2º Teste de Aprendizagem Automática

3 páginas com 8 perguntas e 2 folhas de resposta. Duração: 2 horas
DI, FCT/UNL, 10 de Dezembro de 2015

Pergunta 1 [2 valores] A seguinte expressão indica o limite superior provável, com probabilidade $1-\delta$, para o erro verdadeiro de uma hipótese $\hat{h}$ obtida por minimização do erro empírico, em função da soma de dois termos. O primeiro termo é o menor erro verdadeiro possível no espaço finito de hipóteses $\mathcal{H}$. O segundo termo é função do número total de hipóteses em $\mathcal{H}$, o tamanho do conjunto de treino, m, e a probabilidade δ.

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

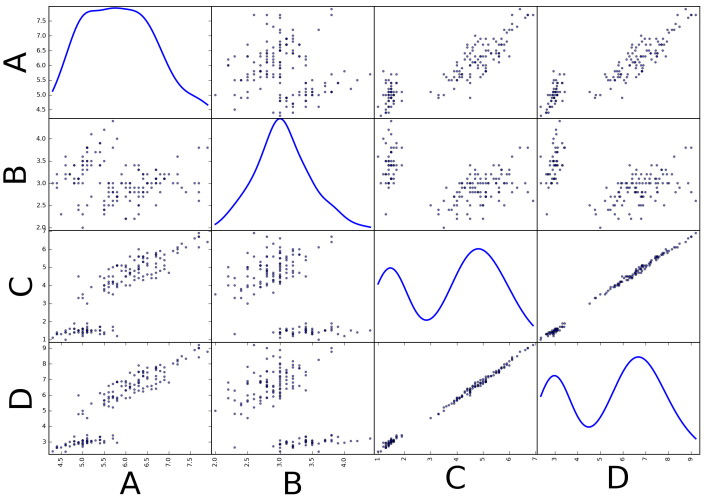
Considere também a seguinte frase: "Treinando dois modelos de classificação de forma a minimizar o erro no conjunto de treino, aquele que apresentar o menor erro no conjunto de treino terá provavelmente também o menor erro verdadeiro."

1.a) Indique que termo da expressão acima tem de ser dominante para que a frase seja **verdadeira**, e **justifique a sua resposta**.

1.b) Indique que termo da expressão acima tem de ser dominante para que a frase seja **falsa**, e **justifique a sua resposta**.

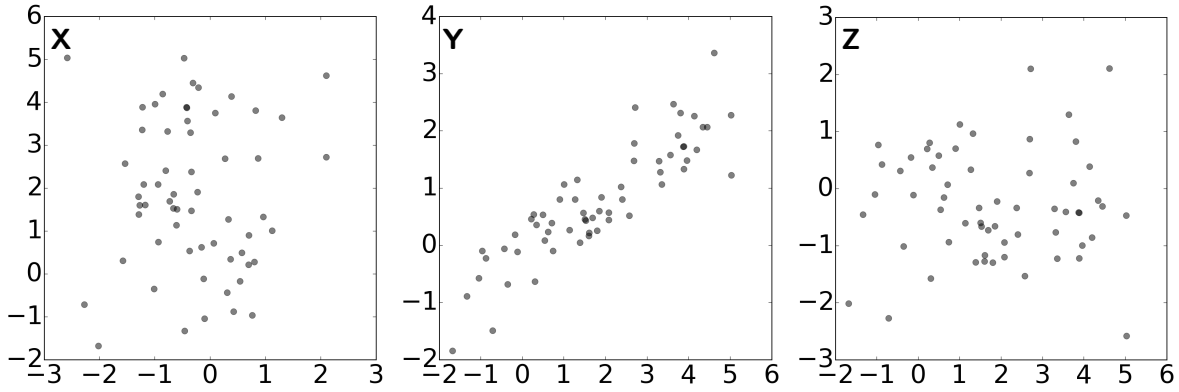
Pergunta 2 [2 valores] Em geral, nesta disciplina temos treinado classificadores minimizando o erro de classificação no conjunto de treino. Em que situações é que minimizar o erro de classificação durante o treino não é a melhor abordagem? Justifique a sua resposta e indique o que se deve fazer nesses casos.

Pergunta 3 [2 valores] A figura mostra a matriz de plots de todos os pares de atributos de um conjunto de dados com quatro atributos, A, B, C e D. Os gráficos na diagonal são Kernel Density Estimations das distribuições de cada atributo. Sabendo que este conjunto de dados não está etiquetado e será usado apenas para aprendizagem não supervisionada, e querendo filtrar um dos atributos para reduzir o conjunto a três atributos, indique, justificando, qual dos atributos eliminaria.

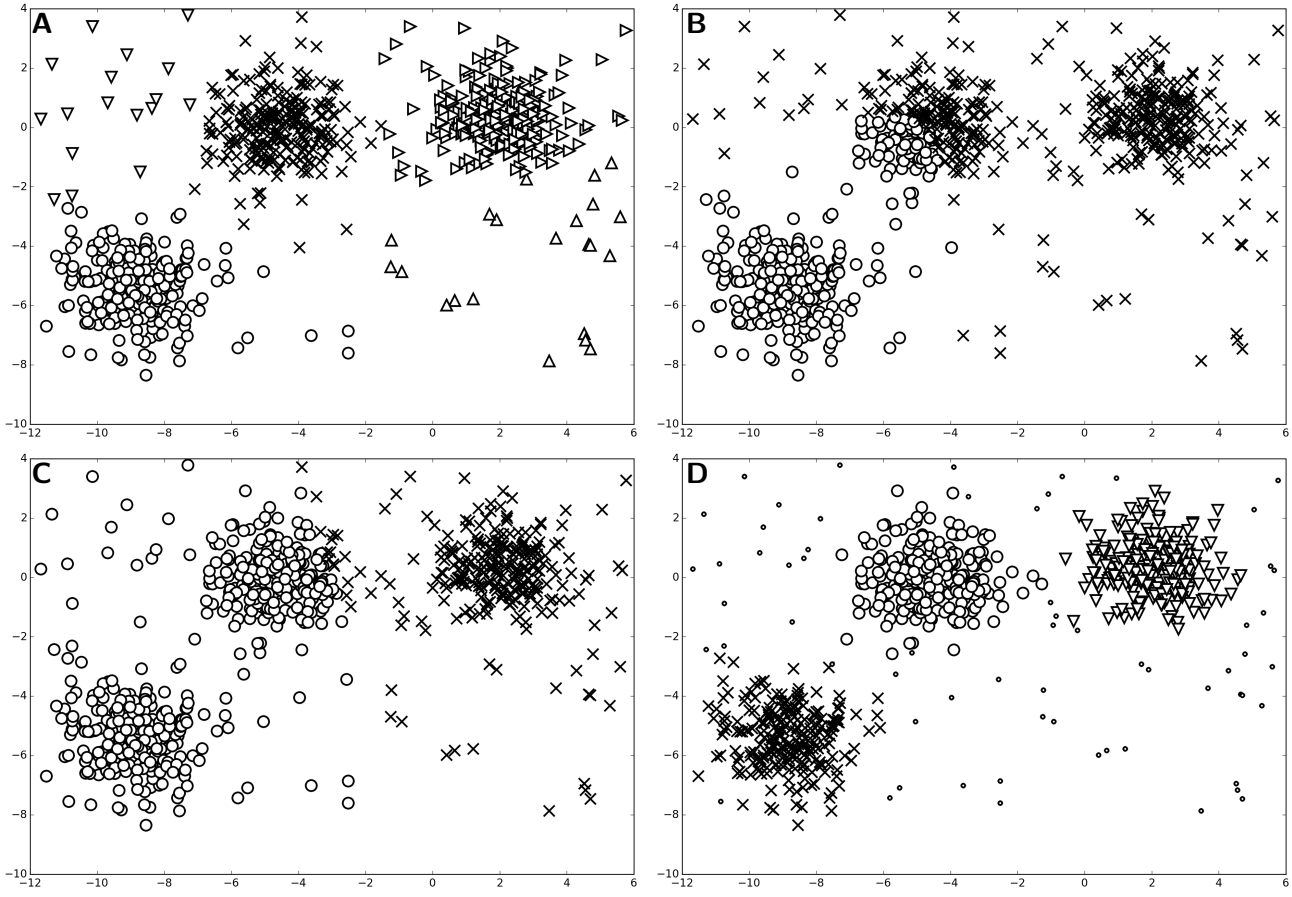


Pergunta 4 [2 valores] Descreva o algoritmo de treino de um mapa auto-organizável (*self-organizing map*, SOM). Assuma que eu já conheço a arquitectura desta rede, focando apenas o treino.

Pergunta 5 [2 valores] Para reduzir a dimensionalidade de um conjunto de dados foi feita uma análise de componente principal (Principal Component Analysis). O conjunto de dados foi projectado nos dois vectores próprios da matriz de covariância que tinham os maiores valores próprios. Aquele que tinha o maior valor próprio foi usado para traçar a coordenada x (horizontal) e o outro, com o segundo maior valor próprio, foi usado a coordenada y (vertical). Indique, **justificando**, qual dos três gráficos X, Y ou Z, representa o resultado desta projecção.



Pergunta 6 [6 valores] O mesmo conjunto de dados foi agrupado com quatro algoritmos de *clustering* diferentes. Leia as quatro alíneas desta pergunta primeiro, consultando o gráfico que se segue, e depois responda a cada uma **justificando a sua 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 (só ocorre no gráfico D).



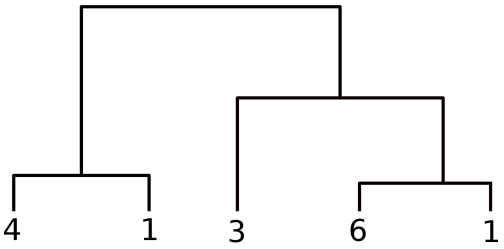
6.a) Qual dos quatro gráficos (A, B, C, D) mostra o resultado da aplicação do algoritmo **Density-based spatial clustering of applications with noise (DBSCAN)**, com um valor de 10 para o número mínimo de vizinhos para os pontos de *core*?

6.b) Qual dos quatro gráficos (A, B, C, D) mostra o resultado de *clustering* usando uma **mistura de duas distribuições Gaussianas**?

6.c) Qual dos quatro gráficos (A, B, C, D) mostra o resultado de *clustering* usando o algoritmo de **k-means** com um valor de $k = 2$?

6.d) Qual dos quatro gráficos (A, B, C, D) mostra o resultado de *clustering* usando o algoritmo de **Affinity Propagation** com um valor inicial de propensão (a diagonal da matriz de similaridade) igual a -200?

Pergunta 7 [2 valores] O gráfico à direita representa o dendrograma completo produzido por um algoritmo de *clustering* hierárquico. Os números nas folhas do dendrograma indicam o número de elementos do conjunto de treino que ficaram no *cluster* correspondente a cada folha. Indique, justificando, se este é o resultado de um algoritmo de *clustering* aglomerativo (*agglomerative clustering*) ou divisivo (*divisive clustering*)



Pergunta 8 [2 valores] Responda a **uma** das duas perguntas abaixo:

- Num modelo oculto de Markov, a variável oculta pode adoptar um de N estados e emitir um de K símbolos em cada estado. Quantos e quais valores é preciso especificar para instanciar completamente este modelo?
- Quando uma rede neuronal artificial tem muitas camadas, o algoritmo de *backpropagation* torna-se demasiado ineficiente e instável para treinar a rede a partir de um conjunto inicial de pesos gerados aleatoriamente. Descreva uma forma usada em *deep learning* para contornar este problema.

AA, Teste 2, 2015-12-10

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) O primeiro termo, que caracteriza o "bias", que é o menor erro verdadeiro de qualquer hipótese no espaço de hipóteses H. deve ser dominante, uma vez que quanto menor for o bias, menor será o erro no conjunto de treino, e, consequentemente, o erro verdadeiro será também menor.

1b) O segundo termo, que corresponde a uma função do tamanho do espaço de hipóteses e do conjunto de treino que determina a variância do agente de aprendizagem ("learner"), deve ser dominante, uma vez que quanto maior for a variância, maior será a distância entre o erro de treino e o erro verdadeiro dos exemplos, logo, um menor erro de treino não corresponderá a um menor erro verdadeiro.

2) Quando o objectivo é encontrar alguma estrutura nos dados, minimizar o erro de treino não permite encontrar a estrutura, ou seja, cumprir o objectivo em si. Neste caso, seria mais vantajoso aplicar algoritmos de unsupervised learning.

3) Com base na figura, eliminaria o atributo D. Os atributos A e B são os que menos se encontram relacionados com os restantes, e C e D estão fortemente correlacionados entre si e apresentam alguma correlação com os atributos A e B. Todavia o atributo D encontra-se um pouco mais relacionado com estes, sendo que é o atributo que deverá ser eliminado.

4) Para treinar: 1. Atribui-se coeficientes pequenos e aleatórios ou algo que represente a distribuição dos exemplos no problema que estamos a considerar; 2. Para cada exemplo tenta-se obter a Best Matching Unit (BMU); 3. A BMU deve ser ajustada de modo a aproximar-se do ponto e todas as unidades vizinhas desta devem se deslocar na direção do ponto, sendo que as unidades vizinhas mais próximas aproximam-se mais; 4. O coeficiente de aprendizagem de cada unidade diminui a cada iteração para que a malha (estrutura formada pelos neurónios) estabilize.

5) O gráfico X representa o resultado desta projeção, uma vez que é o gráfico com maior dispersão segundo a vertical.

AA, Teste 2, 2015-12-10

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	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○

6a) DBSCAN

D, uma vez que é o único gráfico parcial, ou seja, que não classifica todos os pontos.

6b) Mistura de Gaussianas

C, visto que em cada distribuição é feita um média dos pontos e o seu desvio padrão, sendo por isso a distribuição dos clusters bastante dividida entre "cruzes" e "círculos".

6c) k-means

A, uma vez que é o gráfico com clustering melhor, excluindo o D, para a forma globular dos clusters, e o K-means tem melhor performance com clusters globulares.

6d) Affinity Propagation

B, uma vez que cada ponto procura um representante, que pode ser ele mesmo, para os clusters, agrupando-se os vários pontos num só, que é o representante desse cluster, até não existirem mais pontos disponíveis.
No entanto, na imagem podemos observar pontos de um cluster perto de pontos de outro, o que é normal visto que os pontos apenas procuram o seu representante, que pode estar entre pontos de outros clusters.

7) Este é o resultado de um clustering divisivo, pois, segundo o dendrograma, começou-se com um único cluster e foi-se subdividindo esse em sub-clusters menores.

8)

1. Para instanciar completamente este modelo iríamos necessitar de especificar 3 valores, a matriz N x N, em que N representam os estados do modelo, a matriz N x K, em que K representam os valores desses estados, e as N probabilidades de valores de cada estado.
2. Para contornar o problema do algoritmo de backpropagation não se conseguir adaptar adequadamente à data devido ao excesso de camadas, pois ocorre demasiado overfitting, ao usar deeplearning podemos evitar este problema utilizando técnicas numéricas como diferentes funções de ativação, usar derivadas de ordem superior, boa inicialização dos parâmetros, ou seja, os parâmetros são próximos do resultado final, ou momentum(balanço), em que vamos aumentando a variação dos parâmetros da rede sempre que nessa direção se for reduzindo o erro, ganhando-se "balanço" e alterando cada vez mais depressa os valores nessa mesma direção. Quando se altera a direção, o momentum volta a 0 e o processo repete-se.