu_{ki} + (1 - x_{ni}) log(1 - \mu_{ki})))$

Solution:
The correct answer is (a).
We have K mixture components where each component is a vector of d independent Bernoullis. In other words,

$$p(x|\pi, \mu) = \Sigma_{k=1}^{K} \pi_k p(x|\mu) = \Sigma_{k=1}^{K} \pi_k \Pi_{i=1}^{d} \mu_{ki}^{x_i} (1 - \mu_{ki})^{1-x_i}$$

Expected value of the complete data log-likelihood can be written as:

$$\mathbb{E}[log(p(x, z|\pi, \mu))] = \Sigma_{n=1}^{N} \Sigma_{k=1}^{K} r_{nk} \left(log\pi_k + \Sigma_{i=1}^{d} (x_{ni} log\mu_{ki} + (1 - x_{ni}) log(1 - \mu_{ki})) \right)$$

where r_{nk} denotes the posterior probability from the E step. Note that the derivative of Bernoulli distribution is $\frac{x_{ni}}{\mu_{ki}} - \frac{(1-x_{ni})}{(1-\mu_{ki})}$. Taking the derivative with respect to μ_{ki} and setting it to zero gives you

$$\mu_{ki} = \frac{\Sigma_{n=1}^{N} r_{nk} x_{ni}}{\Sigma_{n=1}^{N} r_{nk}}$$

11. Now, suppose you want to do MAP estimation. What is the E step? Select the correct answer.

(a) $r_{nk} = \frac{\pi_k \Pi_{i=1}^{d} \mu_{ki}^{x_{ni}} (1-\mu_{ki})^{1-x_{ni}}}{\Sigma_{k=1}^{K} \pi_k \Pi_{i=1}^{d} \mu_{ki}{}^{x_{ni}} (1-\mu_{ki})^{1-x_{ni}}}$

(b) $r_{nk} = \frac{\Pi_{i=1}^{d} \mu_{ki}^{x_{ni}} (1-\mu_{ki})^{1-x_{ni}}}{\Sigma_{k=1}^{K} \Pi_{i=1}^{d} \mu_{ki}{}^{x_{ni}} (1-\mu_{ki})^{1-x_{ni}}}$

(c) $r_{nk} = \frac{\pi_n \Pi_{i=1}^{d} \mu_{ki}^{x_{ni}} (1-\mu_{ki})^{1-x_{ni}}}{\Sigma_{n=1}^{N} \pi_n \Pi_{i=1}^{d} \mu_{ki}{}^{x_{ni}} (1-\mu_{ki})^{1-x_{ni}}}$

(d) $r_{nk} = \frac{\Pi_{i=1}^{d} \mu_{ki}^{x_{ni}} (1-\mu_{ki})^{1-x_{ni}}}{\Sigma_{n=1}^{N} \Pi_{i=1}^{d} \mu_{ki}{}^{x_{ni}} (1-\mu_{ki})^{1-x_{ni}}}$

Solution:
The correct answer is (a).
The E Step is the same for the MLE case, namely

$$r_{nk} = \frac{\pi_k \Pi_{i=1}^{d} \mu_{ki}^{x_{ni}} (1 - \mu_{ki})^{1-x_{ni}}}{\Sigma_{k=1}^{K} \pi_k \Pi_{i=1}^{d} \mu_{ki}^{x_{ni}} (1 - \mu_{ki})^{1-x_{ni}}}$$

12. What is the M step for μ_{ki} using MAP? You can assume a $Beta(\alpha, \beta)$ prior. Select the correct answer.

(a) $\mu_{ki} = \frac{\Sigma_{n=1}^{N} (r_{nk} x_{ni}) + \alpha - 1)}{\Sigma_{n=1}^{N} (r_{nk}) + \alpha + \beta - 2}$

(b) $\mu_{ki} = \frac{\Sigma_{n=1}^{N} (r_{nk} x_{ni}) + \alpha)}{\Sigma_{n=1}^{N} (r_{nk}) + \alpha + \beta - 1}$

(c) $\mu_{ki} = \frac{\Sigma_{n=1}^{N} (r_{nk} x_{ni}) + \alpha)}{\Sigma_{n=1}^{N} (r_{nk}) + \alpha + \beta}$

(d) $\mu_{ki} = \frac{\Sigma_{n=1}^{N} (r_{nk} x_{ni}) + \beta)}{\Sigma_{n=1}^{N} (r_{nk}) + \alpha + \beta}$

Solution:
The correct answer is (a).

According to Bayes' theorem:

$$p(\theta|\mathbf{X}) \propto p(\mathbf{X}|\theta)p(\theta)$$

$$logp(\theta|\mathbf{X}) = logp(\mathbf{X}|\theta) + logp(\theta) + c$$

where c is an arbitrary constant.

Therefore, we need to add a log prior to the expected value of the complete data log-likelihood. The function we need to maximize is $\mathbb{E}[log(p(x, z|\pi, \mu))] + logp(\mu)$, where $p(\mu) = \Pi_{k=1}^{K}\Pi_{i=1}^{d}p(\mu_{ki})$ and

$$p(\mu_{ki}) = \frac{\mu_{ki}^{\alpha-1}(1-\mu_{ki})^{\beta-1}}{\mathcal{B}(\alpha, \beta)}$$

We can write

$$logp(\mu) = \Sigma_{k=1}^{K}\Sigma_{i=1}^{d}(\alpha-1)log\mu_{ki} + (\beta-1)(1-log\mu_{ki}) - log\mathcal{B}(\alpha, \beta)$$

We take derivative of the following expression with respect to μ_{ki} and set it to zero:

$$\Sigma_{n=1}^{N}\Sigma_{k=1}^{K}r_{nk}\left(log\pi_k + \Sigma_{i=1}^{d}(x_{ni}log\mu_{ki} + (1-x_{ni})log(1-\mu_{ki}))\right) +$$

$$\Sigma_{k=1}^{K}\Sigma_{i=1}^{d}(\alpha-1)log\mu_{ki} + (\beta-1)log(1-mu_{ki})$$

which gives

$$\mu_{ki} = \frac{\Sigma_{n=1}^{N}(r_{nk}x_{ni}) + \alpha - 1)}{\Sigma_{n=1}^{N}(r_{nk}) + \alpha + \beta - 2}$$

Problem 3 (A Different Perspective on EM):

In this question you will show that EM can be seen as an iterative algorithm which maximizes a lower bound on the log-likelihood. We will treat any general model $P(X, Z)$ with observed variables X and latent variable Z. For the sake of simplicity, we will assume that Z is discrete and takes values in $1, 2, ..., m$. If we observe X, the goal is to maximize the log-likelihood

$$l(\theta) = logP(\mathbf{x}; \theta) = log\Sigma_{z=1}^{m}P(\mathbf{x}, z; \theta)$$

with respect to the parameter vector θ. $Q(Z)$ denotes any distribution over the latent variables.

13. For $Q(z) > 0$ when $P(\mathbf{x}, z) > 0$, find a lower bound for the likelihood, $l(\theta)$. Hint: Consider using the Jensen's inequality.

    (a) $\mathbb{E}_Q[logP(X, Z)] - \Sigma_{z=1}^{m}Q(z)logQ(z)$
    (b) $\mathbb{E}_Q[logP(X, Z)] + \Sigma_{z=1}^{m}Q(z)logQ(z)$
    (c) $\mathbb{E}_Q[logP(X, Z)]$
    (d) $\mathbb{E}_Q[logP(X, Z)] + \Sigma_{z=1}^{m}Q(\mathbf{x})logQ(\mathbf{x})$

    Solution:
    The correct answer is (a).

$$
\begin{aligned}
l(\theta) &= logP(\mathbf{x};\theta) \\
&= log\Sigma_{z=1}^{m} P(\mathbf{x}, z;\theta) \\
&= log\Sigma_{z=1}^{m} \frac{P(\mathbf{x}, z;\theta)}{Q(z)} Q(z) \\
&= log\mathbb{E}_{Z\sim Q}[\frac{P(\mathbf{x}, z;\theta)}{Q(z)}] \\
&\geq \mathbb{E}_{Z\sim Q}[log\frac{P(\mathbf{x}, z;\theta)}{Q(z)}] \\
&= \mathbb{E}_{Z\sim Q}[logP(\mathbf{x}, z;\theta)] - \Sigma_{z=1}^{m} Q(z) logQ(z),
\end{aligned}
$$

where for the inequality we have used Jensen's inequality.

14. For a fixed θ, pick the distribution $Q^*(Z)$ which maximizes the lower bound derived in the previous question. Show by yourself that bound is exact for this specific distribution. Hint: Do not forget to add Lagrange multipliers to make sure that Q^* is a valid distribution.

    (a) $P(Z|\mathbf{x};\theta)$
    (b) $P(Z;\theta)$
    (c) $P(\mathbf{X}|z;\theta)$
    (d) $P(\mathbf{X}, Z;\theta)$

    Solution:
    The correct answer is (a).
    Now, assume that we want to maximize the abovewith respect to Q, and let us add a multiplier λ to make sure that Q sums up to 1. Then, we have the following Lagrangian

    $$\mathcal{L}(Q,\lambda) = \Sigma_{z=1}^{m} Q(z) logP(\mathbf{x}, z;\theta) - \Sigma_{z=1}^{m} Q(z) logQ(z) + \lambda(\Sigma_{z=1}^{m} Q(z) - 1)$$

    By setting the derivative of the Lagrangian with respect to $Q(z)$ to zero, we have

    $$\frac{\partial}{\partial Q(z)} \mathcal{L}(Q,\lambda) = logP(\mathbf{x}, z;\theta) - 1 - logQ(z) + \lambda = 0 \implies Q(z) = e^{\lambda-1} P(\mathbf{x}, z;\theta)$$

    . Hence, we have that $Q(z) \propto P(\mathbf{x}, z;\theta)$ and this is exactly the posterior $P(Z|\mathbf{x};\theta)$, which we had to show. It is also easy to see that the bound is tight, as

    $$\mathbb{E}_{Z\sim Q}[log\frac{P(\mathbf{x}, z;\theta)}{Q(z)}] = \Sigma_{z=1}^{m} Q(z) log\frac{P(\mathbf{x}, z;\theta)}{Q(z)} = \Sigma_{z=1}^{m} P(Z|\mathbf{x};\theta) log\frac{P(Z|\mathbf{x};\theta)P(\mathbf{x};\theta)}{P(Z|\mathbf{x};\theta)} = logP(\mathbf{x};\theta)$$

15. Mark the following statements True or False.

    (a) Optimizing the lower bound on likelihood with respect to $Q(.)$ is exactly the E-step.
    (b) Optimizing the lower bound on likelihood with respect to $Q(.)$ is exactly the M-step.
    (c) Optimizing the lower bound on likelihood with respect to θ for fixed $Q(.)$ is exactly the E-step.
    (d) Optimizing the lower bound on likelihood with respect to θ for fixed $Q(.)$ is exactly the M-step.
    (e) The lower bound on likelihood monotonically increases after each step of optimisation.
    (f) The lower bound on likelihood monotonically decreases after each step of optimisation.

Solution:
(a), (d) and (e) are True statements.
We can easily see the EM algorithm as optimizing the lower bound with respect to Q and θ in an alternating manner. Specifically, if we optimize with respect to Q we have shown that the optimal Q is the posterior, and this is exactly the E-step. Optimizing with respect to θ for fixed Q is clearly equivalent to the M-step. As the lower bound is monotonically increased at every step the EM algorithm has to converge.