NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS
MACHINE LEARNING
PROGRESS EXAMINATION, DECEMBER 2022

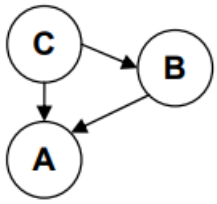
Instructions:

- Time allocated: **120 minutes.**
- Please write your name and your serial number.
- In the multiple-choice questions 1-9, identify which of the choices are true statements. In some of the questions, there is more than one correct answer.
- When you finish, scan or take a photo of your paper and send it by replying to the same e-mail message by which you received the questions.

1. In a Gaussian mixture model:
   a. Our objective is to estimate the weighting factors as well as the means and covariance matrices of the Gaussians
   b. The preferred method for training the model is the least squares method
   c. We seek to perform a classification task
   d. The preferred method for training the model is the Expectation-Maximization method

2. Which of the following are true?
   a. In a Hopfield model the number of synaptic weights grows linearly with the number of neurons
   b. A Hopfield model is used mainly for classification
   c. A Hopfield model is used mainly for pattern retrieval
   d. The joint probability of a Bayesian network is equal to the product of the (conditional) probabilities associated to its nodes
   e. It is impossible for a biased estimator to perform better than an unbiased estimator
   f. 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

3. In the depicted Bayesian network, the joint probability of all variables is given by the formula:
   a. P(A) P(B) P(C)
   b. P(A|B,C) P(C|B) P(B)
   c. P(B|C,A) P(C) P(B) P(A)
   d. P(A|B) P(A|C) P(C|B)
   e. P(A|B,C) P(C|B)
   f. None of the above



4. In a simple one-layered perceptron with weight vector $\boldsymbol{w}$ and bias w_0:
   a. $\boldsymbol{w}$ is parallel to the separating hyperplane in the pattern space
   b. $\boldsymbol{w}$ is perpendicular to the separating hyperplane in the pattern space
   c. w_0 is always positive
   d. Multiplying w_0 as well as all components of $\boldsymbol{w}$ by a common factor ρ changes classification results

5. When we estimate the probability density function of a distribution based on samples drawn from the distribution, the result of the Maximum a Posteriori Probability (MAP) method tends to approach the result of the Maximum Likelihood method when:
   a. The number of patterns in the training set is very large
   b. There is little uncertainty in our apriori estimate for the MAP parameter
   c. Our a priori estimate for the parameter in MAP has a large standard deviation
   d. There are too few patterns in the training set
   e. The dimensionality of our problem is large

6. Consider a generalized regression problem where data points are generated by a 2nd degree polynomial. We employ a 5th degree polynomial to perform regression. This model is likely to result in:
    a. Small variance, large bias
    b. Small variance, small bias
    c. Large variance, large bias
    d. Large variance, small bias

7. We are given a set of 400 pairs (x_i,y_i), i=1,…,400 and we seek to perform generalized linear regression using the ridge regression method. In truth, data are generated by a 5th degree polynomial in x with added noise. We employ 10-fold cross validation and use our well-known formula for estimating the parameter vector $\boldsymbol{\theta}$:

$$\widehat{\boldsymbol{\theta}} = \left(\Phi^{\mathrm{T}}\Phi + \lambda \mathrm{I}\right)^{-1}\Phi^{\mathrm{T}}\boldsymbol{y}$$

    In the above formula, the number of rows and columns in matrix Φ and the number of rows in the unit matrix I are, respectively:
    a. 400,5,5
    b. 400,400,6
    c. 5,360,5
    d. 360,6,6
    e. 360,5,5
    f. 6,6,6
    g. None of the above

8. A potential advantage of the ridge regression method over the least squares method is:
    e. Ridge regression eliminates bias from the estimation
    f. We can avoid overfitting using a regularization term
    g. Prior knowledge related to the specific problem is utilized
    h. There are less parameters to be estimated
    i. Ridge regression produces a smoother fitting curve

9. In a linear regression task, the Maximum Likelihood method can give different results from the Least Squares method when
    a. The number of dimensions is smaller than the number of data in the training set
    b. The noise is white
    c. There are too few instances in the training set
    d. The noise is colored (not white)
    e. There are outliers in our data

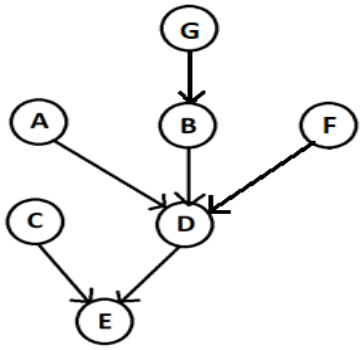
10. Explain very briefly why it is possible in a real-world classification problem for the Bayes classifier to perform more poorly than other classifiers, despite its theoretical optimality.

11. The two-dimensional patterns from two equiprobable classes ω_1 and ω_2 originate from gaussian distributions with means $\boldsymbol{\mu_1} = \begin{bmatrix}0\\0\end{bmatrix}$ and $\boldsymbol{\mu_2} = \begin{bmatrix}3\\3\end{bmatrix}$ respectively and common covariance matrix Σ. The eigenvectors of Σ are $\begin{bmatrix}1\\0\end{bmatrix}$ and $\begin{bmatrix}0\\1\end{bmatrix}$ with corresponding eigenvalues 2 and 1.

    a. Comment on the truth or falsehood of the following statement: In the described situation, the Bayes classifier coincides with the naïve Bayes classifier.

    b. We want to classify pattern $\boldsymbol{x} = \begin{bmatrix}1\\2\end{bmatrix}$ using a Bayes classifier. Draw a diagram to show qualitatively the isolevel curves for the two classes, that pass through point x. From your diagram, decide to which class you would classify x.

12. 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).

13. 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
$f_1(\theta) = \frac{1}{\sqrt{2\pi}}\exp\left[\frac{-(\theta-4)^2}{2}\right]$	$f_2(\theta) = \frac{1}{\sqrt{6\pi}}\exp\left[\frac{-(\theta-5)^2}{6}\right]$	$f_3(\theta) = \frac{1}{2\sqrt{\pi}}\exp\left[\frac{-(\theta-7)^2}{4}\right]$

    The true value of θ is $\theta_0 = 5$. 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.

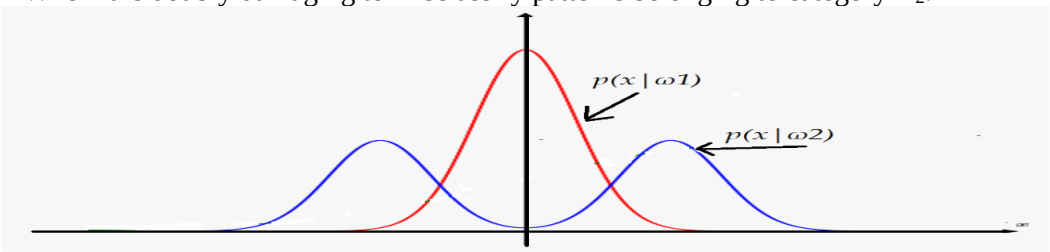
14. Consider the following probability density function:

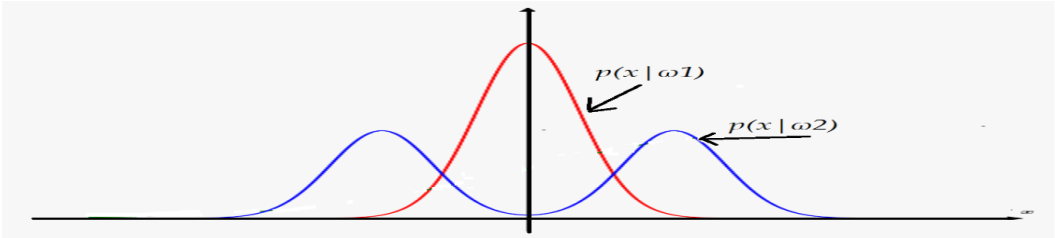
$$p(x) = \frac{\theta \mathrm{e}^{-x}}{\left(1+e^{-x}\right)^{(\theta+1)}}, \quad \theta \geq 0$$

    ,

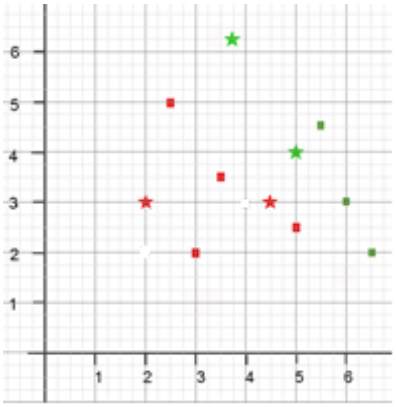
    We are given N random samples x_1, x_2, ..., x_N drawn from this distribution. Calculate the maximum likelihood estimate for the parameter θ in terms of N and the samples x_i, i=1,2,…,N.

15. Probability density functions are shown for 2 equiprobable categories ω_1 and ω_2. Show which intervals of the x axis are classified by a Bayes classifier to each of the two categories:
    a. When the two categories are equally important
    b. When it is doubly damaging to misclassify patterns belonging to category ω_2.





16. In the following diagram you can see a classification problem with two classes. Square dots belong to the training set, while asterisks belong to the test set. All red symbols belong in reality to class 1, while all green symbols belong to class 2.



Calculate the weights and bias after performing **one** epoch of the perceptron algorithm (incremental mode) on the training set starting from zero initial weights and bias. Draw the separating line and the weight vector on the diagram. Which patterns in the test set are correctly classified following this one epoch of training?

17. Consider the linear regression problem:

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

We have the following samples in our training set: (x_1,y_1)=(-1,2) and (x_2,y_2)=(1,1). Find θ and θ_0 using rigde regression with regularization parameter λ. What happens when λ tends to infinity? Comment on the result.

END OF EXAM