NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS
MACHINE LEARNING PROGRESS EXAMINATION, DECEMBER 2023

Instructions:

- Time allocated: **180 minutes.**
- Please write your name and your serial number (if you have one).
- Solve 7 out of the 8 numbered problems. Solution of the bonus problem gives additional marks.
- After you finish answering the questions, it would be helpful if you could use a multi-page scanner app of your choice to take photos of the pages of your paper with your mobile phone and organize all pages in a single pdf file. You should then submit this file in the slot "Mid-term Exam December 2023 submissions" under "Assignments" on eclass.

1. Explain very briefly
   a. what the main difference is between the Bayes classifier and the Naïve Bayes classifier
   b. what the main advantage is of Ridge Regression over the Least Squares method
   c. what are latent variables in the Expectation-Maximization method
   d. what is meant by a machine learning model acting as an Associative Memory.

2. Which of the following statements are true?
   a. In a Gaussian mixture model, the preferred method of training is the Expectation-Maximization method
   b. It is impossible for a biased estimator to perform better than an unbiased estimator
   c. The joint probability of a Bayesian network is equal to the sum of the (conditional) probabilities associated to its nodes
   d. In a generalized linear regression task, Bayesian inference can provide us with both a value and an error for each of our test set estimates
   e. In a Hopfield model the number of synaptic weights grows linearly with the number of neurons
   f. In a generalized linear regression task, Least Squares and Maximum Likelihood will give identical results under white Gaussian noise
   g. In the perceptron algorithm, the weight vector is parallel to the separating hyperplane.

3. Patterns in two equiprobable classes ω_1,ω_2 originate from distributions with the following probability density functions:

$$\text{Class } \omega_1\text{:} \quad p(x|\omega_1) = \begin{cases} 3\exp(-3x), x > 0 \\ 0, x \leq 0 \end{cases}$$

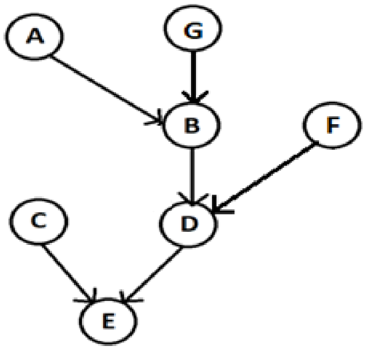
$$\text{Class } \omega_2\text{:} \quad p(x|\omega_2) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(x-3)^2}{2}\right]$$

   a. Find the intervals on the *x* axis assigned to each class by a Bayes classifier.
   b. Calculate the probability of error corresponding to patterns belonging to class ω_1 being wrongly classified.
   c. We are more interested in correctly classifying patterns belonging in class ω_1 than patterns belonging in class ω_2. Therefore, we introduce risk factors $\lambda_{12} = \frac{8}{3}, \lambda_{21} = 1$. Taking into account the risk, find the modified intervals assigned to each class.
   d. Recalculate the probability of error corresponding to patterns belonging to class ω_1 being misclassified and verify that it has indeed decreased compared to its previous value.

You can use the approximations: $\ln(3) + \ln(\sqrt{2\pi}) \cong 2, \ln(8) + \ln(\sqrt{2\pi}) \cong 3$. It may help to plot the probability density functions in a common graph, e.g. using https://www.desmos.com/ calculator.

4. We wish to classify the following 2-D patterns: $\begin{bmatrix}1\\3\end{bmatrix}$ and $\begin{bmatrix}3\\4\end{bmatrix}$ to a class ω₁ (target +1) and: $\begin{bmatrix}3\\1\end{bmatrix}$ to another class ω₂ (target -1) using the perceptron algorithm.
   a. Starting from initial weight vector $\begin{bmatrix}0\\0\end{bmatrix}$ and bias 0, perform **one** epoch of the perceptron algorithm (incremental, ε=1) and tabulate your results.
   b. Give the equation of the decision line that you found and draw it.
   c. Draw the final weight vector.
   d. Has the algorithm classified the patterns successfully?

5. In the Bayesian network shown below, all variables are binary. For example, variable A can take the values A_1 (A is true) and A_0 (A is false).



   a. For each of the nodes A,B,C,D,E,F,G which are the (conditional) probabilities associated with the network that you should know in order to be able to perform inference?
   b. Which formula gives the joint probability *P*(A,B,C,D,E,F,G) for this network?
   c. Inference: Find $P(E_1|B_1)$ in terms of the (conditional) probabilities of question (a).

6. Consider the linear regression problem:

$$y = \theta x + \theta_0 + \eta$$

We have the following samples in our training set: $\begin{bmatrix}x_1\\y_1\end{bmatrix} = \begin{bmatrix}-1\\3\end{bmatrix}$ and $\begin{bmatrix}x_2\\y_2\end{bmatrix} = \begin{bmatrix}1\\1\end{bmatrix}$. Find θ and θ_0 using Ridge Regression with regularization parameter λ. What would the result be if the least squares method was used instead? What happens when λ tends to infinity? Comment on the result.

7. Two-dimensional patterns from two equiprobable classes ω₁ and ω₂ originate from gaussian distributions with means $\boldsymbol{\mu_1} = \begin{bmatrix}0\\0\end{bmatrix}$ and $\boldsymbol{\mu_2} = \begin{bmatrix}4\\4\end{bmatrix}$ respectively and common covariance matrix $\Sigma = \begin{bmatrix}4 & 0\\0 & 1\end{bmatrix}$. We want to classify pattern $\boldsymbol{x} = \begin{bmatrix}4\\3\end{bmatrix}$ using a Bayes classifier.

Draw a diagram to show qualitatively the isolevel curves for the two classes passing through point $\boldsymbol{x}$. From your diagram, or otherwise, decide to which class you would classify $\boldsymbol{x}$.

8. In a parameter estimation task, the parameter *θ* is estimated using three different estimation algorithms. For each of the three algorithms, multiple training sets are used and it is found that the distribution of the results over the different training sets can be approximated as follows:

Algorithm 1	Algorithm 2	Algorithm 3
$f_1(\theta) = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-(\theta-5)^2}{2}\right]$	$f_2(\theta) = \frac{1}{\sqrt{6\pi}} \exp\left[\frac{-(\theta-6)^2}{6}\right]$	$f_3(\theta) = \frac{1}{2\sqrt{\pi}} \exp\left[\frac{-(\theta-8)^2}{4}\right]$

The true value of θ is $\theta_0 = 6$. Which of the three estimation algorithms is preferable in terms of mean square error? Which algorithm leads to the most biased result? Which to the least biased? Comment on whether introducing bias helped in this situation and under what circumstances.

Bonus problem:

The purpose of this problem is to show once again that analysis of probability density functions using Gaussians can facilitate finding solutions to Machine Learning problems. We have seen this in class mainly with classification problems. Here, on the other hand, we have a regression problem: We wish to estimate *y* versus *x*, given our training set. We model the joint probability density function of our variables using a mixture of *M* Gaussians. We thus run the Expectation-Maximization algorithm and obtain our joint PDF in the form:

$$p(x,y) = \sum_{i=1}^{M} C_i \exp\left[\frac{-(x-X_i)^2 - (y-Y_i)^2}{2\sigma_i{}^2}\right]$$

where the number *M* of Gaussians, the coefficients C_i, the means X_i,Y_i and the standard deviations σ_i are known.

a. Calculate the marginal probability $p(x) = \int_{-\infty}^{\infty} p(x,y)dy$ and the conditional probability $p(y|x)$.
b. Derive an analytical formula for the MSE optimal estimate $y = g(x)$.

You can use the integrals: $\int_{-\infty}^{\infty} \exp[-a(y-b)^2]\, dy = \left(\frac{\pi}{\alpha}\right)^{\frac{1}{2}}$and $\int_{-\infty}^{\infty} y \exp[-a(y-b)^2]\, dy = b\left(\frac{\pi}{\alpha}\right)^{1/2}$.

END OF EXAM