“Machine Learning and Computational Statistics”

6^{th} Homework

Exercise 1 (multiple choices question): In mixture modeling, an unknown probability density function (pdf) $p(\boldsymbol{x})$ is expressed as a linear combination of K different pdfs, $p(\boldsymbol{x}|k)$, that is, $p(\boldsymbol{x})=\sum_{k=1}^{K}P_k p(\boldsymbol{x}|k)$, where P_k are weighting parameters associated with the $p(\boldsymbol{x}|k)$’s. Which of the following statements is/are true?

1. The only constraint for the parameters P_k is that they should be non-negative.

2. If $\sum_{k=1}^{K}P_k=1$, then it is guaranteed that $p(\boldsymbol{x})$ has the properties of a pdf.

3. In general, $p(\boldsymbol{x})$ can be a multimodal pdf.

4. In the case where (i) the $p(\boldsymbol{x}|k)$’s are Gaussian pdfs and (ii) all P_k’s are zero except one, which is positive, then $p(\boldsymbol{x})$ is definitely multimodal.

5. If the weighting parameters have a probability interpretation, then it is guaranteed that $p(\boldsymbol{x})$ has the properties of a pdf.

Exercise 2 (multiple choices question): Consider a machine learning task for which a set of N observations, $\boldsymbol{x}_1,\dots,\boldsymbol{x}_N$, is available. The aim is to estimate the probability density function (pdf), $p(\boldsymbol{x})$, that generates them. Assume that $p(\boldsymbol{x})$ is expressed as a linear combination of K different pdfs, $p(\boldsymbol{x}|k)$, that is, $p(\boldsymbol{x})=\sum_{k=1}^{K}P_k p(\boldsymbol{x}|k)$, where P_k are weighting parameters associated with the $p(\boldsymbol{x}|k)$’s. Which of the following statements regarding the interpretation of the above linear combination of pdfs is/are true?

1. Each observation $\boldsymbol{x}_n$ is drawn from a single mixture $p(\boldsymbol{x}|k)$.

2. The specific mixture $p(\boldsymbol{x}|k)$ that emits an observation, $\boldsymbol{x}_n$, it is known.

3. The probability that an observation $\boldsymbol{x}_n$ is drawn from the k-th mixture $p(\boldsymbol{x}|k)$ equals to P_k

4. The probability that $\boldsymbol{x}_n$ occurs is P_k.

NOTE: For exercises 3-12 below, the parameter vectors $\boldsymbol{\theta}_j$’s are denoted by $\boldsymbol{\xi}_j$’s and the set of $\boldsymbol{\Theta}$ of $\boldsymbol{\theta}_j$’s is denoted by $\boldsymbol{\Xi}$.

Exercise 3 (multiple choices question): Consider a machine learning task for which a set of N observations, $\boldsymbol{x}_1,\dots,\boldsymbol{x}_N$, is available. The aim is to estimate the probability density function (pdf) $p(\boldsymbol{x})$ that generates them. Assume that $p(\boldsymbol{x})$ is expressed as a linear combination of K different distributions, $p(\boldsymbol{x}|k)$, that is, $p(\boldsymbol{x}) = \sum_{k=1}^{K} P_k p(\boldsymbol{x}|k)$, where P_k are weighting parameters associated with the $p(\boldsymbol{x}|k)$'s. Which of the following statements is/are true?

1. For a large enough value of K and independently of the choice of the parameter values of the mixtures, $p(\boldsymbol{x})$ can approximate arbitrarily close any continuous pdf.

2. The number K of the mixtures is always known a priori.

3. In mixture modeling, N hidden variables are involved; the labels of the mixtures from which the observations are originated.

4. Under certain conditions, a Gaussian mixture model can approximate any continuous pdf for a large enough number, K, of mixtures.

Exercise 4 (multiple choices question): Consider a machine learning task for which a set of N observations, $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$, is available. The aim is to estimate the probability density function (pdf), $p(\boldsymbol{x})$, that generates them. Assume that $p(\boldsymbol{x})$ is expressed as $p(\boldsymbol{x}) = \sum_{k=1}^{K} P_k p(\boldsymbol{x}|k; \boldsymbol{\xi}_k)$, where the distributions $p(\boldsymbol{x}|k; \boldsymbol{\xi}_k)$ are of the same parametric form, each one parameterized via a vector $\boldsymbol{\xi}_k$, $k = 1, \dots, K$. Also, P_k's are the weighting parameters associated with the $p(\boldsymbol{x}|k; \boldsymbol{\xi}_k)$'s. Which of the following statements is/are true?

1. For the approximation of $p(\boldsymbol{x})$, via the mixture modelling approach, only the estimation of $\boldsymbol{\xi}_k$'s, $k = 1, \dots, K$, based on $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$, is required.

2. Each latent variable, k_n, associated with $\boldsymbol{x}_n$, $n = 1, \dots, N$, is an integer taking values in the set $\{1, \dots, N\}$.

3. Obtaining K is achieved via a cost function optimization task.

4. The set of the unobserved variables in the mixture modelling framework comprises the indices of the mixture pdfs from which the observations have been drawn.

Exercise 5 (multiple choices question): Consider a machine learning task for which a set of N l-dimensional observation vectors, $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$, is available. The aim is to estimate the probability density function (pdf) $p(\boldsymbol{x})$ that generates them. Assume that $p(\boldsymbol{x})$ is expressed as $p(\boldsymbol{x}) = \sum_{k=1}^{K} P_k p(\boldsymbol{x}|k; \boldsymbol{\xi}_k)$, where $p(\boldsymbol{x}|k; \boldsymbol{\xi}_k) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, {\sigma_k}^2 I)$ are Gaussian distributions , each one parameterized via $\boldsymbol{\mu}_k$ and ${\sigma_k}^2$, $k = 1, \dots, K$ (i.e., $\boldsymbol{\xi}_k = [{\boldsymbol{\mu}_k}^T, {\sigma_k}^2]^T$), and P_k is the weighting parameter associated with $p(\boldsymbol{x}|k; \boldsymbol{\xi}_k), k = 1,2, \dots K$. What is the total number of parameters that need to be estimated in order to determine $p(\boldsymbol{x})$?

1. $(l + 1) \cdot K$

2. $(K+2)\cdot l$

3. $(l+2)\cdot K$

4. $l\cdot K$

Exercise 6 (multiple choices question): Consider a machine learning task for which a set of N l-dimensional observation vectors, $\boldsymbol{x}_1,\dots,\boldsymbol{x}_N$, is available. The aim is to estimate the probability density function (pdf) $p(\boldsymbol{x})$ that generates them. Assume that $p(\boldsymbol{x})$ is expressed as $p(\boldsymbol{x}) = \sum_{k=1}^{K} P_k p(\boldsymbol{x}|k;\boldsymbol{\xi}_k)$, where $p(\boldsymbol{x}|k;\boldsymbol{\xi}_k) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k,\Sigma_k)$ are Gaussian distributions , each one parameterized by $\boldsymbol{\mu}_k$ and Σ_k, $k=1,\dots,K$ (i.e., $\boldsymbol{\xi}_k$ comprises all the elements of comprises all the elements of $(\boldsymbol{\mu}_k,\Sigma_k)$), and P_k is the weighting parameter associated with $p(\boldsymbol{x}|k;\boldsymbol{\xi}_k), k = 1,2,\dots K$. What is the total number of parameters that need to be estimated in order to determine $p(\boldsymbol{x})$?

Hint: Σ_k is completely described by its $\frac{l(l+1)}{2}$ diagonal and upper diagonal elements and $\boldsymbol{\xi}_k$ comprises no duplicates.

1. $\left(\frac{l(l+3)}{2}+1\right)\cdot K$

2. $\frac{l(l+1)}{2}\cdot l + K$

3. $\frac{l(l+1)}{2}\cdot K$

4. $\frac{l(l+1)}{2}\cdot K + l$

Exercise 7 (multiple choices question): Consider a machine learning task for which a set $\mathcal{X}$ of N l-dimensional i.i.d. observation vectors, $\boldsymbol{x}_1,\dots,\boldsymbol{x}_N$, is available. The aim is to estimate the probability density function (pdf) $p(\boldsymbol{x})$ that generates them. Assume that $p(\boldsymbol{x})$ is expressed as $p(\boldsymbol{x}) = \sum_{k=1}^{K} P_k p(\boldsymbol{x}|k;\boldsymbol{\xi}_k)$, where $p(\boldsymbol{x}|k;\boldsymbol{\xi}_k) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k,{\sigma_k}^2 I)$ are Gaussian distributions, each one parameterized by $\boldsymbol{\mu}_k$ and ${\sigma_k}^2$, $k=1,\dots,K$ (i.e., $\boldsymbol{\xi}_k = [{\boldsymbol{\mu}_k}^T,{\sigma_k}^2]^T$), and P_k is the weighting parameter associated with $p(\boldsymbol{x}|k;\boldsymbol{\xi}_k)$'s. Let $\boldsymbol{\Xi} = \left[{\boldsymbol{\xi}_1}^T,\dots,{\boldsymbol{\xi}_K}^T\right]^T$ and $\boldsymbol{P} = [P_1,\dots,P_K]^T$. Let also $\mathcal{K}$ denote the (unobserved) set comprising the labels k_n of the distributions from which $\boldsymbol{x}_n$'s are drawn (the pair $(\mathcal{X},\mathcal{K})$ constitutes the complete data set associated with the problem at hand). Which of the following expressions is/are valid for the log-likelihood of the complete data set, $\ln p(\mathcal{X},\mathcal{K};\boldsymbol{\Xi},\boldsymbol{P})$?

1. $\sum_{n=1}^{N} \ln p\left(\boldsymbol{x}_n, k_n; \xi_{k_n}\right)$

2. $\sum_{n=1}^{N} \ln\left(p\left(\boldsymbol{x}_n|k_n;\boldsymbol{\xi}_{k_n}\right)P_{k_n}\right)$

3. $\sum_{n=1}^{N} \ln p\left(\boldsymbol{x}_n | k_n; \boldsymbol{\xi}_{k_n}\right)$

4. $\sum_{n=1}^{N} \ln\left(p\left(k_n | \boldsymbol{x}_n; \boldsymbol{\xi}_{k_n}\right) P_{k_n}\right)$

Exercise 8 (multiple choices question): Consider a machine learning task for which a set $\mathcal{X}$ of N l-dimensional i.i.d. observation vectors, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, is available. The aim is to estimate the probability density function (pdf) $p(\boldsymbol{x})$ that generates them. Assume that $p(\boldsymbol{x})$ is expressed as $p(\boldsymbol{x}) = \sum_{k=1}^{2} P_k p(\boldsymbol{x}|k; \boldsymbol{\xi}_k)$, where $p(\boldsymbol{x}|k; \boldsymbol{\xi}_k) = \mathcal{N}(\boldsymbol{x}|, {\sigma_k}^2 I)$ are Gaussian distributions, each one parameterized by $\boldsymbol{\mu}_k$ and ${\sigma_k}^2$, $k = 1{,}2$ (i.e., $\boldsymbol{\xi}_k = [{\boldsymbol{\mu}_k}^T, {\sigma_k}^2]^T$), and P_k is the probability that a sample has been drawn from $p(\boldsymbol{x}|k; \boldsymbol{\xi}_k)$. Let $\boldsymbol{\varXi} = \left[{\boldsymbol{\xi}_1}^T, {\boldsymbol{\xi}_2}^T\right]^T$ and $\boldsymbol{P} = [P_1, P_2]^T$. Let, also, $\mathcal{K}$ denote the (unobserved) set comprising the labels k_n of the distributions from which $\boldsymbol{x}_n$'s are drawn and $P(k|\boldsymbol{x}; \boldsymbol{\varXi}, \boldsymbol{P})$ be posterior probability of k, given $\boldsymbol{x}$, for $k = 1, 2$. Finally, it is given that half of the data points stem from the first mixture and the other half from the second one (but it is not known which of them stem from which pdf). Which of the following statements is/are true?

1. $P_1 \neq P_2$.

2. $P(1|\boldsymbol{x}; \boldsymbol{\varXi}, \boldsymbol{P}) = 1 - \frac{p(\boldsymbol{x}|2; \boldsymbol{\xi}_2) \cdot P_2}{p(\boldsymbol{x}; \boldsymbol{\varXi}, \boldsymbol{P})}$

3. If $\frac{\|\boldsymbol{x} - \boldsymbol{\mu}_1\|^2}{{\sigma_1}^2} = \frac{\|\boldsymbol{x} - \boldsymbol{\mu}_2\|^2}{{\sigma_2}^2}$, it always holds $P(1|\boldsymbol{x}; \boldsymbol{\varXi}, \boldsymbol{P}) = P(2|\boldsymbol{x}; \boldsymbol{\varXi}, \boldsymbol{P})$

4. if $p(\boldsymbol{x}|2; \boldsymbol{\xi}_2) \cdot P_2 > p(\boldsymbol{x}|1; \boldsymbol{\xi}_1) \cdot P_1$, then $p(\boldsymbol{x}|2; \boldsymbol{\xi}_2) \cdot P_2 > p(\boldsymbol{x}; \boldsymbol{\varXi}, \boldsymbol{P})$

Exercise 9 (multiple choices question): Consider a machine learning task for which a set $\mathcal{X}$ of N l-dimensional i.i.d. observation vectors, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, is available. The aim is to estimate the probability density function (pdf) $p(\boldsymbol{x})$ that generates them. Assume that $p(\boldsymbol{x})$ is expressed as $p(\boldsymbol{x}) = \sum_{k=1}^{K} P_k p(\boldsymbol{x}|k; \boldsymbol{\xi}_k)$, where $p(\boldsymbol{x}|k; \boldsymbol{\xi}_k) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, {\sigma_k}^2 I)$ are Gaussian distributions , each one parameterized by $\boldsymbol{\mu}_k$ and ${\sigma_k}^2$, $k = 1, \ldots, K$ (i.e., $\boldsymbol{\xi}_k = [{\boldsymbol{\mu}_k}^T, {\sigma_k}^2]^T$), and P_k is the probability that a sample has been drawn from $p(\boldsymbol{x}|k; \boldsymbol{\xi}_k)$. Let $\boldsymbol{\varXi} = \left[{\boldsymbol{\xi}_1}^T, \ldots, {\boldsymbol{\xi}_K}^T\right]^T$ and $\boldsymbol{P} = [P_1, \ldots, P_K]^T$. Let also $\mathcal{K}$ denote the (unobserved) set comprising the labels k_n of the distributions from which $\boldsymbol{x}_n$'s are drawn and $P(k|\boldsymbol{x}; \boldsymbol{\varXi}, \boldsymbol{P})$ be the posterior probability of k, given $\boldsymbol{x}$, for $k = 1, \ldots, K$. We employ the EM algorithm to estimate $\boldsymbol{\varXi}$ and $\boldsymbol{P}$. Which of the following pairs "quantity ⟷ (expectation or maximization) EM step" are compatible to each other?

1. P_k ⟷ Maximization step

2. P_k ⟷ Expectation step

3. $\boldsymbol{\mu}_k \leftrightarrow$ Expectation step

4. $\boldsymbol{\mu}_k \leftrightarrow$ Maximization step

5. $P(k|\boldsymbol{x};\boldsymbol{\Xi},\boldsymbol{P}) \leftrightarrow$ Expectation step

6. $P(k|\boldsymbol{x};\boldsymbol{\Xi},\boldsymbol{P}) \leftrightarrow$ Maximization step

7. ${\sigma_k}^2 \leftrightarrow$ Maximization step

8. ${\sigma_k}^2 \leftrightarrow$ Expectation step

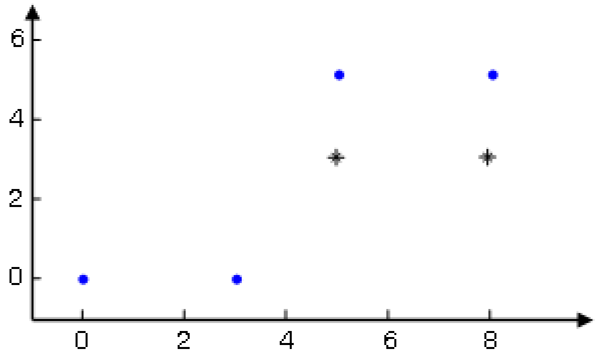
Exercise 10 (multiple choices question): Consider a machine learning task for which a set $\mathcal{X}$ of N l-dimensional i.i.d. observation vectors, $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$, is available. The aim is to estimate the probability density function (pdf) $p(\boldsymbol{x})$ that generates them. Assume that $p(\boldsymbol{x})$ is expressed as $p(\boldsymbol{x}) = \sum_{k=1}^{K} P_k p(\boldsymbol{x}|k;\boldsymbol{\xi}_k)$, where $p(\boldsymbol{x}|k;\boldsymbol{\xi}_k) = \mathcal{N}\left(\boldsymbol{x}\middle|\boldsymbol{\mu}_k, \sigma_k^2 I\right)$ are Gaussian distributions, each one parameterized by $\boldsymbol{\mu}_k$ and σ_k^2, $k = 1, \dots, K$ (i.e., $\boldsymbol{\xi}_k = \left[\boldsymbol{\mu}_k^T, \sigma_k^2\right]^T$), and P_k is the probability that a sample has been drawn from $p(\boldsymbol{x}|k;\boldsymbol{\xi}_k)$. Let $\boldsymbol{\Xi} = [\boldsymbol{\xi}_1^T, \dots, \boldsymbol{\xi}_K^T]^T$ and $\boldsymbol{P} = [P_1, \dots, P_K]^T$. Let $\gamma_{kn} := P(k|\boldsymbol{x}_n;\boldsymbol{\Xi},\boldsymbol{P})$ be the posterior probability of k, given $\boldsymbol{x}$, for $k = 1, \dots, K$, and $\gamma_{kn}^{(j)}, \boldsymbol{\mu}_k^{(j)}, \sigma_k^{2(j)}, P_k^{(j)}$, denote the values of the respective variables at the j-th iteration of the EM algorithm. Which of the following statements is/are true?

1. The value$\gamma_{kn}^{(j)}$ is computed based on $\boldsymbol{\mu}_k^{(j)}, \sigma_k^{2(j)}, P_k^{(j)}$.

2. $P_k^{(j)}$ is the mean of the posterior probabilities $\gamma_{kn}^{(j)}$.

3. $\boldsymbol{\mu}_k^{(j+1)}$ is the mean of all data points weighted by their associated $\gamma_{kn}^{(j)}$

4. $\sigma_k^{2(j+1)}$ is the mean of the squared Euclidean distances of all data points from $\boldsymbol{\mu}_k^{(j+1)}$ weighted by their associated $\gamma_{kn}^{(j)}$, scaled by the factor $\frac{1}{l}$.



Exercise 11 (multiple choices question): Consider a machine learning task for which the set $\mathcal{X}$ comprises the following four 2-dimensional i.i.d. observation vectors: $\boldsymbol{x}_1 = [0,\ 0]^T, \boldsymbol{x}_2 = [3,\ 0]^T$, $\boldsymbol{x}_3 = [5,\ 5]^T$, $\boldsymbol{x}_4 = [8,\ 5]^T$. The aim is to estimate the probability density function (pdf) $p(\boldsymbol{x})$ that generates them. Assume that $p(\boldsymbol{x})$ is expressed as $p(\boldsymbol{x}) = \sum_{k=1}^{2} P_k p(\boldsymbol{x}|k;\boldsymbol{\xi}_k)$, where $p(\boldsymbol{x}|k;\boldsymbol{\xi}_k) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, I)$ are Gaussian distributions with covariance matrices equal to I (the 2×2 identity matrix), each one parameterized by $\boldsymbol{\mu}_k$, $k = 1, 2$ (in this case it is $\boldsymbol{\xi}_k := \boldsymbol{\mu}_k$), and P_k is the probability that a sample has been drawn from $p(\boldsymbol{x}|k; \boldsymbol{\xi}_k)$. Let $\boldsymbol{\Xi} = \left[{\boldsymbol{\xi}_1}^T, {\boldsymbol{\xi}_2}^T\right]^T$ and $\boldsymbol{P} = [P_1, P_2]^T$. Let $\gamma_{kn} := P(k|\boldsymbol{x}_n; \boldsymbol{\Xi}, \boldsymbol{P})$ be the posterior probability of k, given $\boldsymbol{x}$, for $k = 1, 2$, and ${\gamma_{kn}}^{(j)}, {\boldsymbol{\mu}_k}^{(j)}, {P_k}^{(j)}$, denote the values of the respective variables at the j-th iteration of the EM algorithm. Assume that the initial estimates of $\boldsymbol{\mu}_k$ and P_k , $k = 1,2$. Let ${\boldsymbol{\mu}_1}^{(0)} = [5,\ 3]^T, {\boldsymbol{\mu}_2}^{(0)} = [8,\ 3]^T$, ${P_1}^{(0)} = 0.1,\ {P_2}^{(0)} = 0.9$. The estimates of ${\gamma_{kn}}^{(0)}$, ${\boldsymbol{\mu}_k}^{(1)}, {P_k}^{(1)}$, $k = 1,2$, $n = 1, \ldots, 4$, are (precision in one decimal)

1. $\begin{bmatrix} {\gamma_{11}}^{(0)} = 0.9 & {\gamma_{21}}^{(0)} = 0.1 \\ {\gamma_{21}}^{(0)} = 0.0 & {\gamma_{22}}^{(0)} = 1.0 \\ {\gamma_{31}}^{(0)} = 1.0 & {\gamma_{32}}^{(0)} = 0.0 \\ {\gamma_{41}}^{(0)} = 1.0 & {\gamma_{42}}^{(0)} = 0.0 \end{bmatrix}, {\boldsymbol{\mu}_1}^{(1)} = \begin{bmatrix} 2.6 \\ 1.6 \end{bmatrix}, {\boldsymbol{\mu}_2}^{(1)} = \begin{bmatrix} 7.7 \\ 5.0 \end{bmatrix}, {P_1}^{(1)} = 0.3,\ {P_2}^{(1)} = 0.7$

2. $\begin{bmatrix} {\gamma_{11}}^{(0)} = 1.0 & {\gamma_{21}}^{(0)} = 0.0 \\ {\gamma_{21}}^{(0)} = 1.0 & {\gamma_{22}}^{(0)} = 0.0 \\ {\gamma_{31}}^{(0)} = 0.9 & {\gamma_{32}}^{(0)} = 0.1 \\ {\gamma_{41}}^{(0)} = 0.0 & {\gamma_{42}}^{(0)} = 1.0 \end{bmatrix}, {\boldsymbol{\mu}_1}^{(1)} = \begin{bmatrix} 5.0 \\ 3.0 \end{bmatrix}, {\boldsymbol{\mu}_2}^{(1)} = \begin{bmatrix} 8.0 \\ 3.0 \end{bmatrix}, {P_1}^{(1)} = 0.7,\ {P_2}^{(1)} = 0.3$

3. $\begin{bmatrix} {\gamma_{11}}^{(0)} = 1.0 & {\gamma_{21}}^{(0)} = 0.0 \\ {\gamma_{21}}^{(0)} = 1.0 & {\gamma_{22}}^{(0)} = 0.0 \\ {\gamma_{31}}^{(0)} = 0.9 & {\gamma_{32}}^{(0)} = 0.1 \\ {\gamma_{41}}^{(0)} = 0.0 & {\gamma_{42}}^{(0)} = 1.0 \end{bmatrix}, {\boldsymbol{\mu}_1}^{(1)} = \begin{bmatrix} 2.6 \\ 1.6 \end{bmatrix}, {\boldsymbol{\mu}_2}^{(1)} = \begin{bmatrix} 7.7 \\ 5.0 \end{bmatrix}, {P_1}^{(1)} = 0.5,\ {P_2}^{(1)} = 0.5$

4. $\begin{bmatrix} {\gamma_{11}}^{(0)} = 1.0 & {\gamma_{21}}^{(0)} = 0.0 \\ {\gamma_{21}}^{(0)} = 1.0 & {\gamma_{22}}^{(0)} = 0.0 \\ {\gamma_{31}}^{(0)} = 0.9 & {\gamma_{32}}^{(0)} = 0.1 \\ {\gamma_{41}}^{(0)} = 0.0 & {\gamma_{42}}^{(0)} = 1.0 \end{bmatrix}, {\boldsymbol{\mu}_1}^{(1)} = \begin{bmatrix} 2.6 \\ 1.6 \end{bmatrix}, {\boldsymbol{\mu}_2}^{(1)} = \begin{bmatrix} 7.7 \\ 5.0 \end{bmatrix}, {P_1}^{(1)} = 0.7,\ {P_2}^{(1)} = 0.3$

Exercise 12 (multiple choices question): Consider a machine learning task for which a set $\mathcal{X}$ of N l-dimensional i.i.d. observation vectors, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, is available. The aim is to estimate the probability density function (pdf), $p(\boldsymbol{x})$, that generates them. Assume that $p(\boldsymbol{x})$ is expressed as $p(\boldsymbol{x}) = \sum_{k=1}^{K} P_k p(\boldsymbol{x}|k; \boldsymbol{\xi}_k)$, where $p(\boldsymbol{x}|k; \boldsymbol{\xi}_k) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k; \Sigma_k)$ are Gaussian distributions, each one parameterized by the mean vector $\boldsymbol{\mu}_k$ and the covariance matrix Σ_k, $k = 1, \ldots, K$ (in this case, $\boldsymbol{\xi}_k$ comprises $\boldsymbol{\mu}_k$ and (the entries of) Σ_k), and P_k is the probability that a sample has been drawn from $p(\boldsymbol{x}|k; \boldsymbol{\xi}_k)$. Let $\boldsymbol{\Xi} = \left[{\boldsymbol{\xi}_1}^T, \ldots, {\boldsymbol{\xi}_K}^T\right]^T$ and $\boldsymbol{P} = [P_1, \ldots, P_K]^T$. Let $\gamma_{kn} := P(k|\boldsymbol{x}_n; \boldsymbol{\Xi}, \boldsymbol{P})$ be the posterior probability of k, given $\boldsymbol{x}$, for $k = 1, \ldots, K$, and ${\gamma_{kn}}^{(j)}, {\boldsymbol{\mu}_k}^{(j)}, {\Sigma_k}^{(j)}, {P_k}^{(j)}$, denote the values of the respective variables at the j-th iteration of the EM algorithm. Which of the following statements is/are true?

1. The EM algorithm can only handle normal pdfs with isotropic covariance matrices ($\Sigma_k = \sigma_k{}^2 I$).

2. It is $\Sigma_k{}^{(j+1)} = \frac{\sum_{n=1}^{N} \gamma_{kn}{}^{(j)} (\boldsymbol{x}_n - \boldsymbol{\mu}_k{}^{(j+1)})(\boldsymbol{x}_n - \boldsymbol{\mu}_k{}^{(j+1)})^T}{\sum_{n=1}^{N} \gamma_{kn}{}^{(j)}}$

3. The EM algorithm can always handle successfully the cases where the estimate of K is different from the actual number of the pdfs comprising $p(\boldsymbol{x})$.

4. Only the data points for which $p(\boldsymbol{x} | k; \boldsymbol{\xi}_k{}^{(j)}) = \max_{m=1,\dots,K} p(\boldsymbol{x} | m; \boldsymbol{\xi}_m{}^{(j)})$ contribute to the estimate of $\boldsymbol{\mu}_k{}^{(j)}, \Sigma_k{}^{(j)}, P_k{}^{(j)}$.

Exercise 13:

Consider the Rayleigh distribution $p(x;\theta) = 2\,\theta\ x\, exp(-\theta\, x^2) u(x)$, (where $u(x) = 1(0)$, if $x \geq 0$ (< 0)).

Given

- a set of N measurements $x_1, \dots, x_N$, for the random variable x that follows the Rayleigh distribution, and
- the a priori probability for the parameter θ, which is a normal distribution, $N(\theta_0, \sigma_0{}^2)$ (where θ_0, $\sigma_0{}^2$ are known),

(a) Compute the MAP estimate of the parameter θ.

(b) How this estimate becomes for the case were (i) $N \rightarrow \infty$, (ii) $\sigma_0{}^2 \gg$ and (c) $\sigma_0{}^2 \ll$? Give a short justification.

(c) Compare the MAP estimate of θ with the ML estimate from exercise 1.

(d) For $N = 5$ and $x_1 = 2$, $x_2 = 2.2$, $x_3 = 2.7$, $x_4 = 2.4$, $x_5 = 2.6$, $\theta_0 = 1.8$ and $\sigma_0{}^2 = 1$ estimate the θ_{MAP}. Utilizing this estimate, determine $\hat{p}(x)$, for $x = 2.3$ and $x = 2.9$. Compare the results with those obtained in exercise 4 of Homework 5, where the ML estimate where considered.

Exercise 14:

Consider a data set $Y = \{x_1, \dots, x_N\}$, whose elements have been drawn independently from the exponential distribution

$$p(x|\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

The parameter λ of the distribution is modelled by a prior gamma distribution, i.e.,

$$p(\lambda) \equiv p(\lambda; a, b) = \begin{cases} \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}, & \lambda \geq 0 \\ 0, & \lambda < 0 \end{cases}$$

(a) **Determine** the likelihood $p(Y|\lambda)$.
(b) **Form** the product of the prior and the likelihood and determine the MAP estimate of λ.
(c) **Give** the form of $p(x)$, in terms of the MAP estimate of λ, determined in (b).
(d) **Determine** the posterior distribution for λ, $p(\lambda|\, Y)$, in the light of the Y.
(e) **Compare** the form of the resulting posterior $p(\lambda|\, Y)$ with that of the prior $p(\lambda)$ of λ and comment briefly.
(f) **Prove** that $p(x|Y)$ is a lomax distribution.

For the following, assume that $Y = \{2.8, 2.4, 2.9, 2.6, 2.1, 2.2\}$, $a = 2$ and $b = 2$.

(g) **Write down** the $p(x)$ of (c) for the above Y.
(h) **Write down** the $p(x|Y)$ of (f) for the above Y.
(i) **Compute** $p(x)$ (from (g)) and $p(x|Y)$ (from (h)), for $x = 2.5$.

Hints: (i) For (a) and (b) work as in exercise 13.

(ii) For (d): A pdf of the form $C\lambda^r e^{-s\lambda}$ is a gamma distribution with parameters r and s.

(iii) For (f): (I) It is $\int_0^\infty t^b e^{-at} dt = \frac{\Gamma(b+1)}{a^{b+1}}$ and (II) $\Gamma(z+1) = z\Gamma(z)$. (III) The lomax distribution is defined as $p(x; c, d) = \frac{cd^c}{(x+d)^{c+1}}$, for $x \geq 0$ and 0 otherwise. (IV) In order to completely determine $p(x|Y)$, the values c, d of the distribution need to be determined.

Exercise 15:

Consider the model $x = \theta + \eta$ ($x, \theta, \eta \in R$) and a set of measurements $Y = \{x_1, \dots, x_N\}$, which are noisy versions of θ. Assume that we have prior knowledge about θ saying that it lies close to θ_0. Formulating the ridge regression problem for this case as follows

$min_\theta\, J(\theta) = \sum_{n=1}^N (x_n - \theta)^2$, subject to $(\theta - \theta_0)^2 \leq \rho$

Prove that

$$\theta_{RR} = \frac{\sum_{n=1}^N x_n + \lambda\theta_0}{N + \lambda}$$

where λ is a user defined parameter.

Hint: Define the Lagrangian function $L(\theta) = \sum_{n=1}^N (x_n - \theta)^2 + \lambda((\theta - \theta_0)^2 - \rho)$ (λ is the Lagrange multiplier corresponding to the constraint).

Exercise 16:

Consider the case where the data at hand are modeled by a pdf of the form

$$p(\boldsymbol{x}) = \sum\nolimits_{j=1}^{m} P_j p(\boldsymbol{x} \mid j), \quad \sum_{j=1}^{m} P_j = 1, \quad \int_{-\infty}^{+\infty} p(\boldsymbol{x} \mid j) = 1$$

where $m = 3$ and $P_j, j = 1,2,3$, are the a priori probabilities of the pdfs $p(\boldsymbol{x}|j)$, which involved in the definition of $p(\boldsymbol{x})$. In the "parameter updating" part of the EM-algorithm, which allows the estimation of the parameters of $p(\boldsymbol{x}|j)$'s as well as P_j's, we need to solve the problem

$$[P_1, P_2, P_3] = argmax_{[P_1,P_2,P_3]} \sum_{i=1}^{N} \sum_{j=1}^{3} P(j|\boldsymbol{x_i}) \ln P_j\, ,\ subject\ to\ \sum_{j=1}^{3} P_j = 1,$$

for fixed $P(j|\boldsymbol{x_i})$'s. Prove that, independently of the form adopted for each $p(\boldsymbol{x}|j)$, the solution of the above problem is

$$P_j = \frac{1}{N} \sum_{i=1}^{N} P(j|\boldsymbol{x_i})\, ,\ j = 1,2,3.$$

Hint: In this case we have an equality constraint. Work as follows:

1. Define the Lagrangian function
   $L(P_1, P_2, P_3) = \sum_{i=1}^{N} \sum_{j=1}^{3} P(j|\boldsymbol{x_i}) \ln P_j + \lambda(\sum_{j=1}^{3} P_j - 1),$
2. Solve the equations $\frac{\partial L(P_1,P_2,P_3)}{\partial P_j} = 0, j = 1,2,3$, expressing each P_j in terms of λ.
3. Substitute P_j's in the constraint equation $\sum_{j=1}^{3} P_j = 1$ and solve with respect to λ.
4. Compute P_j's from the equations derived in step 2 above.

Notes:

- In the case of equality constraints, the final solution **does not** involve the Lagrangian multipliers.
- The extension to the general case of m individual pdfs is straightforward.

Exercise 17:

Consider again the setup of exercise 16, where now $p(\boldsymbol{x}|j)$'s are normal distributions with means $\boldsymbol{\mu}_j$ and **fixed** covariance matrices $\Sigma_j, j = 1, \dots, m$. Prove that the solution of the optimization problems

$$\boldsymbol{\mu}_j = argmax_{\boldsymbol{\mu}_j} \sum\nolimits_{i=1}^{N} P(j|\boldsymbol{x_i}) \ln\left(p\left(\boldsymbol{x}_i|j;\boldsymbol{\mu}_j\right)\right),\ j = 1,\dots,m$$ [1]

is

$$\boldsymbol{\mu}_j = \frac{\sum_{i=1}^{N} P(j|\boldsymbol{x_i})\boldsymbol{x_i}}{\sum_{i=1}^{N} P(j|\boldsymbol{x_i})}, j = 1,\dots,m$$

Hint: Take the gradient of $\sum_{i=1}^{N} P(j|\boldsymbol{x_i}) \ln\left(p\left(\boldsymbol{x}_i|j;\boldsymbol{\mu}_j\right)\right)$ with respect to $\boldsymbol{\mu}_j$, set it equal to zero and solve for $\boldsymbol{\mu}_j$.

Exercise 18 (**python code**):

Consider the two data sets $X_1$and X_2 contained in the attached file "Dataset.mat", each one of them containing 4-dimensional data vectors, in its rows. The vectors of X_1 stem from the pdf $p_1(\boldsymbol{x})$, while those of X_2 stem from the pdf $p_2(\boldsymbol{x})$.

(a) Based on X_1, estimate the values of $p_1(\boldsymbol{x})$ at the following points:
$\boldsymbol{x}_1 = (2.01, 2.99,\ 3.98,\ 5.02)$, $\boldsymbol{x}_2 = (20.78,\ -15.26,\ 19.38,\ -25.02)$, $\boldsymbol{x}_3 = (3.08, 3.88,\ 4.15,\ 6.02)$.

(b) Based on X_2, estimate the values of $p_2(\boldsymbol{x})$ at the following points:
$\boldsymbol{x}_1 = (0.05, 0.15, -0.12, -0.08)$, $\boldsymbol{x}_2 = (7.18,\ 7.98,\ 9.12, 9.94)$, $\boldsymbol{x}_3 = (3.48, 4.01,\ 4.55,\ 4.96)$, $\boldsymbol{x}_4 = (20.78,\ -15.26,\ 19.38,\ -25.02)$.

In both the above cases use a parametric approach.

Hints:

- To load the data sets use the script "HW6.ipynb".
- Use the Sklearn.mixture.GaussianMixture class (https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html), if you are willing to use Gaussian mixtures modelling.

It could be proved useful for the modelling of each pdf to compute the mean of each data set and then to consider the distances of the data vectors from it. However, other methods can also be applied.

[1] Note that, since Σ_j is fixed, it is $\boldsymbol{\theta}_j \equiv \boldsymbol{\mu}_j$.