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

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

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

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

Solution:
The correct answers are (a) and (c).

Assume that the dataset \(D\) 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. \(w_1 = \text{Solution:} \quad w_1 = 0.47\)

3. \(w_2 = \text{Solution:} \quad w_2 = 0.53\)

Given the above \(R\) for the expectation step, write the result of the maximization step for the centers \(\mu_1, \mu_2\). Round your answer to two decimal points.
4. $\mu_1 = \text{Solution:}
\mu_1 = 3.57$

5. $\mu_2 = \text{Solution:}
\mu_2 = 16.25$

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

where $N_k = \sum_{n=1}^{N} \gamma_k(x_n)$.

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 $\sigma_1^2, \sigma_2^2$. Round your answer to two decimal points.

6. $\sigma_1^2 = \text{Solution:}
\sigma_1^2 = 16.53$

7. $\sigma_2^2 = \text{Solution:}
\sigma_2^2 = 23.44$

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

where $N_k = \sum_{n=1}^{N} \gamma_k(x_n)$.

For this example,

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

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

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

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

$$\hat{\mu}_1 = \frac{1}{1}(1) = 1$$

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

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

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

10. What is the M step for $\mu_{ki}$ using MLE? Select the correct answer. Here, $r_{nk}$ is the responsibility of the data point $x_n$ belonging cluster center $\mu_k$, as computed in the E step.
11. Now, suppose you want to do MAP estimation. What is the E step? Select the correct answer.

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

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

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

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

Solution:
The correct answer is (a).

The E step is the same for the MLE case, namely

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

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

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

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

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

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

Solution:
The correct answer is (a).
According to Bayes’ theorem:

\[ p(\theta | X) \propto p(X | \theta) p(\theta) \]

\[ \log p(\theta | X) = \log p(X | \theta) + \log p(\theta) + c \]

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 \( \mathbb{E}[\log p(x, z | \pi, \mu)] + \log p(\mu) \), where \( p(\mu) = \prod_{k=1}^{K} \prod_{i=1}^{d} p(\mu_{ki}) \) and

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

We can write

\[ \log p(\mu) = \sum_{k=1}^{K} \sum_{i=1}^{d} (\alpha - 1) \log \mu_{ki} + (\beta - 1)(1 - \log \mu_{ki}) - \log B(\alpha, \beta) \]

We take derivative of the following expression with respect to \( \mu_{ki} \) and set it to zero:

\[ \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left( \log \pi_k + \sum_{i=1}^{d} (x_{ni} \log \mu_{ki} + (1 - x_{ni}) \log (1 - \mu_{ki})) \right) + \sum_{k=1}^{K} \sum_{i=1}^{d} (\alpha - 1) \log \mu_{ki} + (\beta - 1) \log (1 - \mu_{ki}) \]

which gives

\[ \mu_{ki} = \frac{\sum_{n=1}^{N} r_{nk} x_{ni} + \alpha - 1}{\sum_{n=1}^{N} r_{nk} + \alpha + \beta - 2} \]

**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

\[ l(\theta) = \log P(x; \theta) = \log \sum_{z=1}^{m} P(x, z; \theta) \]

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

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

(a) \( \mathbb{E}_Q[\log P(X, Z)] - \sum_{z=1}^{m} Q(z) \log Q(z) \)

(b) \( \mathbb{E}_Q[\log P(X, Z)] + \sum_{z=1}^{m} Q(z) \log Q(z) \)

(c) \( \mathbb{E}_Q[\log P(X, Z)] \)

(d) \( \mathbb{E}_Q[\log P(X, Z)] + \sum_{z=1}^{m} Q(x) \log Q(x) \)

**Solution:**

The correct answer is (a).
\[ l(\theta) = \log P(x; \theta) \]
\[ = \log \sum_{z=1}^{m} P(x, z; \theta) \]
\[ = \log \sum_{z=1}^{m} \frac{P(x, z; \theta)}{Q(z)} \cdot Q(z) \]
\[ = \log \mathbb{E}_{Z \sim Q}[\frac{P(x, z; \theta)}{Q(z)}] \]
\[ \geq \mathbb{E}_{Z \sim Q}[\log \frac{P(x, z; \theta)}{Q(z)}] \]
\[ = \mathbb{E}_{Z \sim Q}[\log P(x, z; \theta)] - \sum_{z=1}^{m} Q(z) \log Q(z), \]

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

14. For a fixed \( \theta \), 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) \( P(Z|x; \theta) \)
(b) \( P(Z; \theta) \)
(c) \( P(X|z; \theta) \)
(d) \( P(X, Z; \theta) \)

Solution:
The correct answer is (a).

Now, assume that we want to maximize the above with respect to \( Q \), and let us add a multiplier \( \lambda \) to make sure that \( Q \) sums up to 1. Then, we have the following Lagrangian

\[ L(Q, \lambda) = \sum_{z=1}^{m} Q(z) \log P(x, z; \theta) - \sum_{z=1}^{m} Q(z) \log Q(z) + \lambda (\sum_{z=1}^{m} Q(z) - 1) \]

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

\[ \frac{\partial}{\partial Q(z)} L(Q, \lambda) = \log P(x, z; \theta) - 1 - \log Q(z) + \lambda = 0 \implies Q(z) = e^{\lambda - 1} P(x, z; \theta) \]

. Hence, we have that \( Q(z) \propto P(x, z; \theta) \) and this is exactly the posterior \( P(Z|x; \theta) \), which we had to show.

It is also easy to see that the bound is tight, as

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

15. Mark the following statements True or False.

(a) Optimizing the lower bound on likelihood with respect to \( Q(\cdot) \) is exactly the E-step.
(b) Optimizing the lower bound on likelihood with respect to \( Q(\cdot) \) is exactly the M-step.
(c) Optimizing the lower bound on likelihood with respect to \( \theta \) for fixed \( Q(\cdot) \) is exactly the E-step.
(d) Optimizing the lower bound on likelihood with respect to \( \theta \) for fixed \( Q(\cdot) \) 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 $\theta$ 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 $\theta$ 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.