Instituto Superior Técnico

Machine Learning (Aprendizagem Automática)

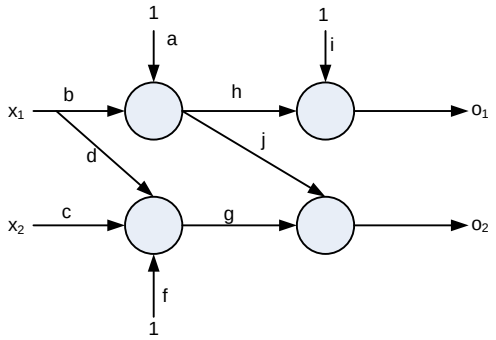
Exam of 25/1/2016. Duration: 3 hours

Important notes:

- **Solve the problems in separate sheets.** You may solve several items of each problem in the same sheet.
- Present all answers in a clear, ordered and detailed manner.
- In calculations, present all steps, with a brief explanation of each one.
- **Justify all answers and all calculation steps.**
- Keep at least four significant digits in all calculations.

Problem 1

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



x_1	x_2	d_1	d_2
-1	1	0.5	0.5
1	-1	0	-0.5

a) (1.2 points) Draw the backpropagation network. Don't forget to include the gains of all branches, as well as the input variables.

b) (1.8 points) Compute the value of weight **d** after the first update using backpropagation in **real-time mode**, assuming that, initially, all the weights are equal to 0.5, and that the training is performed using the fixed step size parameter $\eta = 0.1$. The cost function is

$$C = \frac{1}{2}\sum_{k=1}^{K} \left\| \bar{o}^k - \bar{d}^k \right\|^2.$$

c) (1 point) Assume that you had a validation set available. Explain how you would use that set to help you choose the MLP's final weights after several iterations of training.

Problem 2

The following data points are to be used to estimate a mixture of Gaussians with 2 components:

$$x^1 = [-3,-1]^T, \quad x^2 = [-1,-2]^T, \quad \text{and} \quad x^3 = [3,1]^T.$$

Consider the initial conditions $\bar{\mu}_1 = [0,0]^T$, $\bar{\mu}_2\ [-1,1]^T$, $V_1 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$, $V_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $w_1 = w_2 = 0.5$ (V_i are the covariance matrices of the Gaussian distributions). The values of the two Gaussian densities at the data points, for the initial parameter values, are supplied:

	$g_1(x)$	$g_2(\mathrm{x})$
x^1	$6.532\text{x}10^{-3}$	$2.915\text{x}10^{-3}$
x^2	$2.279\text{x}10^{-2}$	$1.678\text{x}10^{-3}$
x^3	$6.532\text{x}10^{-3}$	$5.339\text{x}10^{-6}$

a) (1 point) Compute the log-likelihood obtained using the initial parameter values.

b) (1.8 points) Find the values of the parameters $w_1, w_2, \bar{\mu}_1$ and V_1 that result from the first iteration of the EM algorithm.

c) (1 point) Consider, for a given data set, the density estimates that would be obtained by running the EM algorithm until convergence, using mixtures of 2 and of 3 Gaussians. Which of the two estimates would you expect to have a higher likelihood?

Problem 3

Consider a classification problem in two dimensions, with two classes, in which the training set is given by

x_1	x_2	Class	x_1	x_2	Class
0	1	-1	$\sqrt{3}$	2	$+1$
1	0	-1	2	0	$+1$
-1	0	-1	-2	0	$+1$
0	3	-1			

a) (0.4 points) Plot the training patterns. Are the classes linearly separable?

b) (0.8 points) Consider the following mapping from input space to feature space: $\varphi(\bar{x}) = (x_1^2, x_2^2)$. Find the corresponding kernel.

c) (0.4 points) Plot the patterns in the feature space.

d) (1.3 points) Find, in the feature space, the linear support vector machine for these data. Find the equations of the classification boundary and of the margin boundaries, and plot those boundaries on the plot from item b). You may find the classifier by inspection, but you must explain what you have done.

e) (1.3 points) Find the corresponding support vector machine in the input space. Plot the classification boundary and the margin boundaries on the plot from item a). Indicate, in the plot, the support vectors, and the classification regions for both classes.

f) (0.8 points) Find, in terms of the input space coordinates, the mathematical condition under which the classifier will classify a pattern into class +1.

Problem 4

Consider three-dimensional patterns, $\bar{x} = (x_1, x_2, x_3)$, with $x_1, x_2 \in \{0,1\}$ and $x_3 \in \{0,1,2\}$, and a binary classification, $C \in \{C_1, C_2\}$. Assume that there is an unknown joint probability distribution of $\bar{x}$ and C, and that the following training patterns and classifications were obtained by independently sampling from that distribution.

$\bar{x}$	C	Number
(0,0,0)	C_1	2
(0,0,0)	C_2	1
(0,0,1)	C_1	1
(0,1,0)	C_2	2

$\bar{x}$	C	Number
(0,1,1)	C_1	1
(0,1,1)	C_2	2
(1,0,0)	C_1	1
(1,0,1)	C_1	1

$\bar{x}$	C	Number
(1,0,1)	C_2	1
(1,1,0)	C_2	1
(1,1,1)	C_1	1
(1,1,1)	C_2	2

The columns labeled "Number" indicate how many exemplars of the corresponding pair $(\bar{x}, C)$ exist.

a) (1 point) Find the classification of the pattern (0,1,1) using an approximation of the Bayes classifier with probabilities estimated from this training set.

b) (1 point) Find the classification of the pattern (0,1,2) according to the Naïve Bayes classifier for these data, using x_1, x_2 and x_3 as features.
Suggestion: use Laplacian smoothing.

Alternative: (0.7 points) If you're not able to find the classification of the pattern (0,1,2), find the classification of the pattern (0,1,1).

Problem 5

Consider the training set $\{(3,-1), (2,1), (5,3), (6,1)\}$.

a) (1.2 points) Find the principal directions of this set. You may indicate those directions by means of vectors.

b) (1 point) Find the variances of the principal components of these data.

c) (1 point) Consider the determination of the first principal direction of a probability distribution. Explain which is the quantity that is optimized in that process, whether it is minimized or maximized, and why optimizing that quantity is reasonable in this case. Be as precise and detailed as possible in your explanation, using mathematical expressions where appropriate.

Problem 6 (2 points)

Consider a linear support vector machine that has three support vectors, $\bar{x}^1, \bar{x}^2$ and $\bar{x}^3$, in which $\bar{x}^1$ is of class C_1, and $\bar{x}^2$ and $\bar{x}^3$ are of class C_2. As you know, the machine's classification boundary can be expressed in the form $\bar{w} \cdot \bar{x} + b = 0$, in which $\bar{x}$ is an input pattern, $\bar{w}$ is a suitable vector, and b is a suitable scalar. Find, as a function of the support vectors, a vector $\bar{w}$ that can be used to represent the classification boundary in the above mentioned form.