1º Teste de Aprendizagem Automática

2 páginas com 6 perguntas e 2 folhas de resposta. Duração: 1h 30m
DI, FCT/UNL, 12 de Dezembro de 2016

Pergunta 1 [3 valores]

1.a) In a classification problem, if we have a hypothesis class $\mathcal{H}$ with a finite number of hypotheses, we can say that the true error ($E(\hat{h})$) of the hypothesis with the smallest empirical error will be greater than the value in the expression below with a probability of δ (where $\hat{E}(h)$ is the empirical error of a hypothesis h):

$$\left(\min_{h\in\mathcal{H}} \hat{E}(h)\right) + 2\sqrt{\frac{1}{2m}\ln\frac{|\mathcal{H}|}{\delta}}$$

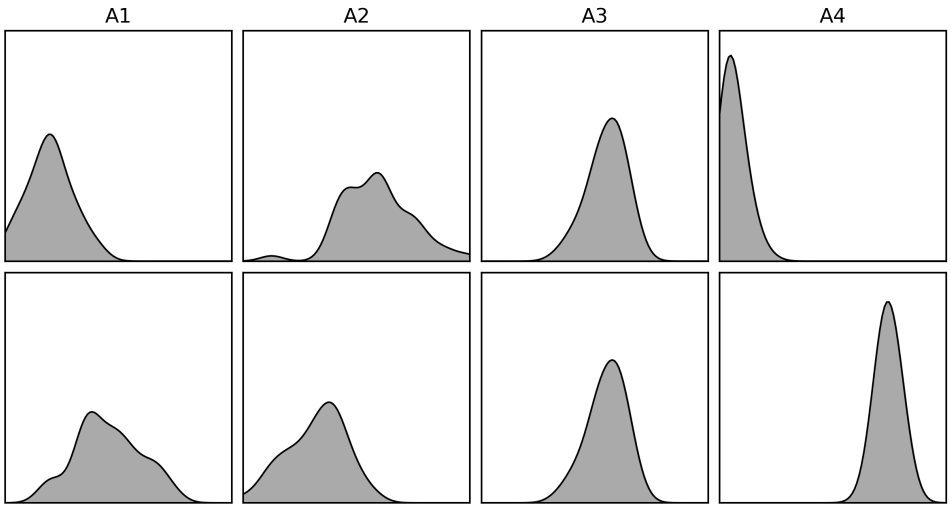
However, if the number of hypotheses in $\mathcal{H}$ is infinite (for example, if the parameters are continuous) we cannot use the number of hypotheses for this estimate of the generalization error. In this case, we need another measure for the capacity of the hypothesis class to adjust to the points to classify. What measure is used in this case? (You do not need to specify the name but you must explain the measure).

1.b) The expression below is an estimate of the probability that the true error of the hypothesis with the smallest empirical error ($E(\hat{h})$) is greater than its empirical error ($\hat{E}(\hat{h})$) plus an additional term. Explain why, in general, classification models with many parameters (such as deep neural networks, for example) have a good performance with large volumes of data but do not perform well when the data set is small.

$$P\left(E(\hat{h}) > \hat{E}(\hat{h}) + \mathcal{O}\left(\sqrt{\frac{VC(\mathcal{H})}{m}\ln\frac{m}{VC(\mathcal{H})} + \frac{1}{m}\ln\frac{1}{\delta}}\right)\right) = \delta$$

Pergunta 2 [4 valores]

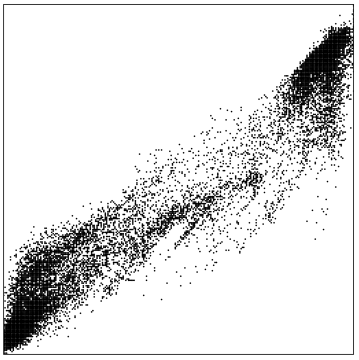
2.a) In a classification problem with two classes and four attributes, we want to discard one of the attributes to reduce the problem to 3 dimensions. The figure below shows the kernel density estimation plots for the distribution of values of each attribute (A1 a through A4). All plots are in the same scale. The top row shows the distributions for class 1 and the bottom row for class 2. Which attribute would you discard? Justify your answer.



2.b) Suppose that instead of selecting 3 attributes we reduced the dimensionality of this problem transforming the data with principal component analysis. What is the relative orientation of the 3 vectors corresponding to the three principal components and what is the measure that this process tries to preserve as much as possible?

Pergunta 3 [5 valores]

Suppose we have a set of points distributed as represented in the figure on the right. These points are samples from continuous values that we want to quantize into 20 points representing this distribution. For each algorithm below, indicate whether or not the algorithm is appropriate for this task. Justify each answer considering how each algorithm works. If the algorithm is appropriate indicate the values of the necessary parameters or explain how they can be computed.



3.a) K-Means.

3.b) DBSCAN

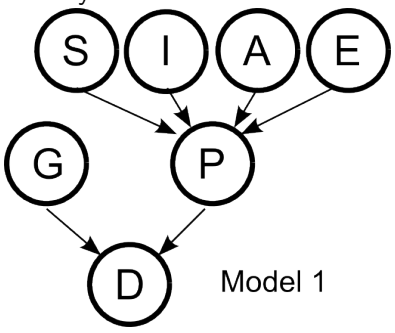
3.c) Affinity Propagation.

Pergunta 4 [3 valores]

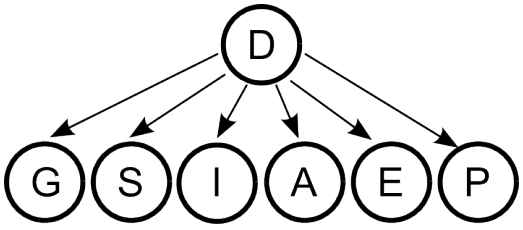
We have a set of 238 bivalve fossils. For each exemplar we have three anatomical measurements of the shell and, based on these values, we want to obtain a phylogenetic tree grouping the exemplars in species, genera, families and so forth in a hierarchy of groups. Suggest an appropriate algorithm for this task and explain how the algorithm works.

Pergunta 5 [3 valores]

In order to study the causal factors of diabetes, we want to compute the joint probability of the diabetic state (D) and the following causal factors: exercise (E), diet (A), blood pressure (P), age (I), general health (S) and blood glucose level (G). Consider the two graphical models below, where the arrows represent the conditional dependency relations between variables (bayesian networks):



Model 1



Model 2

5.a) Write, for each model, the expression of the joint probability distribution of all the variables as a function of their conditional dependencies, according to the respective model.

Note that you do not need to write "P(D,E,A,P,I,S,G) =". Just write the expression for each model.

5.b) Suppose you have 152 forms filled, for 152 persons, 76 of which are diabetic and 76 are not. In these forms, each parameter E, A, P, I, S, G is marked with an integer value from 0 through 9 and the variable D, for the diabetes diagnostic, is binary. Under these conditions, which of the two models should we use to estimate the joint probability distribution? Justify your answer.

Pergunta 6 [2 valores]

We have have a set of 500 examples and we assume that each example was generated by one of three gaussian random distributions. We know the attribute values for each example but we do not know from which distribution each example comes from. Briefly explain (no need for equations) how the Expectation-Maximization can be used to compute the parameters of the three gaussian distributions as well as the relative contribution of each distribution to the overall mixture.

AA-Teste2 2016-12-21

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 dígito do seu número o círculo correspondente. Por fim indique o número de filas de alunos à sua frente e o número de alunos à sua direita pintando o circulo correspondente abaixo.

Nome:

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

1a)

1b)

2a)

2b)

3a)

3b)

(3c na próxima folha. 3c on next sheet)

AA-Teste2 2016-12-21

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 dígito do seu número o círculo correspondente. Por fim indique o número de filas de alunos à sua frente e o número de alunos à sua direita pintando o circulo correspondente abaixo.

Nome:

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

3c)

4)

5a)

5b)

6)