NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS
MACHINE LEARNING (M124)
FINAL EXAMINATION, SEPTEMBER 2021

GROUP A

Problem A1

A classifier is used to classify instances $\boldsymbol{x}$ into two classes A and B. Thus, the classifier assigns $\boldsymbol{x}$ to A if $P(\boldsymbol{x}|A)>P(\boldsymbol{x}|B)$ and to B if $P(\boldsymbol{x}|A)<P(\boldsymbol{x}|B)$.

a) Under which condition is this classifier identical to the Bayes classifier?
b) Provided the condition holds and this is indeed a Bayes classifier, we have discussed in class that it is considered an optimum classifier. Yet, in real life problems, it can exhibit a suboptimal performance compared to other classifiers. Provide reasons why this is possible.

Problem A2

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 a Hopfield model grows linearly with the number of neurons.
c) The joint probability in a Bayesian network in equal to the sum of the (conditional) probabilities associated to its nodes.
d) With the assumption of white noise, the least squares method gives identical results with the maximum likelihood method in regression problems.
e) It is possible that a biased estimator perform better than an unbiased estimator.
f) For a k-nearest neighbour classifier used to classify instances into 2 classes, it is advisable that k is even.

Problem A3

We are given a set of 500 pairs (x_i,y_i), i=1,…,500 and we seek to perform generalized linear regression using the ridge regression method. In truth, data are generated by a 4th degree polynomial in x with added noise. We employ 10-fold cross validation and use our well known formula for estimating the parameter vector $\boldsymbol{\theta}$:

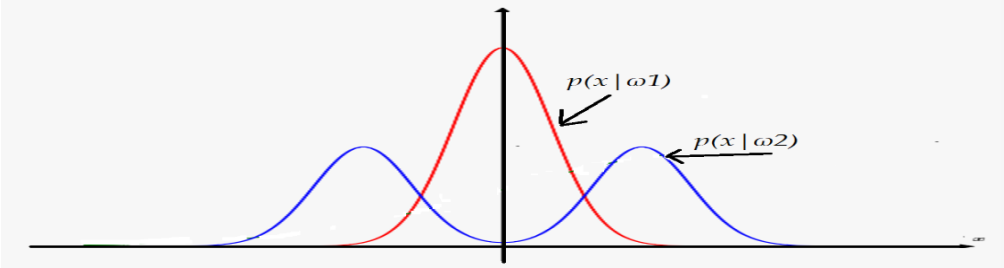
$$\widehat{\boldsymbol{\theta}} = \left(\Phi^{\mathrm{T}}\Phi + \lambda \mathrm{I}\right)^{-1}\Phi^{\mathrm{T}}\boldsymbol{y}$$

What is the number of rows and the number of columns in matrix Φ? What is the number of columns in the unit matrix I?

Problem A4

Probability density functions are shown for two equiprobable categories ω1 and ω2. 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.



Problem A5

The one-dimensional Pareto distribution has the following probability density function:

$$f(x) = a / x^{a+1} \text{ for } x \geq 1$$

The «shape parameter» α takes on positive values. We are given *N* random samples $X_1, X_2, \ldots, X_n$ originating from this distribution. Show that the maximum likelihood estimate of α is:

$$a_{ML} = \frac{N}{\sum_{i=1}^{N} \ln(X_i)}$$

GROUP B

Problem B1

We use a support vector machine with polynomial kernel to classify a number of instances into two non-linearly separable classes. Write down the dual Lagrangian related to this problem. Identify two parameters likely to affect the classification accuracy of the support vector machine.

Problem B2

We train a multilayer perceptron to classify a number of instances into different classes using the back propagation method. Upon completion of the training process, we observe 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 above is an example of:

a) overfitting
b) underfitting
c) none of the above

Problem B3

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

Problem B4

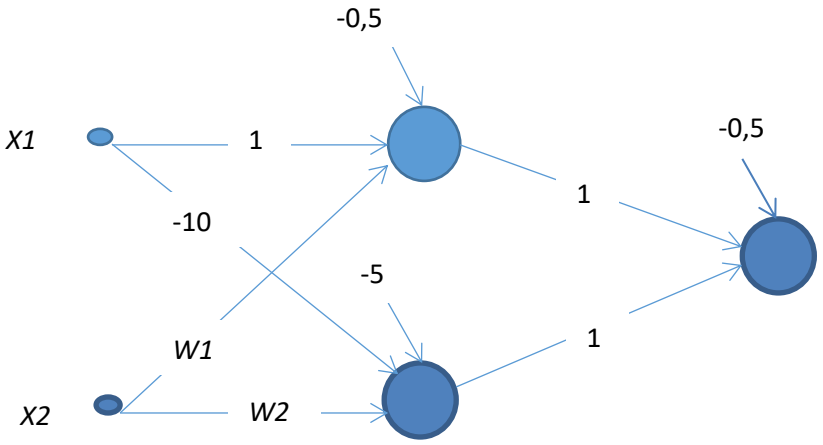
Training of a Hidden Markov Model is usually achieved by applying

a) A variant of the Expectation-Maximization method
b) Back Propagation
c) Quadratic Programming
d) A variant of the nearest neighbour method

Problem B5

The depicted feedforward network realizes the XOR function (output 0 when both inputs are 0 or 1, output 1 when one input is 1 and the other 0). All neurons have activation functions of the form:

$$f(z)=\begin{cases}1, z \geq 0\\ 0, z<0\end{cases}$$



a) Show that no matter what the values of W1, W2 are, the network gives us the desired output for the pairs (X1,X2)=(1,0) and (X1,X2)=(0,0).
b) With the data shown in the diagram and given that the network realizes the XOR function, which are the output values of the two intermediate neurons for the input pairs (X1,X2)=(0,1) and (X1,X2)=(1,1)?
c) Find the allowed values of W1 και W2, so that the network realizes the XOR function.
d) For W1=-1 and W2=10, draw a diagram to show which regions in the 2-dimensional input space are classified to each of the two classes.
e) Is it feasible to train this network with the back propagation algorithm? If not, which modification is necessary to make this possible?

Problem B6

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

a) Convolution with the kernel A given below (receptive field:2, stride:1)
b) Application of the non-linear function f: f(z)=max(0,z)-min(0,z)
c) 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.