

1º Teste de Aprendizagem Automática

3 páginas, 5 perguntas, 2 folhas de resposta. Duração: 2 horas
DI, FCT/UNL, 2 de Maio de 2015

Pergunta 1 [4 valores] A expressão abaixo relaciona o erro de treino com um limite provável para o erro de generalização em função do tamanho do conjunto de treino (N), da dimensão Vapnik–Chervonenkis do classificador (D_{VC}) e da probabilidade do erro de generalização ultrapassar esse limite (δ).

$$P(E_{teste} \leq E_{treino} + S) = 1 - \delta$$

$$S = \sqrt{\frac{D_{VC}(\log(2N/D_{VC}) + 1) - \log(\delta/4)}{N}}$$

Modelo	E_{treino}	S
A	0,12	0,24
B	0,07	0,26
C	0,04	0,40
D	0,21	0,21
E	0,34	0,18
F	0,02	0,49

O quadro à direita mostra, para vários modelos A a F, o respectivo erro de treino e o valor calculado de S. Em todos os casos foi usado o mesmo conjunto de treino com 1000 pontos e o cálculo foi feito para um δ de 0,05.

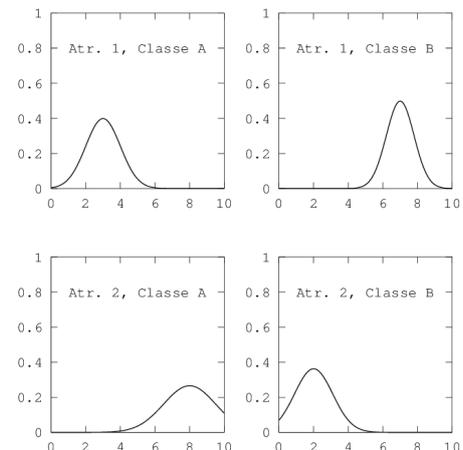
- Indique o modelo (de A a F) que sofre mais de **underfitting**. Justifique a sua resposta.
- Indique o modelo (de A a F) para o qual o risco de **overfitting** é maior. Justifique a sua resposta.
- Indique, justificando, qual dos modelos A a F recomendaria.

Pergunta 2 [3 valores] Foi criado um classificador de Naïve Bayes usando um conjunto de treino com as seguintes características:

- Os pontos estão categorizados em duas classes, A e B;
- Cada ponto tem dois atributos contínuos, Atr.1 e Atr.2, variando de 0 a 10;
- As classes A e B estão representadas na mesma proporção no conjunto de treino, 50% cada uma.

A figura à direita mostra as distribuições de probabilidades dos atributos Atr.1 e Atr.2 em cada classe A e B.

Indique os atributos de um exemplo hipotético que este classificador classificaria na classe A e os atributos de outro exemplo hipotético que este classificador classificaria na classe B. Justifique a sua resposta.



Pergunta 3 [6 valores] Considere o seguinte modelo de classificação onde $g(x)$ é o valor de saída para o exemplo x e os valores w_n são os $M+1$ coeficientes do modelo (contando com w_0 também), sendo M a dimensão dos vectores de entrada:

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

Para obter cada hipótese o modelo é treinado apresentando os exemplos repetidas vezes e por ordem aleatória. De cada vez que se apresenta um exemplo x_t , o modelo é actualizado alterando cada coeficiente w_n da seguinte forma:

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

onde $y(x_t)$ é a classe verdadeira do exemplo x_t , que pode ser 0 ou 1, e x_t^n é o valor do atributo de índice n de x_t , considerando este valor igual a 1 se n for 0. O valor η é uma constante que controla o ritmo da aprendizagem. Depois do treino, considera-se que o exemplo x está na classe 1 se $g(x)$ for maior que 0.5, ou na classe 0 caso contrário ($g(x)$ é um valor entre 0 e 1).

3.a) Desenhe um conjunto de pontos, em duas dimensões, que **pode** ser classificado sem erros com este modelo depois de um treino adequado. Represente a classe 0 por \circ e a classe 1 por \times .

3.b) Desenhe um conjunto de pontos, em duas dimensões e com duas classes, que **não pode** ser classificado sem erros com este modelo, por muito que se tente treiná-lo, mas que possa ser classificado sem erros por outros classificadores mais adequados. Represente a classe 0 por \circ e a classe 1 por \times .

3.c) Alguns conjuntos que não podem ser classificados sem erros com o modelo descrito nesta pergunta podem ser classificados sem erros com uma extensão deste modelo que acrescenta uma camada de funções como $g(x)$ imediatamente antes da função $g(x)$, que assim passa a receber como valores de entrada os valores de saída dessas funções adicionais. Com o algoritmo de treino descrito nesta pergunta, pode-se continuar a melhorar o classificador simplesmente acrescentando mais camadas? Justifique a sua resposta.

Pergunta 4 [4 valores] Para classificar o conjunto de dados representado na sua folha de resposta, foi treinado um classificador obtendo os valores de α que minimizam uma das seguintes expressões (A ou B) onde N é o número de exemplos, y o valor da classe de cada exemplo (representados com um círculo para a classe -1 e uma cruz para a classe +1) e x o vector com as coordenadas de cada ponto.

$$A) \min_{\alpha} \left(\frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N \alpha_n \alpha_m y_n y_m K(x_m, x_n) - \sum_{n=1}^N \alpha_n \right) \quad K(x, z) = e^{\left(\frac{-\|x - z\|^2}{0.4} \right)}$$

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

Sabe-se que foram impostas as seguintes restrições durante a minimização:

$$\alpha_n \geq 0, \quad n = 1, \dots, N \quad \sum_{n=1}^N \alpha_n y_n = 0$$

Infelizmente, não foi registado qual dos dois modelos foi usado para classificar o conjunto.

4.a) Na imagem na sua folha de resposta a linha mais grossa indica a superfície de decisão e as linhas mais finas as margens. Assinale os pontos correspondentes a valores de α maiores que zero (trace um círculo à volta dos pontos que seleccionar).

4.b) Identifique, justificando, qual dos modelos (A ou B) descritos acima foi usado para classificar este conjunto, resultando na fronteira de decisão representada na imagem.

4.c) Suponha que, para diminuir o risco de overfitting, queria permitir que este classificador classificasse incorrectamente alguns pontos do conjunto de treino, simplificando assim a fronteira de decisão. Como faria isto?

Pergunta 5 [3 valores] Num problema de regressão foram criadas 200 réplicas do conjunto de treino por bootstrapping, com 60 pontos cada uma. Dois modelos, A e B, foram instanciados em cada réplica e, assim, calculados os valores X e Y indicados a seguir:

$$X = \frac{1}{60} \sum_{t=1}^{60} [\bar{g}(x_t) - y_t]^2 \quad Y = \frac{1}{60 \times 200} \sum_{t=1}^{60} \sum_{i=1}^{200} [\bar{g}(x_t) - g_i(x_t)]^2$$

Para o modelo A, $X = 0.8$ e $Y = 0.1$ e para o modelo B, $X = 0.02$ e $Y = 12.5$. Indique, justificando, qual dos dois modelos utilizaria para fazer uma regressão pelo método de *bagging* (bootstrap aggregating).

AA 2014/15 Teste 1

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"/>															

Numero: _____

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"/>				

1a)

1b)

1c)

2)

3a) Pode ser classificado sem erros

3b) Não pode ser classificado sem erros

AA 2014/15 Teste 1

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.

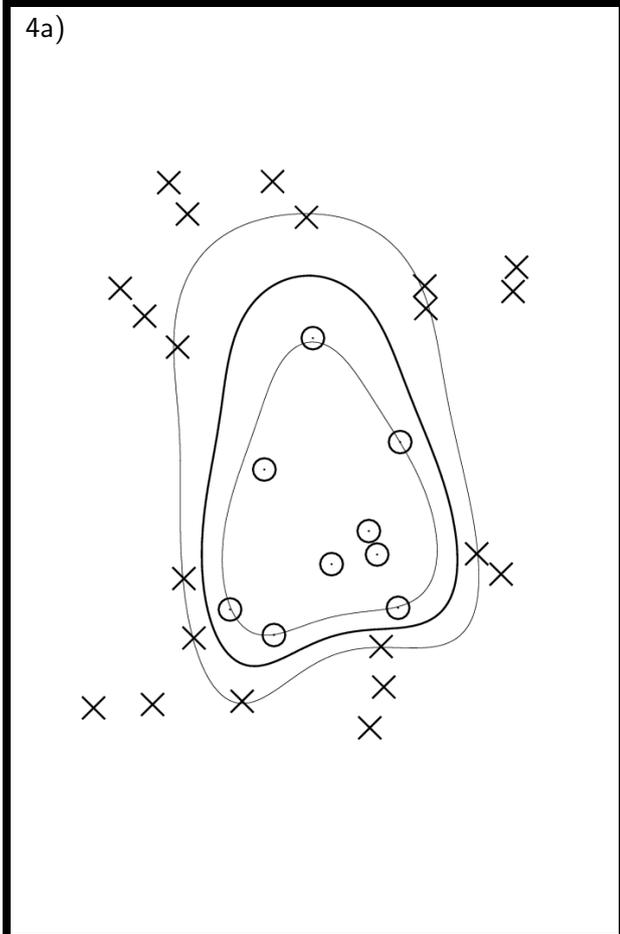
Numero: _____

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

Nome:

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

3c)



4b)

4c)

5