1º Teste de Aprendizagem Automática

3 pages, 5 questions, 2 answer sheets. Duration: 2 hours
DI, FCT/UNL, 30 April de 2014

Question 1 [4 points] 200 replicas of a set of 50 examples of known value were created by random sampling with reposition. For each of six different regression models (A through F) 200 hypotheses were computed adjusting each model to each of the replicas of the initial set. Then the following X and Y values were computed for each model:

$$X = \frac{1}{50}\sum_{t=1}^{50}[\bar{g}(x_t) - y_t]^2 \qquad Y = \frac{1}{50\times 200}\sum_{t=1}^{50}\sum_{i=1}^{200}[\bar{g}(x_t) - g_i(x_t)]^2$$

where t is the index of the example in the initial set, $\bar{g}(x_t)$ is the average of the predicted values for x_t by the 200 hypotheses obtained, y_t the real value for point x_t and $g_i(x_t)$ the predicted value for point x_t by the hypotheses obtained training the model with replica i.

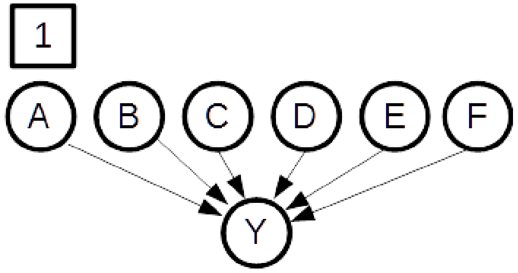
Model	X	Y
A	52	0.012
B	13	0.23
C	2	0.51
D	0.4	7
E	0.2	86
F	0.1	2067

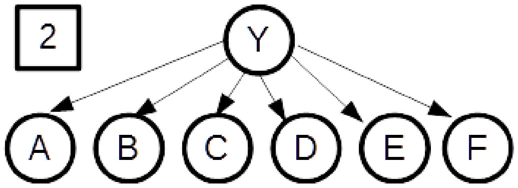
1.a) Indicate one model for which the generalization error should be mostly due to its inability to adjust to the examples. Justify your answer.

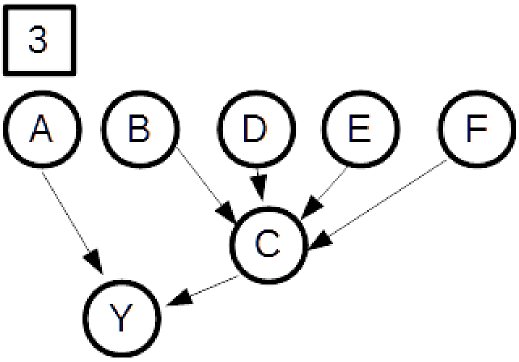
1.b) Indicate one model for which the generalization error should be mostly due to its excessive capacity to adjust to the examples in the training set. Justify your answer.

1.c) Indicate which model should have the lowest generalization error. Justify your answer.

Question 2 [4 points] The owner of a kiosk, a bayesian networks fan, wants to create a classifier to decide whether or not to grant credit to each client who requests it. He can describe each client with six categorical attributes (A through F) and each feature has four attributes. For instance, attribute A is the vehicle the client drives and has the categories luxury, heavy, utility and motorcycle. He doesn't know which of the three Bayes networks below he should use (1, 2 or 3) where Y is the class for each client (reliable or non-reliable).







2.a) Pick one of the networks (indicating which) and write down the expression for the joint probability distribution of the variables in that network.

2.b) Knowing the kiosk owner already has a database of 22 clients, with all attributes, and knowing that 10 paid previous debts on time and 12 did not, suggest the best of the three networks he should use to predict the reliability of future costumers. Justify your answer.

Question 3 [6 points] Consider the following classification model where $g(x)$ is the output for example x e the values w_n are the M+1 coefficients for the model (also counting w_0), where M is the dimension of the input vectors:

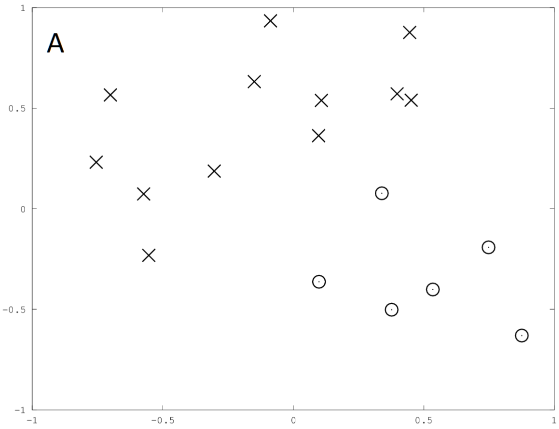
$$g(x) = \frac{1}{1+e^{-net(x)}} \qquad net(x) = w_0 + \sum_{i=1}^{M} w_i x_i$$

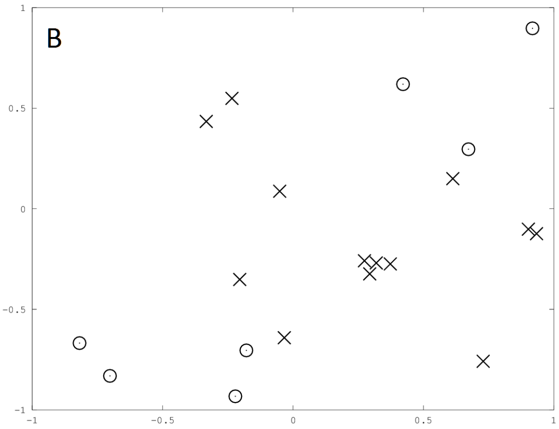
To obtain each hypothesise the model is trained presenting all examples several times in random order. Each time an example x_t is presented the model is updated changing each coefficient w_n with this equation:

$$\Delta w_n = \eta\Big(y(x_t) - g(x_t)\Big)g(x_t)\Big(1 - g(x_t)\Big)x_t^n$$

where $y(x_t)$ is the true class of example x_t, either 0 or 1, and x_t^n is the value of attribute n of x_t, considering this value to be 1 if n is 0. The constant η controls the learning rate. After training, an example x is predicted to belong to class 1 if $g(x)$ is greater than 0.5, or 0 otherwise ($g(x)$ outputs a value between 0 and 1).

Consider the two following sets of examples, A and B, where a circle represents a point in class 0 and a cross a point in class 1:





3.a) Indicate which of the two sets can be classified without errors using the model described above. Justify your answer.

3.b) Explain, briefly, how to modify the model, using the same kind of functions, to classify the other set that this model cannot classify without errors (you don't need to specify the equations, just to show you understand the concepts).

3.c) Explain what modifications you would have to make to the function that updates the coefficients if you built the model requested in 3.b (you don't need to specify the equations, just to show you understand the concepts).

Question 4 [4 points] To classify the set of examples represented in your answer sheet, a classifier was trained by computing the values of α that minimize one of the following expressions (A or B) where N is the number of examples, y the class of each example (a circle for class -1 and a cross for class $+1$) and x the vector with the coordinates of each point.

$$A)\ min_{\alpha}\left(\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m x_n^T x_m - \sum_{n=1}^{N}\alpha_n\right)$$

$$B)\ min_{\alpha}\left(\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m K(x_m,x_n) - \sum_{n=1}^{N}\alpha_n\right) \qquad K(x,z) = e^{\left(\frac{-||x-z||^2}{2\sigma^2}\right)}$$

We know that the minimization respected the following constraints:

$$0 \leq \alpha_n \leq 10, \quad n = 1, ..., N \qquad\qquad \sum_{n=1}^{N}\alpha_n y_n = 0$$

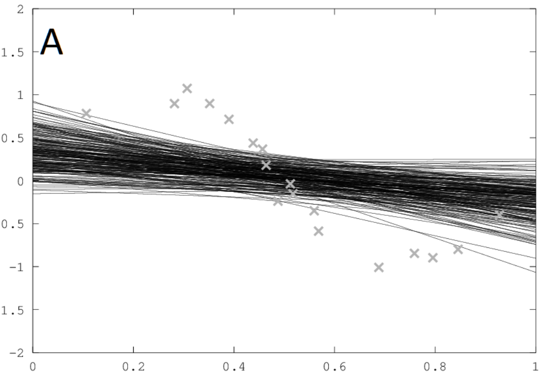
Unfortunately, there is no record of which of the two models was used to classify this set.

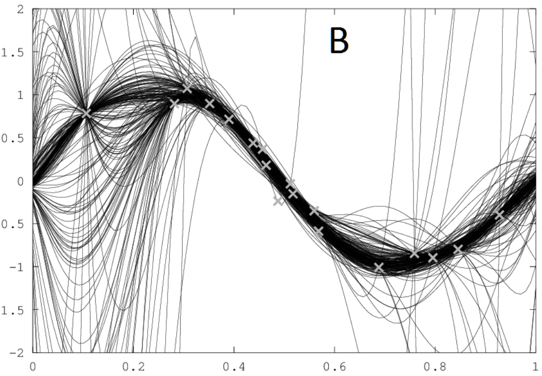
4.a) In the image on your answer sheet the thick line represents the decision boundary and the thinner lines the margins. Mark those points for which the corresponding α value is greater than one (circle around the points to mark).

4.b) Identify which of the two models (A ou B) was used to classify this set, resulting in the decision boundary depicted in the figure. Justify your answer.

4.c) In the conditions described above, and eventually with some other set of points, would it be possible to obtain a classifier that would incorrectly classify some point in the training set? Justify your answer.

Question 5 [2 points] The figures below show the result of training two regression models (black lines) using 200 replicas obtained by bootstrapping from the set of points shown in grey.





Indicate, justifying your answer, which of the two models (A or B) you would choose to create a regression function using bagging (the average of the 200 regression functions obtained by training the model on the replicas of the training set).

AA - Teste 1 - 2014-04-30

Numero:

0	○	○	○	○	○
1	○	○	○	○	○
2	○	○	○	○	○
3	○	○	○	○	○
4	○	○	○	○	○
5	○	○	○	○	○
6	○	○	○	○	○
7	○	○	○	○	○
8	○	○	○	○	○
9	○	○	○	○	○

Preencha o seu nome abaixo e o seu número à direita. Pinte por baixo de cada digito do seu numero o circulo correspondente. Por fim indique o número de alunos à sua frente e à sua direita pintando o circulo correspondente nos números abaixo.

Nome:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Alunos à Frente	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Alunos à Direita	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

1a)

1b)

1c)

2a)

2b)

2c)

AA - Teste 1 - 2014-04-30

Numero:					
0	○	○	○	○	○
1	○	○	○	○	○
2	○	○	○	○	○
3	○	○	○	○	○
4	○	○	○	○	○
5	○	○	○	○	○
6	○	○	○	○	○
7	○	○	○	○	○
8	○	○	○	○	○
9	○	○	○	○	○

Preencha o seu nome abaixo e o seu número à direita. Pinte por baixo de cada digito do seu numero o circulo correspondente. Por fim indique o número de alunos à sua frente e à sua direita pintando o circulo correspondente nos números abaixo.

Nome:

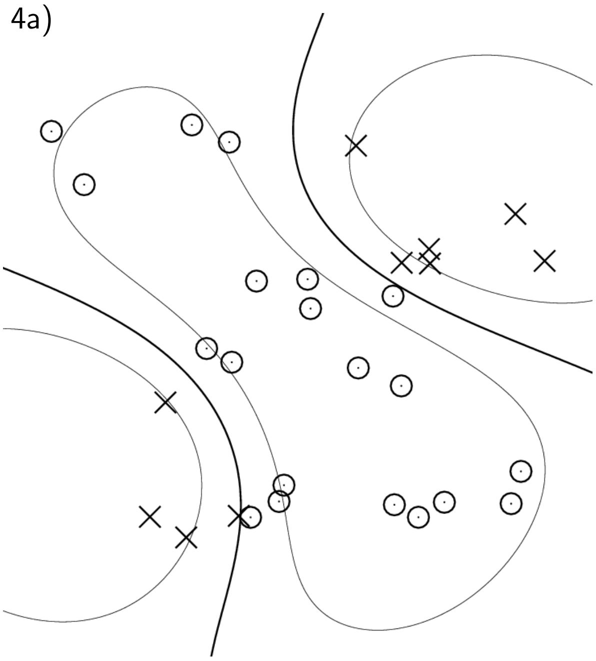
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Alunos à Frente	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Alunos à Direita	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

3a)

3b)

3c)

4a)



4b)

4c)

5