Instituto Superior Técnico

Machine Learning (Aprendizagem Automática)

Exam of 30/1/2017. Duration: 3 hours

Number:	**First and last names:**

Important notes:

- **Present all answers in a very clear, ordered and detailed manner, with all steps of the calculations, and with a brief explanation of each one.**

  If you don't present your answers as indicated, your marks will be reduced, and can be very low, even if your answers are correct.

- Justify all answers, except in multiple-choice questions.
- Keep at least four significant digits in all calculations.
- For problems 1 to 4, the instructions on how to answer are given in the problems themselves. Solve problems 5 to 8 in separate sheets. You may solve several items of each of the latter problems in the same sheet.

Answer table for problems 1 to 4:

	a)	**b)**	**c)**	**d)**	**e)**
Problem 1					
Problem 2					
Problem 3					
Problem 4					

Problem 1 (0.8 marks for each item; wrong answers will be marked -0.5)

Answer this problem by writing, in the answer table above, in the box corresponding to each item, "T" for "True", or "F" for "False".

Indicate whether each of the following assertions is true or false.

a) The K-means method always yields a global optimum of its objective function, at the end of the training process.

b) The EM algorithm, when used to estimate probability densities by means of mixtures of Gaussians, always converges to a global optimum of its objective function.

c) Principal components analysis always yields a global optimum of its objective function.

Problem 2 (0.8 marks for each item; wrong answers will be marked -0.5)

Answer this problem by writing, in the answer table above, in the box corresponding to each item, "T" for "True", or "F" for "False".

Indicate whether each of the following assertions is true or false.

a) The stochastic gradient descent method converges to a stationary point of the objective function if the step-size parameter tends to zero when the iteration number tends to infinity.

b) In multilayer perceptrons, the advantage of on-line learning over batch-mode learning generally increases as the size of the training set increases.

c) The construction phase of the ID3 method chooses, for each node of the decision tree, the attribute that, when used at that node, minimizes the number of errors in the training set (keeping the attributes already chosen at other nodes fixed).

d) A linear, hard-margin classification support vector machine with separable training data from two different classes cannot have just one support vector.

e) A linear, hard-margin classification support vector machine with separable training data from two different classes, in a three-dimensional space, can have just two support vectors.

Problem 3 (0.8 marks for each item; wrong answers will be marked -0.5)

Answer this problem by writing "A" or "B" in the box corresponding to each item, in the answer table at the beginning of this exam.

Consider data represented by a random variable $X \in \mathbb{R}$. These data belong to one of two classes, A and B. The following probability distributions are known:

$$P(A) = \frac{1}{3} \qquad P(B) = \frac{2}{3} \qquad p(x|A) = \begin{cases} 1 - x/2 & \text{if } 0 \le x \le 2 \\ 0 & \text{otherwise} \end{cases} \qquad p(x|B) = \begin{cases} 2\, e^{-2x} & \text{if } x \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

Consider the Bayes classifier for these data. Indicate to which class that classifier will assign each of the following data points:

a) $x = 0.6$ b) $x = 1$ c) $x = 1.9$

Problem 4 (0.8 marks; wrong answers will be marked -0.2)

Answer this problem by writing, in the answer table at the beginning of this exam, an "X" in the box corresponding to the sentence that you chose. Do not choose more than one sentence.

Indicate a true sentence.

a) Principal components analysis maximizes the likelihood of the principal components.

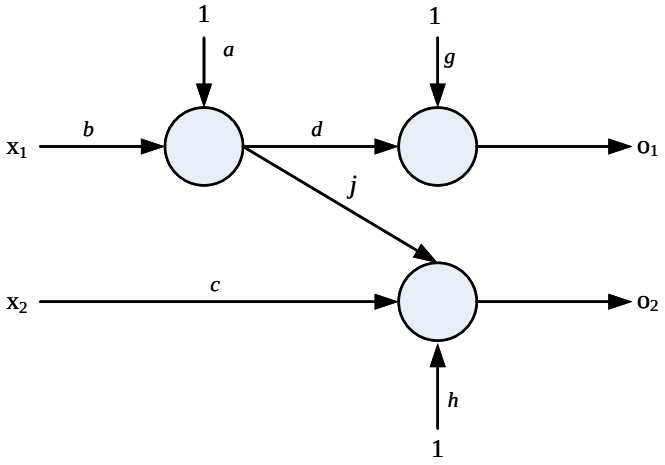
b) Principal components analysis maximizes the likelihood of the training data.

c) Principal components analysis maximizes the variance of the reconstructed data.

d) Principal components analysis maximizes the posterior probability of the principal components.

Problem 5

Consider the following multilayer perceptron (MLP), in which all the units have as activation function $S(s) = \frac{s^2}{1+s^2}$. Also consider the training set given in the following table.



x_1	x_2	d_1	d_2
-1	1	0.5	-0.5
1	-1	-0.5	0.5

a) (1 mark) Draw the backpropagation network. Don't forget to include the gains of all branches, as well as the input variables.

b) (1.6 marks) Compute the value of the weight b after the first update using backpropagation in **batch mode**, assuming that, initially, all the weights were equal to 0.1. Training is performed with the fixed step-size parameter $\eta = 0.2$ and without momentum. The cost function is the total squared error.

a) c) (1 mark) Repeat item b) above, but now using backpropagation in **real time**, and using as cost function the one of item b) plus a regularization term for exponential weight decay with parameter $\lambda = 0.1$. Assume again that, initially, all the weights were equal to 0.1, and that $\eta = 0.2$.

Problem 6

Consider a classification problem with the training set given in the following table, in which the x_i represent components of the input patterns and d represents the desired classification of each pattern.

x_1	x_2	x_3	d	x_1	x_2	x_3	d
1	5	4	0	0	2	0	1
2	1	5	0	2	3	1	1
3	3	6	0	5	2	5	1

a) (2 marks) Construct a support vector machine (SVM) for this problem, using the mapping from input space to feature space $\varphi(\bar{x}) = [x_1 x_2, x_3]^T$. Indicate, in your response, the SVM's condition for classifying the patterns into the class that corresponds to $d = 1$, as well as the equations of the SVM's margin boundaries, all expressed in terms of the coordinates of the input space. You may find the feature-space classifier by inspection, but you should clearly explain how you have found it.

b) (0.8 marks) Find the kernel that corresponds to the mapping from input space to feature space given in item a) above.

Problem 7 (2 marks)

Consider the following data set: $\{\bar{x}^1,\ \bar{x}^2,\ \bar{x}^3\} = \{(-3,-1),(-4,-2),(3,1)\}$. Perform one iteration of the EM algorithm to estimate the parameters of a mixture of two Gaussians to approximate the probability density of the distribution from which these data were drawn. Assume that, initially, $\mu_1 = (0,-1)$, $\mu_2 = (0,1)$ and $w_1 = 0.4$, and that the initial covariances of both Gaussians were equal to the identity matrix.

The values of the two Gaussian densities at the data points, for the initial parameter values, are supplied in the following table, except for $g_1(\bar{x}^{\,2})$, which you will need to compute.

	$g_1(\bar{x})$	$g_2(\bar{x})$
$\bar{x}^1$	0.001768	2.393×10^{-4}
$\bar{x}^2$	– not supplied –	5.931×10^{-7}
$\bar{x}^3$	2.393×10^{-4}	0.001768

Problem 8 (2 marks)

This problem is intended to distinguish the students that deal best with the topics studied in our course. In this problem, you must provide a very detailed response, with a very careful justification of each step.

Consider a linear, hard-margin classification support vector machine. Let $\bar{x}^k,\ k = 1,\cdots,K$ be the input patterns from its training set, which contains patterns from two different classes, and which is assumed to be linearly separable. Assume, for simplicity, that the input patterns are ordered so that the SVM's support vectors are $\bar{x}^k,\ k = 1,\cdots,p$. As you know, the SVM's classification boundary obeys an equation of the form $\bar{w} \cdot \bar{x} + b = 0$, and the vector $\bar{w}$ in this equation can be expressed as a linear combination,

$$\bar{w} = \sum_{k=1}^{p} \alpha_k \bar{x}^k. \tag{1}$$

Show that, for any such vector $\bar{w}$, the coefficients α_k in equation (1) can be chosen so that

$$\sum_{k=1}^{p} \alpha_k = 0.$$

Suggestion: Consider the linear SVM whose training patterns are $\tilde{x}^k = \bar{x}^k - \bar{x}^1,\ k = 1,\cdots,K$, with the same classifications as in the original training set.