NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS
MACHINE LEARNING
PROGRESS EXAMINATION, DECEMBER 2021

Instructions:

- Time allocated: **100 minutes.**
- Please write your name and your serial number (if you have one).
  In a few of the multiple choice 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. A potential advantage of the ridge regression method over the least squares method is:
   a. Ridge regression eliminates bias from the estimation
   b. We can avoid overfitting using a regularization term
   c. Prior knowledge related to the specific problem is utilized
   d. There are less parameters to be estimated
   e. Ridge regression produces a smoother fitting curve

2. Identify which of the following are true and which are false:
   a. In a Hopfield model the number of synaptic weights grows linearly with the number of neurons
   b. The capacity of the Hopfield model grows linearly with the number of neurons
   c. The joint probability of a Bayesian network is equal to the sum of the (conditional) probabilities associated to its nodes
   d. It is possible for a biased estimator to perform better than an unbiased estimator
   e. For a k-nearest neighbor classifier used to classify instances into two classes, it is advisable that k be even.

3. 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 large
   b. The noise is white
   c. There are too many instances in the training set
   d. There are outliers in our data
   e. The noise is not white

4. 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 approaches infinity
   b. Our a priori estimate for the parameter in MAP has a large standard deviation
   c. Our a priori estimate for the parameter in MAP has a small standard deviation
   d. There are too few patterns in the training set

5. Consider a generalized regression problem where data points are generated by a 5th degree polynomial. Characterize as small or large the bias and variance of the estimates produced by the following models:

Model	**Bias**	**Variance**
Linear	Small/Large	Small/Large
5th degree polynomial	Small/Large	Small/Large
30th degree polynomial	Small/Large	Small/Large

6. 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 in matrix Φ is:

a. 6
b. 400
c. 7
d. 360
e. 8
f. 380

7. In the previous problem, the number of columns in the unit matrix I is:

a. 6
b. 400
c. 7
d. 360
e. 8
f. 380

8. In the previous problem, the number of columns in matrix Φ is:

a. 6
b. 400
c. 7
d. 360
e. 8
f. 380

9. 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.

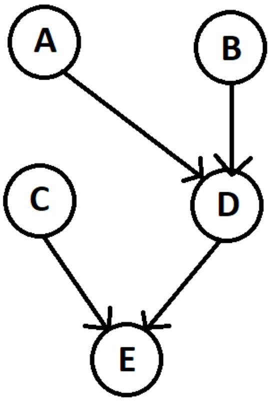
10. We wish to classify the following 2-D patterns: $\begin{bmatrix}1\\3\end{bmatrix}$ and $\begin{bmatrix}2\\4\end{bmatrix}$ to a class ω₁ (target +1) and $\begin{bmatrix}0\\1\end{bmatrix}$ to another class ω₂ (target -1) using the perceptron algorithm. Starting from zero initial weights and bias, perform one epoch of the perceptron algorithm (incremental, ε=1) and tabulate your results. Give the equation of the decision line that you found. Which patterns in the training set are correctly classified and which wrongly? Find the distance of pattern $\begin{bmatrix}0\\1\end{bmatrix}$ from the decision line.

11. 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 $\Sigma = \begin{bmatrix}3 & 0\\0 & 1\end{bmatrix}$. 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, that pass through point $\boldsymbol{x}$. From your diagram, or otherwise, decide to which class you would classify $\boldsymbol{x}$.

12. 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, but now the corresponding covariance matrices for the two distributions are $\Sigma_1 = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$ and $\Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}$.
    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. Compute the equation of the separating curve of the two classes according to the Bayes classifier.
    c. Classify pattern $\boldsymbol{x} = \begin{bmatrix} 1/2 \\ 0 \end{bmatrix}$.

13. 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 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) for this network?
    c. Inference: Find $P(E_1|A_1)$ in terms of the (conditional) probabilities of question (a).



14. The one-dimensional exponential distribution has the following probability density function:

$$f(x) = \theta \exp(-\theta\, x), \quad x > 0$$

You are given *N* random samples $X_1, X_2, \ldots, X_N$ originating from this distribution. Estimate the parameter θ by the maximum likelihood method.

END OF EXAM