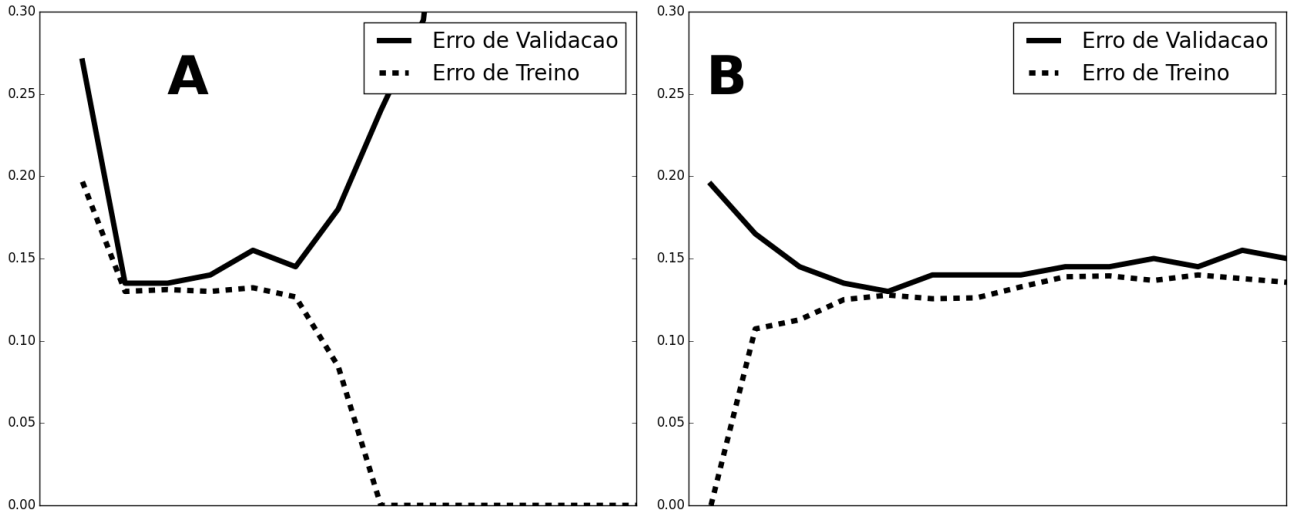
1º Teste de Aprendizagem Automática

3 páginas com 6 perguntas e 2 folhas de resposta. Duração: 2 horas
DI, FCT/UNL, 22 de Outubro de 2015

Pergunta 1 [4 valores] As figuras abaixo mostram o erro de treino e de validação para uma validação cruzada usando 10 *folds*, com o mesmo conjunto de dados. Num caso foi usado um classificador de *k-nearest neighbours* e no outro uma *support vector machine* mas, infelizmente, não foi registado qual gráfico correspondia a qual classificador. O eixo das ordenadas, cujos valores foram omitidos, corresponde ao valor de k no caso do classificador k-NN e ao logaritmo do valor de γ do *kernel* usado para a SVM, cuja expressão é $K(\vec{x}, \vec{z}) = e^{-\gamma||x-z||^2}$



1.a) Indique, justificado, qual dos gráficos (A ou B) corresponde ao classificador k-NN e qual ao classificador SVM.

1.b) Com base nos erros medidos (*Erro de Treino* e *Erro de Validação*), explique como escolheria o melhor valor para o parâmetro do classificador (k no caso do classificador k-NN e γ no classificador SVM).

1.c) Depois de escolher o melhor valor do parâmetro, o erro em que se baseou para essa escolha será um estimador não tendencioso do erro verdadeiro do classificador? Se responder afirmativamente, explique porquê. Se responder pela negativa, explique o que teria sido necessário fazer para obter um estimador não tendencioso do erro verdadeiro do classificador.

Pergunta 2 [4 valores] *Logistic Regression* é um classificador linear que calcula um hiperplano definido por $\vec{w}^T\vec{x} + w_0$ minimizando

$$E(\widetilde{w}) = -\sum_{n=1}^{N} [t_n \ln g_n + (1 - t_n) \ln(1 - g_n)] \qquad g_n = \frac{1}{1 + e^{-(\vec{w}^T \vec{x_n} + w_0)}}$$

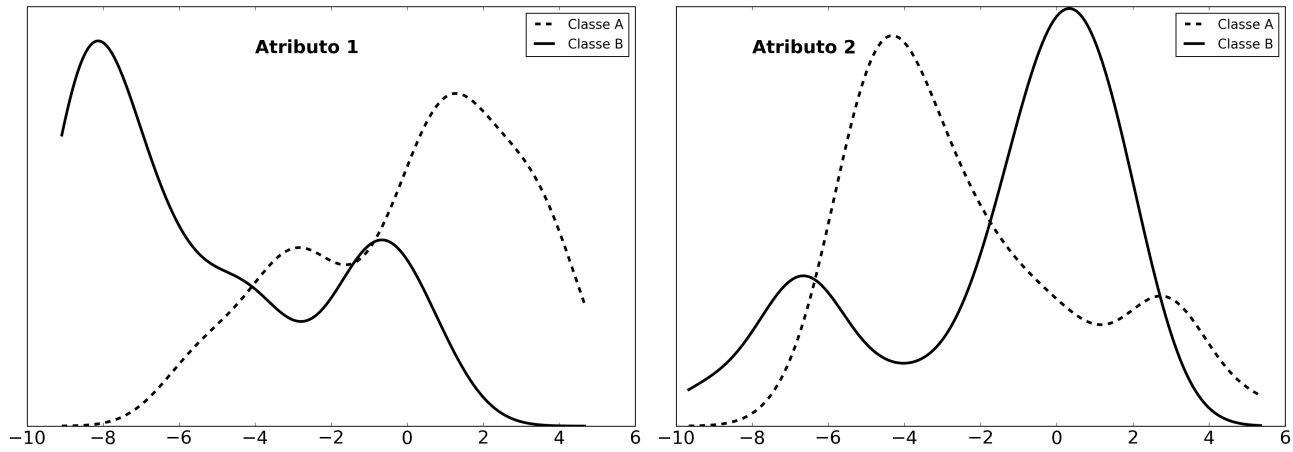
2.a) Suponha que tem um problema de classificação no qual cada exemplo tem dois atributos numéricos contínuos, x_1 e x_2, e as duas classes a distinguir **não são** linearmente separáveis. Explique o que poderia fazer para conseguir separá-las adequadamente com um classificador deste tipo (*Logistic Regression*).

2.b) Depois de optimizar o classificador de *Logistic Regression*, notou-se que cometeu 12 erros de classificação num conjunto com 100 exemplos. No mesmo conjunto, um classificador do tipo *Support Vector Machine* cometeu 10 erros de classificação. Indique, se puder fazê-lo com confiança, qual dos classificadores é o melhor ou, caso contrário, explique porque é que não pode decidir.

Pergunta 3 [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, `Atributo 1` e `Atributo 2`;
- As classes `A` e `B` estão representadas na mesma proporção no conjunto de treino, 50% cada uma.

Os gráficos abaixo mostram as distribuições de probabilidade de cada um dos dois atributos dada cada uma das classes.



3.a) Indique que valores teriam 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.

3.b) Explique porque é que, ao contrário de classificadores discriminantes como *Logistic Regression* e *Support Vector Machine*, o classificador de *Naïve Bayes* permite gerar exemplos artificiais.

Pergunta 4 [3 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:

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

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:

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

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).

4.a) Este classificador pode separar sem erros duas classes que **não sejam** linearmente separáveis? Explique porquê.

4.b) Imagine que tem uma rede de funções destas interligadas, disposta em camadas de forma a que todas as funções numa camada estão ligadas pelos pesos w a todas as funções da camada anterior. Explique como estruturaria a rede e a utilizaria para distinguir K classes com $K > 2$.

Pergunta 5 [3 valores] As figuras na sua folha de resposta representam o resultado do treino de dois classificadores treinados obtendo os valores de α que minimizam a expressão

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

onde N é o número de exemplos, y o valor da classe de cada exemplo (representados com um círculo preenchido a cinzento para a classe 1 e um círculo preenchido a branco para a classe -1) e x o vector com as coordenadas de cada ponto. Foram também impostas as seguintes restrições durante a minimização:

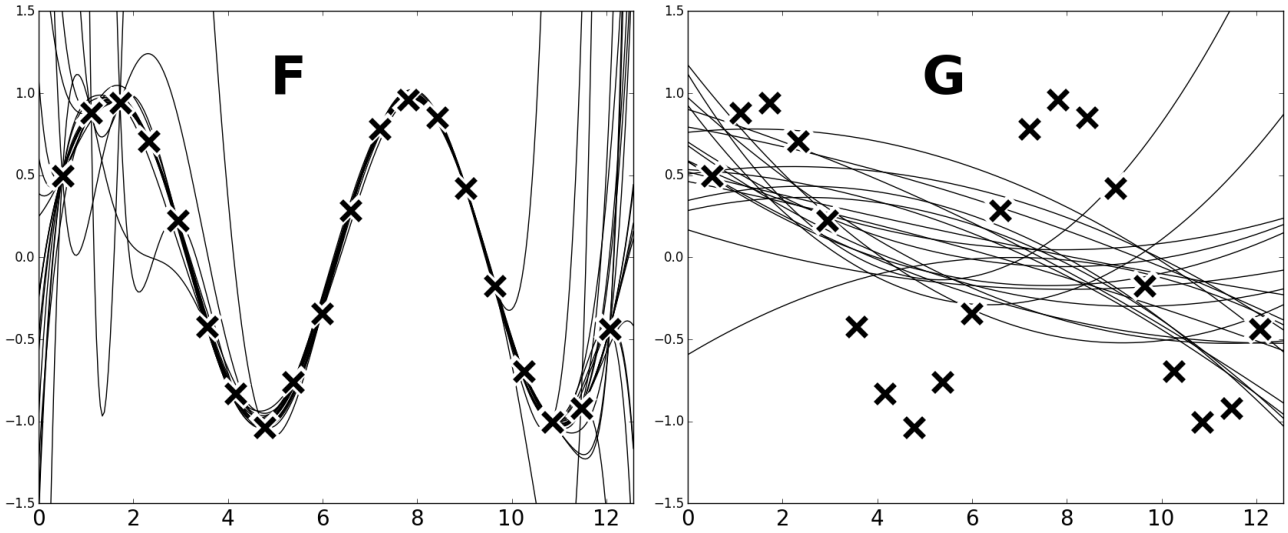
$$0 \leq \alpha_n \leq 10, \quad n = 1, ..., N \qquad\qquad \sum_{n=1}^{N} \alpha_n y_n = 0$$

5.a) Indique qual classificador (A ou B) foi treinado com a função $K(\vec{x_m}, \vec{x_n}) = \vec{x_m}^T \vec{x_n}$ e qual foi treinado com a função $K(\vec{x_m}, \vec{x_n}) = (\vec{x_m}^T \vec{x_n} + 1)^3$, sabendo que uma destas foi usada num dos classificadores e a outra no outro. Justifique a sua resposta.

5.b) Escolha um dos gráficos (A ou B) na sua folha de resposta e assinale, nesse gráfico, **com um círculo**, cada ponto para o qual o valor de α correspondente é **maior que zero** e **menor que 10**.

5.c) Escolha um dos gráficos (A ou B) na sua folha de resposta e assinale, nesse gráfico, **com uma cruz**, cada ponto para o qual o valor de α correspondente é **igual a 10**.

Pergunta 6 [3 valores] Os gráficos F e G abaixo mostram, cada um, 20 instâncias de dois modelos polinomiais de regressão. Cada instância foi obtida treinando o modelo numa réplica do conjunto de treino (representado pelas cruzes) obtida por *bootstrapping*.



6.a) Indique o gráfico (F ou G) e o valor de x (aproximado, um inteiro de 0 a 12) de um ponto que tenha um valor de *bias* maior do que a maioria dos pontos dos gráficos F e G. Justifique a sua resposta.

6.b) Indique o gráfico (F ou G) e o valor de x (aproximado, um inteiro de 0 a 12) de um ponto que tenha um valor de *variance* maior do que a maioria dos pontos dos gráficos F e G. Justifique a sua resposta.

6.c) Qual dos dois modelos, F ou G, escolheria para criar um modelo de regressão pelo método de *bootstrap aggregating*? Justifique a sua resposta.

AA Teste 1 2015-10-22

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) Grafico A pretence ao SVM que usa o gamma na função como exponencial afastando assim o erro de validação do erro de treino

Grafico B pretence ao k-NN que ao aumentar o k nao afectará a diferença (significativamente) entre os dois erros visto que o k so escolhe o numero de "vizinhos" a que se aplica o peso de 1/k.

1b) Escolheria o tivesse um erro de validação mais baixo e que não divergisse muito do erro de treino.
Associaria os parâmetros K e Gama com base nos erros de validação. K e gama seriam os valores para os quais o erro de validação ficasse menor. Não é fiável utilizar o erro de treino porque pode ocorrer overfitting.

1c) Não, seriam necessários dois conjuntos fora do conjunto de treino. Ja temos o conjunto de validaçao, mas tambem é necessario o conjunto de teste para estimar o valor real do erro. O conjunto de validaçao ira servir para escolher a melhor hipotese enquanto que o de teste sera para estimar o erro.
Não, visto que apenas temos o erro de treino e o erro de validação, e para poder descobrir o erro real de forma imparcial precisaríamos de ter o erro de teste, pois os outros dois erros, a partir da média, são tendenciosos visto serem sempre iguais numa forma genérica.

2a) Carregar e standartizar os dados como éfeito anteriormente, de seguida aplicar a função logistica e a função logista de custo para minimizar e obter melhores resultados.

Aplicar a regressão logistica numa maior dimensao, (por exemplo no 3D , ter um plano a separar os pontos em vez de uma linha) adicionando um termo x1 X x2 e aplicando a regressão em 3D. Em seguida, voltamos a projetar o novo conjunto no plano original, verificando se o encaixe da data é o melhor. Caso não seja o melhor encaixe, repetimos o processo até encontrar um melhor, mas tentando evitar o overfitting.

2b) Não será possivel decidir com confiança porque a regressão logistica terá dificuldades onde a a classificação fosse de 0,5 por exemplo. Enquanto que o SVM teria dificuldades se as nao houvesse separabilidade, assim nao podemos decidir com confiança qual dos classificadores é melhor.
Não é possível dizer qual dos classificadores é melhor, visto que apesar de existirem vários testes para comparar classifica dores, como o teste de McNemar, não nos são facultados os dados necessários para o realizar (falta número de acertos).

3a) Atributo 1: Classe A= 2 , Classe B = -8 | Atributo 2: Classe A= -4 , Classe B = 1
Pois os valores para cada classe de cada atributo, submetendo-se ao classificador, este classificaria sempre os maiores valores, visto que o classificador vai estimar os dados computando o arco máximo. Visto estarmos à procura do maior valor para as classes em cada atributo, para a classe A, o classificador devolveria o valor 2 e para a classe B devolveria 1.

3b) Um classificador discriminante prevê a classe de um exemplo a partir da estimativa da probabilidade condicional de um ponto que pertence à classe e dadas as hipóteses. Naïve Bayes é um classificador generativo, pois este primeiro estima a distribuição das probabilidades conjuntas das classes e formula os valores, e só depois prevê a classe a partir dessa probabilidade conjunta.
A razão pela qual este tipo de classificador é chamado generativo é pelo facto desta distribuição da probabilidade conjunta poder ser usada para gerar exemplos artificiais para cada classe.

AA Teste 1 2015-10-22

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) Visto este classificador ser semelhante à regressão logística, se as classes não forem linearmente separáveis existirão sempre erros, mas poderemos tentar corrigi-los ao aplicar a regressão logística numa maior dimensão ao adicionar termos. Em seguida, voltamos a projetar o novo conjunto no plano original, verificando se o encaixe da data é o melhor. Caso não seja o melhor encaixe, repetimos o processo até encontrar um melhor, mas tentando evitar o overfitting.

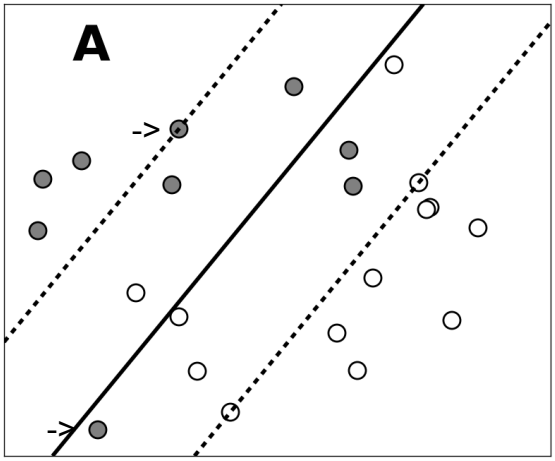
4b) Se o objectivo for distinguir K classes (com K > 2), então seria preciso, em vez de ter um neuron de output, ter K neurons de output, sendo que cada um destes neurons seria treinado para dar sinal de 1 apenas para uma das classes. Por exemplo, para K = 3:

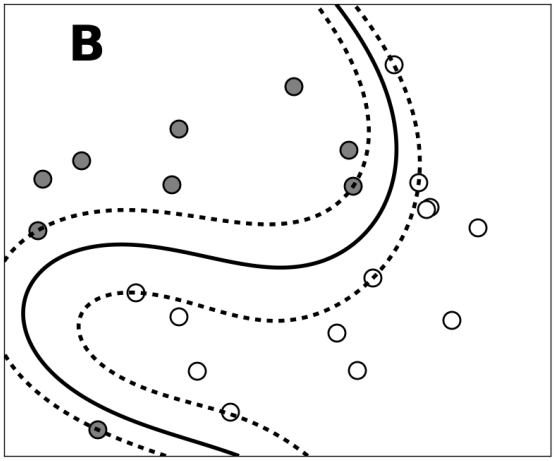
- Output1: 1 quando a classe fosse 1, 0 nas outras classes
- Output2: 1 quando a classe fosse 2, 0 nas outras classes
- Output3: 1 quando a classe fosse 3, 0 nas outras classes

Depois para se classificar um ponto como estando numa das K classes, basta ver qual o neuron de output que tem maior valor para esse ponto.

5a) A primeira função foi utilizada no classificador A e a segunda no classificador B. A primeira função faz uma transformação linear e por isso associamos ao classificador A. A segunda função faz uma transformação para 3 dimensões e por isso associamos ao classificador B.

5b) e 5c)





6a) Bias é a diferença entre a média dos pesos e o peso real. Neste caso, em G, p/x = I, pelo que nenhum dos classificadores prevê corretamente o valor de G (muitas linhas dissociadas)

6b) Variance - alta variância. Diferença entre as várias previsões e a média das mesmas. Neste caso, x=12.

6c)