

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) Num problema de classificação, se tivermos uma classe de hipóteses \mathcal{H} com um número finito de hipóteses, podemos dizer que o erro verdadeiro ($E(\hat{h})$) da hipótese de menor erro empírico só tem uma probabilidade δ de ultrapassar o valor da seguinte expressão (onde $\hat{E}(h)$ é o erro empírico de uma hipótese h):

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

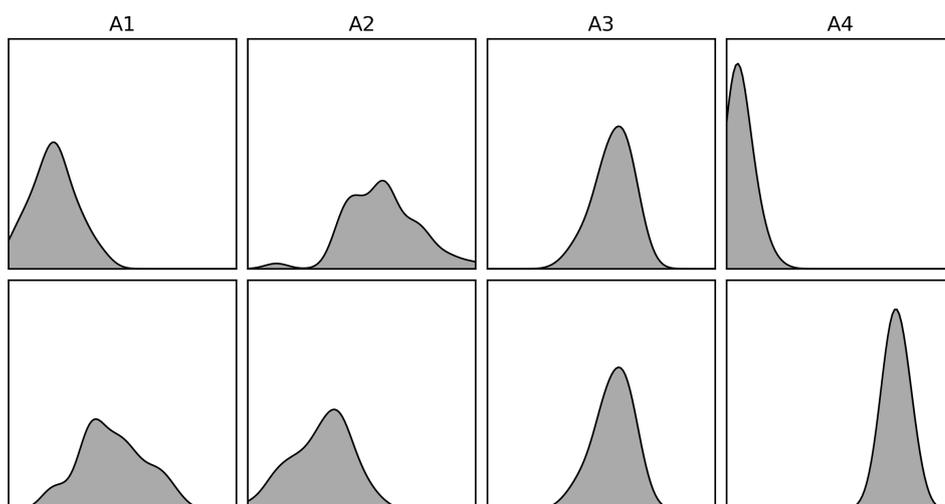
No entanto, se o número de hipóteses em \mathcal{H} for infinito (por exemplo, se houver parâmetros contínuos), não podemos usar o número de hipóteses para estimar o erro de generalização. Nesse caso, precisamos de outra medida da capacidade da classe de hipóteses se ajustar aos pontos a classificar. Qual é a medida usada nesse caso? (Não precisa dizer o nome mas explique o que é).

1.b) A expressão abaixo é uma estimativa da probabilidade do erro verdadeiro da hipótese com menor erro empírico ($E(\hat{h})$) ultrapassar o seu erro empírico ($\hat{E}(\hat{h})$) por mais que um certo valor. Explique porque é que, em geral, modelos de classificação com muitos parâmetros (por exemplo, redes neuronais profundas) têm sucesso com grandes volumes de dados mas não quando o conjunto de dados é pequeno.

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

Pergunta 2 [4 valores]

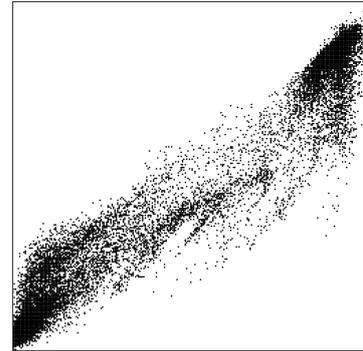
2.a) Num problema de classificação com duas classes e quatro atributos, queremos descartar um dos atributos para reduzir o problema a 3 dimensões. A figura abaixo mostra os gráficos de kernel density estimation da distribuição dos valores de cada atributo (A1 a A4). Todos os gráficos estão na mesma escala. Na fila de cima estão as distribuições para os exemplos da classe 1, na de baixo as distribuições dos exemplos da classe 2. Qual o atributo que descartaria? Justifique a sua resposta.



2.b) Suponha que em vez de seleccionar 3 atributos reduzíamos a dimensionalidade do problema transformando os dados por análise de componente principal (principal component analysis). Qual é a orientação relativa dos vectores correspondentes aos 3 componentes principais e qual é a medida que este processo tenta preservar o mais possível?

Pergunta 3 [5 valores]

Suponha que temos um conjunto de pontos distribuídos como se representa na figura à direita. Estes pontos são exemplo de medidas de valores contínuos que queremos quantizar em 20 pontos representativos desta distribuição. Para cada algoritmo abaixo indique se é ou não é um algoritmo adequado para esta tarefa. Justifique cada resposta considerando o funcionamento do algoritmo respectivo. Se o algoritmo for adequado, indique também os valores dos parâmetros necessários ou como os pode calcular.



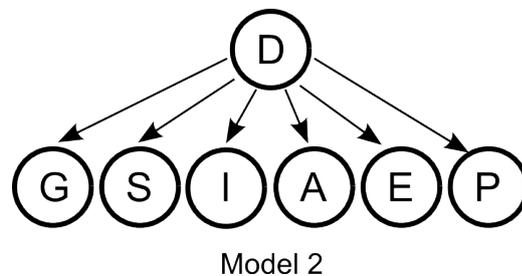
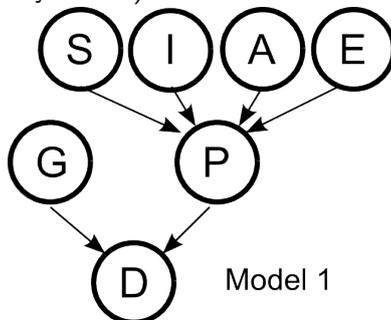
- 3.a) K-Means.
- 3.b) DBSCAN
- 3.c) Affinity Propagation.

Pergunta 4 [3 valores]

Temos um conjunto de 238 fósseis de bivalves. Para cada exemplar fez-se três medições anatómicas da concha e, com base nestes valores, queremos obter uma árvore filogenética que nos permita agrupar os fósseis em espécies, géneros, famílias, numa hierarquia de gupos. Sugira um algoritmo adequado para esta tarefa e explique como esse algoritmo funciona.

Pergunta 5 [3 valores]

Para estudar a importância de factores causais da diabetes, quer-se calcular a distribuição conjunta de probabilidades do estado diabético (D) e os seguintes factores causais: exercício (E), alimentação (A), pressão arterial (P), idade (I), estado geral de saúde (S) e nível de glucose no sangue (G). Considere os dois modelos gráficos abaixo, nos quais as setas representam relações de dependência condicional entre variáveis (redes bayesianas):



5.a) Escreva, para cada modelo, a expressão da distribuição conjunta das variáveis em função das suas dependências condicionais, de acordo com o modelo respectivo.

Não precisa de escrever " $P(D,E,A,P,I,S,G) =$ ". Escreva logo a expressão para cada modelo.

5.b) Suponha que tem 152 questionários preenchidos, para 152 pessoas, 76 das quais são diabéticas e 76 não são diabéticas. Nestes questionários, cada parâmetro E, A, P, I, S, G está indicado com um valor inteiro de 0 a 9 e a variável D, do diagnóstico de diabetes, é binária. Neste caso, qual dos dois modelos devemos usar para estimar a distribuição conjunta de probabilidades? Justifique a sua resposta.

Pergunta 6 [2 valores]

Temos um conjunto de 500 exemplos e assumimos que cada um foi gerado por uma de três distribuições aleatórias gaussianas. Sabemos os valores dos atributos de cada exemplo mas não sabemos qual a distribuição que originou cada exemplo. Explique resumidamente (e sem ser necessário equações) como podemos usar o algoritmo de Expectation-Maximization para obter os parâmetros das três distribuições gaussianas bem como o contributo de cada uma na mistura.

AA-Teste2 2016-12-21

Numero: _____

0	<input type="checkbox"/>				
1	<input type="checkbox"/>				
2	<input type="checkbox"/>				
3	<input type="checkbox"/>				
4	<input type="checkbox"/>				
5	<input type="checkbox"/>				
6	<input type="checkbox"/>				
7	<input type="checkbox"/>				
8	<input type="checkbox"/>				
9	<input type="checkbox"/>				

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 círculo correspondente abaixo.

Nome:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Filas à Frente	<input type="checkbox"/>															
Alunos à Direita	<input type="checkbox"/>															

1a)

1b)

2a)

2b)

3a)

3b)

AA-Teste2 2016-12-21

Numero: _____

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 círculo correspondente abaixo.

Nome:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Filas à Frente	<input type="radio"/>															
Alunos à Direita	<input type="radio"/>															

0	<input type="radio"/>				
1	<input type="radio"/>				
2	<input type="radio"/>				
3	<input type="radio"/>				
4	<input type="radio"/>				
5	<input type="radio"/>				
6	<input type="radio"/>				
7	<input type="radio"/>				
8	<input type="radio"/>				
9	<input type="radio"/>				

3c)

4)

5a)

5b)

6)