NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS
MACHINE LEARNING
FINAL EXAMINATION, FEBRUARY 2021

Instructions:

- Please write your name and your serial number.
- If you have passed the intermediate exam and want to keep its grade, answer only questions in Group B. **You have 1 hour and 45 minutes**.
- If you have not taken, or have not passed, or do not want to keep the grade of the intermediate exam, answer questions in both Group A and Group B. **You have 2 hours and 15 minutes.**
- In some 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. Please try to group all pages into a single pdf file if possible.

GROUP A

1. Explain very briefly
   a. 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
   b. what is the main difference between the Bayes classifier and the naïve Bayes classifier

2. Explain very briefly
   a. Under what circumstances you would use MAP instead of maximum likelihood in a distribution estimation problem
   b. What is the main advantage of ridge regression over the least squares method

3. 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 ω$_1$ (target +1) and : $\begin{bmatrix}0\\1\end{bmatrix}$ to another class ω$_2$ (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.

4. 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.

GROUP B

1. A multilayer perceptron is trained to classify a number of instances into different classes using the backpropagation method. Upon completion of the training process, it is observed that the mean square error over the training set is one order of magnitude lower than the mean square error over the test set. The situation described here is an example of:
    a. Overfitting
    b. Underfitting
    c. None of the above

2. With regard to the previous problem, which of the following methods would you employ in order to improve the situation?
    a. Apply weight pruning techniques
    b. Apply weight elimination techniques
    c. Add more hidden layers
    d. Use a validation set
    e. Add more outputs to the network
    f. Augment the training set with artificial data
    g. Double the number of nodes in each hidden layer

3. Which of the following activation functions can produce a negative output?
    a. Sigmoid logistic 1/(1+exp(-x))
    b. ReLU
    c. Leaky ReLU
    d. Hyperbolic tangent tanh(x)

4. The maximum value of the derivative of the logistic sigmoid function $f(x) = 1/[1 + \exp(-x)]$ is:
    a. 0.25
    b. 0
    c. 1
    d. 0.125
    e. -0.75

5. Indicate which of the following statements are true:

    The kernel trick in support vector machines
    a. Is used to greatly reduce the complexity of evaluating inner products in multidimensional spaces
    b. Allows us to solve non-linearly separable problems with minimal changes to the formulation used to solve linearly separable problems
    c. Can be used in conjunction with every non-linear activation function
    d. Can be used in conjunction with some non-linear activation functions
    e. Can be used with gaussian activation functions
    f. Cannot be used with polynomial functions
    g. Can be used safely with logistic sigmoid functions

6. For a hidden Markov Model with *M* discrete states, we are given a sequence of observations $x_1,x_2,\ldots,x_N$ and we wish to infer the sequence of states $q_1,q_2,\ldots,q_N$ which led to this sequence of observations (type 2 problem). In order to do this, we would need to apply:
   a. The back-propagation algorithm
   b. The expectation-maximization algorithm
   c. The Viterbi algorithm
   d. The MAP algorithm
   e. A quadratic programming algorithm

7. In the previous question, the complexity of our operations would be of order:
   a. MN^2
   b. $\log(MN)$
   c. MN
   d. NM^2
   e. $M\exp(N)$

8. In the previous question, it would be convenient to take logarithms of local probabilities in the trellis diagram in order to facilitate the computations.
   a. Yes
   b. No
   c. Can't say

9. In a speech recognition task, we model different phonemes with different Hidden Markov models. Given a sequence of observations, we wish to perform recognition by finding the total probability of this sequence for each HMM and determining the maximum of these probabilities (type 1 problem). In this plan, is it convenient to take logarithms of local probabilities in the trellis diagram to facilitate the computations?
   a. Yes
   b. No
   c. Can't say

10. Training of a Hidden Markov Model is usually achieved by applying
    a. A variant of the Expectation-Maximization method
    b. Back Propagation
    c. Quadratic Programming
    d. A variant of the nearest neighbours method

11. Compute appropriate synaptic weights and thresholds for a feedforward neural network with one input x, two hidden neurons and one output neuron, so that it achieves the following output:

$$y = \begin{cases} 0, x < -1 \\ 1, -1 < x < 1 \\ 0, x > 1 \end{cases}$$

Assume that the activation function of all neurons is of the form $f(x) = \begin{cases} 0, x < 0 \\ 1, x > 0 \end{cases}$

12. In a 2-D space, we use a linear support vector machine to classify pattern $x_1=(1,1)^{\mathbf{T}}$ into one class A and pattern $x_2=(-1,-1)^{\mathbf{T}}$ into another class B. Write down the dual Lagrangian. Calculate the Lagrange multipliers corresponding to the two patterns. Find the equation of the separating line and plot it in the two-dimensional plane.

13. Repeat the previous problem using a non-linear support vector machine with kernel $\mathrm{K}(\boldsymbol{x},\boldsymbol{y})=(1+\boldsymbol{x}\cdot\boldsymbol{y})^2$. Show that you obtain different Lagrange multipliers, but the same separating line as before.

14. A support vector machine with a Gaussian kernel is used to solve a non-linear classification task. Name two parameters that you would vary in order to achieve best generalization.

15. A specific layer in a convolutional neural network implements the following actions:

    1) Convolution with the kernel A given below (receptive field:2, stride:1)
    2) Application of a ReLU activation function
    3) max pooling on a 2x2 window.

Input:

1	8	-5
-9	-3	5
4	-8	5

Convolutional kernel A:

2	2
-1	-1

Compute the output of the layer (don’t use padding).

END OF EXAM