

Atenção: Preencha cuidadosamente a folha de resposta. Cada pergunta tem um número único; responda apenas dentro da caixa respectiva. Pode usar o enunciado como folha de rascunho, mas não escreva fora das caixas na folha de resposta. O teste dura 1 hora e 30 minutos. Bom trabalho.

Grupo 1

Pergunta 1 (2 valores): A partir de um conjunto 20 pontos (x, y) foram criadas 200 réplicas por *bootstrapping*. A cada réplica foi ajustado um polinómio de 3º grau minimizando o erro quadrático e foram calculados os valores A e B, segundo as expressões abaixo, usando o conjunto das 200 réplicas.

$$A = \frac{1}{N} \sum_{t=1}^N [\bar{g}(x_t) - y_t]^2$$

$$B = \frac{1}{NM} \sum_{t=1}^N \sum_{i=1}^M [\bar{g}(x_t) - g_i(x_t)]^2$$

Nestas expressões N indica o número de pontos no conjunto inicial e em cada réplica (20), M o número de réplicas (200), $g_i(x_t)$ o valor de y para o x do ponto t calculado com o polinómio ajustado à réplica i e $\bar{g}(x_t)$ a média dos y calculados para o x do ponto t pelos 200

polinómios, cada um ajustado à sua réplica.

Suponha que o valor de B foi 0.98 e o valor de A foi de 0.03. Indique, justificando, se devemos reduzir o número de parâmetros (usar um polinómio de menor grau), aumentar o número de parâmetros (usar um polinómio de maior grau) ou manter o número de parâmetros.

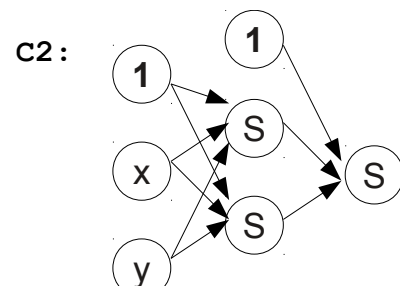
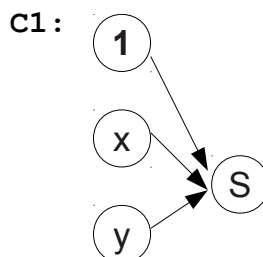
Pergunta 2 (2 valores): Explique porque é que numa regressão polinomial queremos minimizar o quadrado da distância entre a linha definida pelo nosso modelo e os valores reais dos pontos que queremos estimar.

Pergunta 3 (2 valores): Explique porque é que numa regressão logística o que queremos minimizar não é o quadrado da distância entre a linha de decisão definida pelo classificador e os pontos das classes a separar.

Grupo 2

Considere os dois conjuntos de dados A e B, cada um com quatro pontos definidos por duas variáveis, x e y, e uma classe que pode ser 0 ou 1. Considere também os dois classificadores, C1 e C2, em cuja representação o S significa a função sigmóide aplicada ao somatório dos valores de entrada multiplicados cada um pelo seu peso e cada seta indica o peso que multiplica o valor de entrada de onde a seta parte.

A:		B:	
x, y	Classe	x, y	Classe
0, 0	0	0, 0	0
0, 1	1	0, 1	1
1, 1	1	1, 1	0
1, 0	1	1, 0	1



Pergunta 4 (3 valores): Para cada conjunto A e B indique qual ou quais classificadores C1 e C2 podem separar sem erros as duas classes se tiverem os pesos adequados. Justifique a sua resposta explicando porquê.

Pergunta 5 (2 valores): Para que serve aquele 1 nos valores de entrada de cada sigmóide e que diferença faz se é sempre constante?

Grupo 3

Um conjunto de 30 vectores em $\mathbf{R}^2$, em que cada vector é composto por 2 valores numéricos, reparte-se entre duas categorias, $y=0$ e $y=1$, que não podem ser separadas em $\mathbf{R}^2$ por uma linha recta. Há uma função ϕ que mapeia cada vector x em $\mathbf{R}^2$ num vector $\phi(x)$ em $\mathbf{R}^\infty$ – ou seja, com infinitos valores – de tal forma que o produto interno de dois vectores $\phi(x)$ e $\phi(z)$ seja dado pela equação abaixo, à esquerda. Suponha que foi treinada uma *Support Vector Machine* minimizando em função ao vector α a expressão abaixo à direita, sujeita às restrições de que todos os valores de α sejam positivos, que o produto interno de α pelo vector y (das categorias) seja igual a 0 e que nenhum valor de α seja superior a 5.

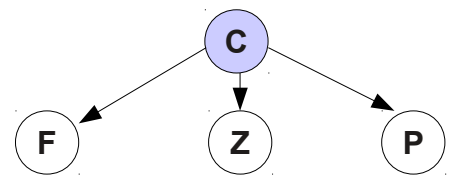
$$\phi(x)^T \phi(z) = e^{\left(\frac{-||x - z||^2}{2\sigma^2} \right)} \quad \min_{\alpha} \left(\frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N \alpha_n \alpha_m y_n y_m \phi(x_n)^T \phi(x_m) - \sum_{n=1}^N \alpha_n \right)$$

Pergunta 6 (2 valores): Quantos valores tem o vector α e como sabemos quais são os vectores de suporte para esta SVM?

Pergunta 7 (2 valores): É possível que a função a minimizar, com as restrições impostas, tenha uma solução mas ainda assim haja vectores no conjunto de treino que sejam incorrectamente classificados pela SVM? Porquê?

Grupo 4

Considere a rede de Bayes representada graficamente à direita. O objectivo é modelar a probabilidade condicional de certos atributos conforme um indivíduo tem ou não cancro no pulmão, para assim obter um classificador que prevê essa doença com base nesses atributos. No diagrama a variável C representa ter ou não ter cancro no pulmão e os atributos a considerar são F , fumar ou não fumar; Z , ter ou não ter cinzeiro em casa; e P , morar numa casa com pelo menos uma pessoa que fume, incluindo o próprio.



Pergunta 8 (2 valores): Este modelo probabilístico descreve adequadamente a relação entre as várias variáveis consideradas? Justifique a sua resposta.

Pergunta 9 (3 valores): Vamos assumir que calculámos os parâmetros do nosso modelo com um conjunto de indivíduos para os quais conhecíamos todas as variáveis, calculando a probabilidade condicional de F , Z e P dado C . Como utilizamos depois esse modelo para prever o se um indivíduo irá ou não ter cancro no pulmão se soubermos, para esse indivíduo, os valores de F , Z e P ?