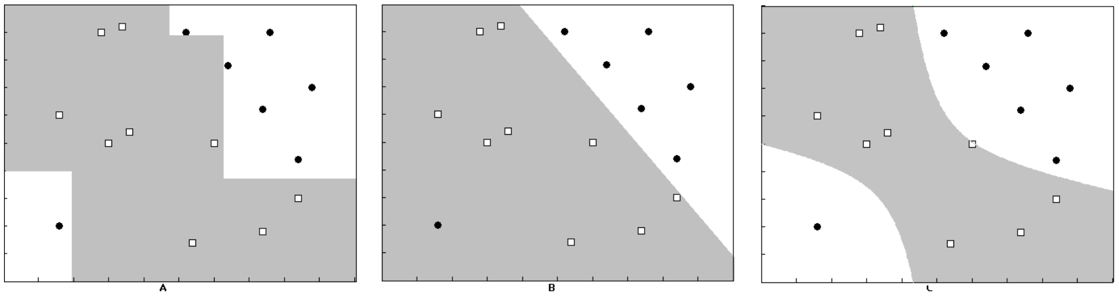
Exam 25 January 2016
Duration: 2h30min

1. Consider a classification problem with two real valued inputs. For each of the following algorithms, specify all of the separators below (depicted in the training set) that it could have generated and explain why. If it could not have generated any of the separators, explain why not.



A. Nearest neighbour, with k=1
B. Probabilistic Gaussian Discriminant, with the same covariance for both classes
C. Probabilistic Gaussian Discriminant, with different covariance for the two classes
D. Decision Tree

2. 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	3	A	0	0	B
0	-3	A	0	-2	B
4	1	A	1	1	B
4	-2	A			

a) Sketch the points of the training set in the input space. You noticed the classes aren't linearly separable but you want to apply a linear classifier. What could you do?
b) What would be the prediction of a nearest neighbour classifier at the point (2;1) using k=1? And with k=3? And with k = 10?
c) In practice, how do you set k in the nearest neighbour?
d) Consider the classifier that classifies points x according to which of the two class means is closer (in the input space and using Euclidean distance). Classify the 7 points given in the table above.

3. Consider the dataset

x₁	**x₂**
0	3
0	-3
4	1
4	-2

x₁	**x₂**
0	0
0	-2
1	1

(It's the dataset provided in question 2 but **without** labels)

a) Iterate the k-means algorithm, for two iterations, using two centers with initial positions (2;-0.25) and (1/3; -1/3). Compute the value of the algorithm's cost function at the initial positions and after each iteration.

b) Prove that the value of the cost function in the k-means algorithm is never increasing after each iteration (the algorithm converges monotonically to a solution)

4. Indicate which, if any, of the statements are correct, and explain your answer. Assuming a linearly separable dataset, Support Vector Machines typically:

(i) when evaluated on the training set, achieve higher accuracy than Perceptrons.

(ii) when evaluated on the test set, achieve higher accuracy than Perceptrons because SVMs can be used with kernels.

5. Consider the dataset

x_1	x_2	y
1	1	2.5
2	1	4
1	2	3.5

a. Using linear regression, estimate w_1 and w_2 of the model $y=w_1x_1+w_2x_2$.

b. Knowing that $\boldsymbol{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ can be written as $\boldsymbol{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \alpha_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \alpha_2 \begin{bmatrix} 2 \\ 1 \end{bmatrix} + \alpha_3 \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, with $\alpha_i \in \boldsymbol{R}$, compute suitable values for α_i.

c. Consider the linear regularized risk minimization problem

$$min_{\boldsymbol{w}} \sum_{i=1}^{n} L(\boldsymbol{w^t x_i}, y_i) + \lambda \left\|\boldsymbol{w}\right\|^2$$

$L\,(.)$ is a generic loss function and $\lambda > 0$ is a user defined coefficient, and $x_i \in R^d$, where d is the dimension of the input space. Prove that the solution **w** belongs to the span of the samples. In other words,

$$\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i \boldsymbol{x}_i$$

where $\boldsymbol{\alpha} \in R^n$ is a vector of coefficients α_i that implicitly defines the solution and *n* is the number of observations in the sample.

d. Applying the result from c) to linear regression we know that an alternative computation of **w** is to first estimate α_i from the data and then obtain **w**. When (relate *n* and *d*) would this last approach be a good option?

6. A friend of mine likes to climb on the roofs of FEUP. To make a good start to the coming week, he climbs on a Sunday with probability 0.98. Being concerned for his own safety, he is less likely to climb today if he climbed yesterday, so

Pr(climb today|climb yesterday) =0.4

If he did not climb yesterday then he is very likely to climb today, so

Pr(climb today|¬climb yesterday) =0.9

Unfortunately, he is not a very good climber, and is quite likely to injure himself if he goes climbing, so

Pr(injury|climb today) =0.8 whereas

Pr(injury|¬climb today) =0.1

a) Explain how my friend's behaviour can be formulated as a Hidden Markov Model. What assumptions are required?

b) You learn that on Monday and Tuesday evening he obtains an injury, but on Wednesday evening he does not. Compute the probability that he climbed on Wednesday. (You can use Matlab to do the computations but write the major steps.)

7. Define an HMM that generates a sequence of the form $A^{k_1}C^{4-k_1}A^{k_2}C^{4-k_2}$… where, e.g., A^{k_1} represents a series of length k$_1$ consisting of only A's. The k$_1$, k$_2$, k$_3$, … are drawn from the set {1,2,3} with equal probabilities. The observation follows a Bernoulli distribution (that is, the observations take <u>only two</u> possible values).