2º Teste de Aprendizagem Automática

3 páginas de enunciado com 6 perguntas mais 2 folhas de resposta. Duração: 1h 30m
DI, FCT/UNL, 21 de Dezembro de 2017

Pergunta 1 [4 valores] Considere um problema de classificação com classes de hipóteses $\mathcal{H}$ com um número infinito de hipóteses. Podemos estimar um limite superior para o erro verdadeiro da hipótese com menor erro empírico $\hat{h}$ em relação ao seu erro empírico $\hat{E}(\hat{h})$. Há apenas uma probabilidade de δ do erro verdadeiro ultrapassar esta soma, onde VC é a dimensão Vapnik–Chervonenkis do classificador e m o tamanho do conjunto de treino:

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

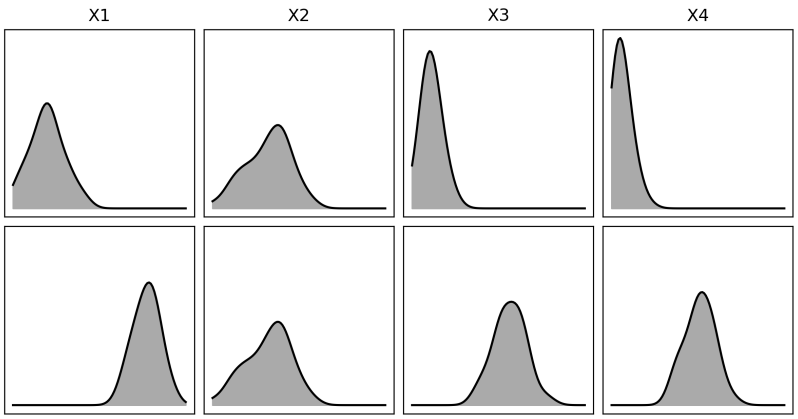
Dimensão	Bias	Var
2	0.308	0.001
3	0.252	0.002
4	0.104	0.005
5	0.012	0.102
6	0.008	0.120
7	0.006	0.328

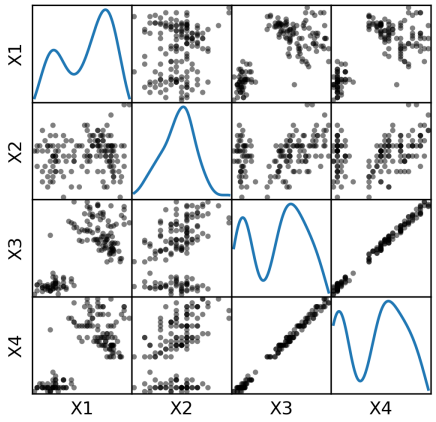
Partindo de um conjunto de dados a duas dimensões, a tabela mostra os valores de *bias* e *variance* calculados com classificadores lineares aplicados aos dados originais (dimensão 2) e a expansões não lineares dos dados originais para mais dimensões (3 a 7).

1.a) Indique a dimensão de um destes classificadores em que o termo $\hat{E}(\hat{h})$ seja o mais importante na estimativa do limite superior do erro verdadeiro. Justifique a sua resposta.

1.b) Indique a dimensão de um destes classificadores em que o termo $\sqrt{\frac{VC(\mathcal{H})}{m}\ln\frac{m}{VC(\mathcal{H})} + \frac{1}{m}\ln\frac{1}{\delta}}$ seja o mais importante na estimativa do limite superior do erro verdadeiro. Justifique a sua resposta.

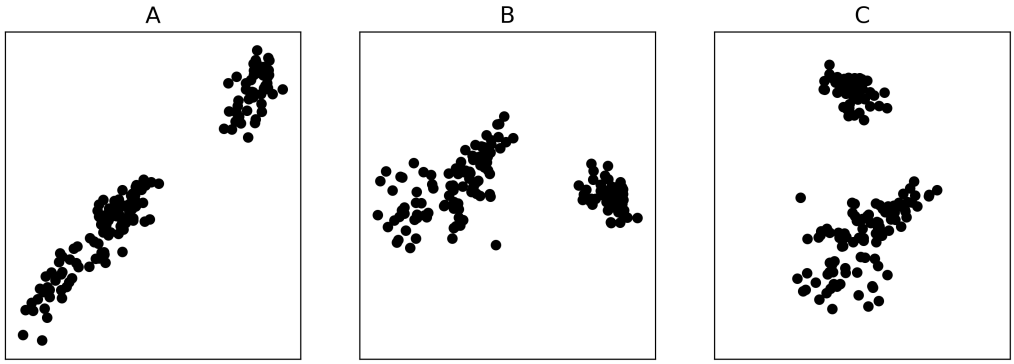
Pergunta 2 [4 valores] Num problema de classificação com duas classes queremos reduzir a dimensão de quatro para dois atributos. O painel da esquerda da figura abaixo mostra os gráficos de kernel density estimation da distribuição dos valores de cada atributo (X1 a X4), na mesma escala, com as distribuições para os exemplos da classe 1 na linha de cima e as distribuições dos exemplos da classe 2 na linha de baixo. O painel da direita mostra a matriz de gráficos de dispersão (*scatter matrix*) para estes dados.



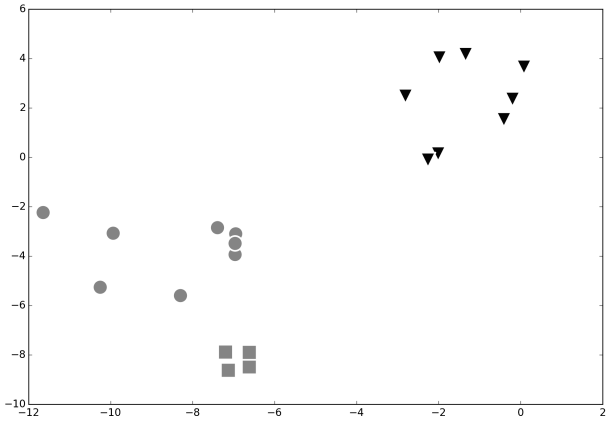


2.a) Quais os dois atributos que seleccionaria, destes quatro? Justifique a sua resposta e se houver alternativas enumere-as.

2.b) Suponha que em vez de seleccionar dois atributos fez uma análise de componentes principais (*Principal Component Analysis*) e projectou os dados nos dois componentes com maior valor próprio, com o principal no eixo horizontal o segundo no eixo vertical. Indique qual dos três gráficos abaixo (A, B, C) corresponde a esta projecção explicando porque é que os outros dois gráficos não podem ser os correctos. Note que as escalas são iguais nos 3 gráficos.



Pergunta 3 [3 valores]
A figura à direita mostra o resultado das duas primeiras iterações de um algoritmo de clustering. A primeira iteração separou os exemplos em dois grupos indicados pela cor, com os triângulos a negro num grupo e os restantes, a cinzento, no outro. A segunda iteração criou dois clusters adicionais com os exemplos do grupo a cinzento, diferenciados pelos quadrados e pelos círculos. Para cada um destes três algoritmos, indique se pode ou não pode ter sido esse o algoritmo usado, justificando a sua resposta:



3.a) Fuzzy C-Means

3.b) Bissecting K-Means

3.c) Agglomerative Clustering

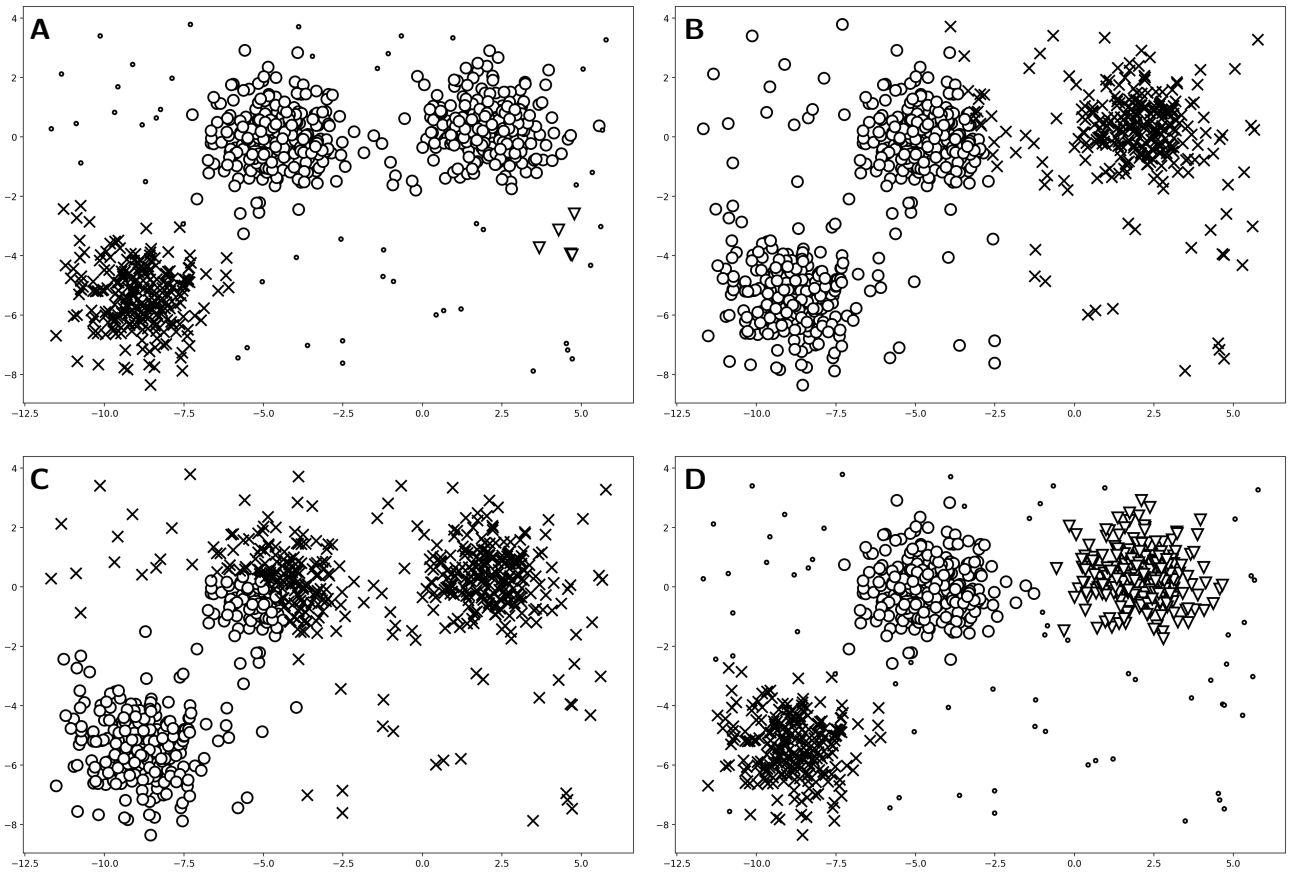
Pergunta 4 [6 valores] O mesmo conjunto de dados bidimensionais foi agrupado com algoritmos de *clustering* diferentes. Em todos os casos foi usada a distância euclideana como medida de distância entre pontos. Leia as quatro alíneas desta pergunta primeiro e observe atentamente os gráficos que se seguem antes de responder **justificando cada resposta**. No gráfico, cada símbolo maior representa um ponto num *cluster*. Os pontos que não foram atribuídos a um *cluster* são representados com círculos pequenos (gráficos A e D).

4.a) Qual gráfico (A, B, C, D) mostra o resultado da aplicação do algoritmo **Density-based spatial clustering of applications with noise (DBSCAN)**, com um valor de 10 para o número mínimo de vizinhos para os pontos de *core*?

4.b) Qual gráfico mostra o resultado da aplicação do algoritmo DBSCAN com um valor de 5 para o número mínimo de vizinhos para os pontos de *core*?

4.c) Qual gráfico mostra o resultado de *clustering* usando o algoritmo de **k-means** com um valor de $k = 2$?

4.d) Qual gráfico mostra o resultado de *clustering* usando uma **mistura de duas distribuições Gaussianas**?



Pergunta 5 [2 valores] O treino de modelos ocultos de Markov (*Hidden Markov Models*) e de algoritmos de clustering como K-Means e misturas de gaussianas (*Gaussian Mixture Models*) exige resolver o problema de maximizar a verosimilhança de um conjunto de parâmetros sem se saber todas as variáveis a modelar (por exemplo, os estado ocultos no HMM ou as atribuições aos clusters no K-Means e GMM). Explique sucintamente como se resolve esse problema.

Pergunta 6 [1 valores] Em teoria, uma rede neuronal com uma camada oculta pode aproximar qualquer função simplesmente aumentando o número de neurónios na camada oculta. Explique porque é que, ainda assim, é geralmente preferível usar uma rede neuronal profundas em vez de uma rede neuronal com uma camada oculta suficientemente grande.

AA, Segundo teste, 2017-12-21

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)

AA, Segundo teste, 2017-12-21

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	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

4a)

4b)

4c)

4d)

5)

6)