Teste 2 - Versão 1
04/06/2024

Aprendizagem Automática
90 minutos

Nome: _____

ID: _____

Problema	Valores	Classificação
1	3	
2	4	
3	2.5	
4	2	
5	2	
6	2	
7	3	
8	2	
Total	20.5	

Problema 1 (Escolha Múltipla Geral, 3 valores)

1.) Para cada uma das questões seguintes circula a(s) opção(ões) correta(s).
Nota: respostas parcialmente corretas não serão contabilizadas.

1.1) Utilizar o mesmo conjunto de dados para treinar um modelo e avaliar o seu desempenho resulta numa: (0.5 valores)

a. avaliação pessimista
b. avaliação otimista
c. avaliação imparcial
d. nenhuma das anteriores

1.2) Supõe que vamos treinar um SVM (Support Vectro Machine) utilizando o conjunto de dados de treino abaixo. Quais serão os vetores de suporte? (0.5 valores)

#	x_1	x_2	y
1	-1	2	-
2	-1	1	-
3	1	3	+
4	1	1	-
5	3	3	+

a) 2, 3, 5
b) 1, 3, 4
c) 3, 4
d) 1, 3, 5
e) nenhuma das anteriores

1.3) Quais dos seguintes algoritmos conseguem aprender fronteiras de decisão não lineares? (0.5 valores)

a) Árvore de decisão (com profundidade igual a 5)
b) AdaBoost com múltiplas árvores de decisão (profundidade igual a 1)
c) Regressão Linear
d) SVM (*hard margin*)

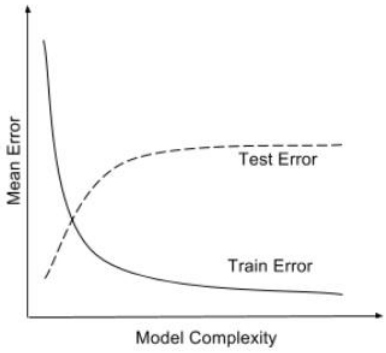
1.4) Quais dos seguintes são verdadeiros sobre os SVMs (*Support Vectro Machines*)? (0.5 valores)

a) Aumentar o hiperparâmetro C tende a diminuir o erro de treino.
b) O SVM *hard margin* é um caso especial do SVM *soft margin* quando o hiperparâmetro C é zero.
c) Aumentar o hiperparâmetro C tende a diminuir a margem.
d) Aumentar o hiperparâmetro C tende a diminuir a sensibilidade a *outliers*.

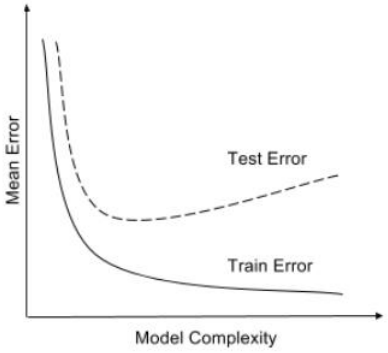
1.5) Considera que treinaste um modelo utilizando um conjunto de dados de treino, D_{train}. Após o treino, procedeste à avaliação do modelo num conjunto de dados de teste independente, D_{test}. Observaste que o erro do modelo ao ser testado com D_{test} é significativamente elevado. Para investigar a causa desse desempenho insatisfatório, decidiste calcular o erro do modelo no conjunto de treino D_{train}. Descobriste, então, que o erro no treino é praticamente nulo. Quais das seguintes opções podem ajudar? (0.5 valores)

a) Aumentar o tamanho de D_{train}
b) Aumentar o tamanho de D_{test}
c) Aumentar a complexidade do modelo
d) Diminuir a complexidade do modelo
e) Concluir que a Aprendizagem de Automática não funciona.

1.6) Imagina que decides desenhar os erros de treino e teste em função da complexidade de um modelo. Com qual das seguintes figuras esperas que o tue gráfico se pareça? (0.5 valores)



(a)



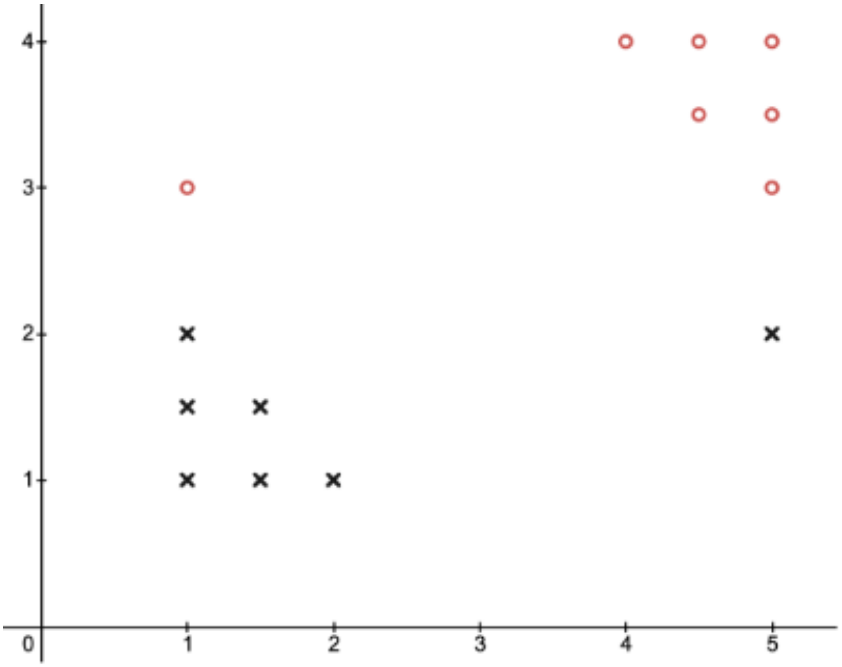
(b)

a) (a)

b) (b)

Problema 2 (SVMs, 4 valores)

2.1) Dados os pontos no gráfico abaixo, desenha e legenda duas linhas: a fronteira de decisão aprendida por um SVM *hard margin* e a fronteira de decisão aprendida por um SVM *soft margin*. (1 Valor)



2.2) Assume que estamos perante um problema multi-classe com 4 classes. Decidimos treinar um SVM “one-versus-one" e outro “one-versus-all".

2.2.1) Quantos classificadores irão ser treinados em cada caso? (0.5 Valores)

2.2.2) Explica sucintamente como a abordagem “one-versus-one" resolve a classificação de uma nova amostra. (1 Valor)

2.3) Ao considerar o valor da variável de *slack* (ξ) de um SVM, como determinamos se um ponto está bem classificado, mal classificado, viola a margem e/ou pode ser um vetor de suporte? As condições a serem analisadas são: (1 Valor)

1. $\xi = 0$

Bem classificado? _____

Violação de margem? _____

Pode ser um vetor de suporte? _____

2. $0 < \xi \leq 1$

Bem classificado? _____

Violação de margem? _____

Pode ser um vetor de suporte? _____

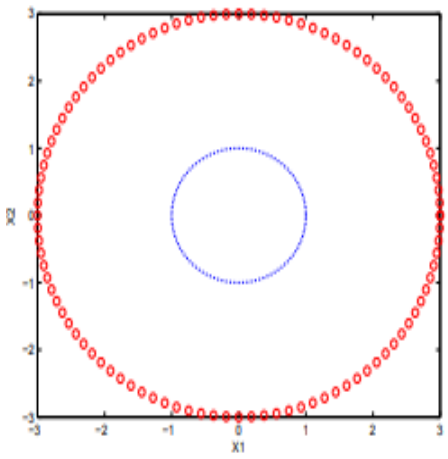
3. $\xi > 1$

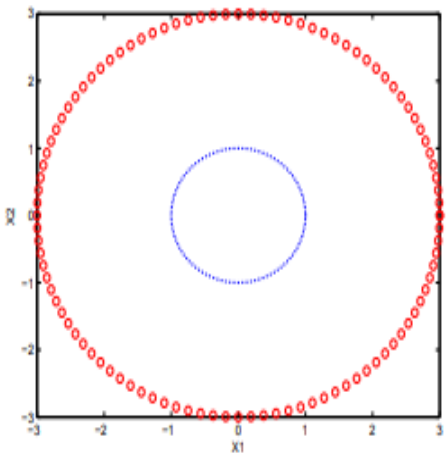
Bem classificado? _____

Violação de margem? _____

Pode ser um vetor de suporte? _____

2.4) Dados os seguintes 2 gráficos, que ilustram um conjunto de dados com duas classes. Desenha a fronteira de decisão ao treinar um classificador SVM com kernels linear e RBF (*radial basis function*), respectivamente. (0.5 Valores)





Problema 3 (Perceptron, 2.5 valores)

3) Suponha que lhe é dado o seguinte conjunto de dados:

x_1	x_2	y
0	0	0
0	1	0
1	0	0
1	1	1

Sabendo que queremos treinar um perceptron com os dados fornecidos e que:

- Os pesos iniciais foram aleatoriamente definidos: w_1=0.9 e w_2=0.9.
- O limiar de ativação (*activation threshold*) foi definido como θ=0.5.
- O *learning rate* ficou definido como α=0.5.

3.1) Qual será o vetor de pesos atualizado (w_1 , w_2) depois de passarmos o exemplo 1 pelo algoritmo do perceptron? Apresenta todos os cálculos efetuados. (1 Valor)

3.2) Qual será o vetor de pesos atualizado (w_1 , w_2) depois de passarmos o exemplo 2 pelo algoritmo do perceptron? Apresenta todos os cálculos efetuados. (1.5 Valores)

Problema 4 (Ensembles, 2 valores)

4.1) Identifique qual técnica (bagging, boosting ou stacking) é descrita em cada uma das seguintes afirmações. (0.5 valores cada)

4.1.1) Neste método várias amostras bootstrap dos dados de treino são criadas e um modelo é treinado para cada amostra. _____

4.1.2) Cada novo modelo é treinado para corrigir os erros dos modelos anteriores, ajustando os pesos das amostras. _____

4.1.3) Este método combina as previsões de vários modelos base diferentes, utilizando um meta-modelo para fazer a predição final. _____

4.1.4) É eficaz na redução de overfitting porque combina as previsões de múltiplos modelos treinados em diferentes subconjuntos de dados. _____

Problema 5 (Otimização de Hiperparâmetros, 2 valores)

5.1) Em aprendizagem automática qual é a diferença entre parâmetros e hiperparâmetros de um modelo? (1 Valor)

5.2) "Com o aumento do número de hiperparâmetros a testar, o custo da *grid search* aumenta exponencialmente.". Comenta esta afirmação. (1 Valor)

Problema 6 (*Imbalanced Learning*, 2 valores)

6.1) Imagina que estás a trabalhar para uma startup de tecnologia que recebe milhares de candidaturas de emprego todos os dias. Um dia decides treinar um modelo de aprendizagem automática para automatizar todo o processo de contratação. O modelo classifica automaticamente currículos dos candidatos e rejeita ou envia ofertas de emprego. Qual das seguintes medidas é mais importante para o teu modelo? Explica. (1 Valor)

$$\text{Recall} = \frac{True\ Positives}{Total\ Positive\ Samples}$$

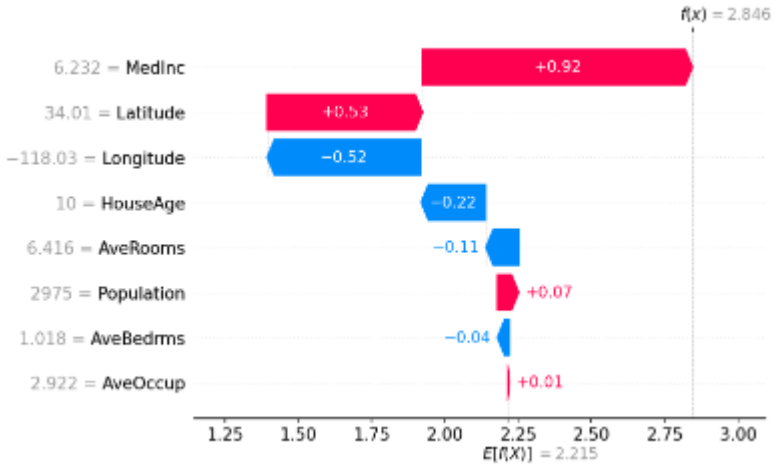
$$\text{Precision} = \frac{True\ Positives}{Total\ Predicted\ Positive\ Samples}$$

6.2) Sucintamente explica o que é o método SMOTE (*Synthetic Minority Over-sampling Technique)* e como ele funciona? (1 Valor)

Problema 7 (XAI, 3 valores)

7.1) Indica dois tipos de modelos de aprendizagem automática que são inerentemente interpretáveis. Sucintamente explica como e porquê. (1.5 Valores)

7.2) Imagina que treinamos um modelo para prever o valor de uma casa. Após o treino do modelo, recorremos à abordagem SHAP (shapley additive explanations) para obter alguma interpretabilidade do modelo. Obtivemos o seguinte gráfico:



Que conclusões consegues tirar do gráfico? (1.5 Valores)

Problema 8 (Viés e Variância, 2 Valores)

Foram estudados vários métodos para controlar o overfitting para diversos classificadores. Abaixo, encontram-se listados vários classificadores e ações que podem afetar o seu bias e variância. Indique (circulando) como o bias e a variância mudam em resposta à ação: (0.5 Valores cada)

8.1) Aumentar a profundidade máxima numa arvore de decisão:

Bias	**Variância**
Diminuir	Diminuir
Aumentar	Aumentar
Permanecer inalterado	Permanecer inalterado

8.2) Aumentar muito o C num SVM:

Bias	**Variância**
Diminuir	Diminuir
Aumentar	Aumentar
Permanecer inalterado	Permanecer inalterado

8.3) Remover alguns exemplos de treino (não incluindo vetores de suporte) num SVM:

Bias	**Variância**
Diminuir	Diminuir
Aumentar	Aumentar
Permanecer inalterado	Permanecer inalterado

8.4) Aumentar o número de árvores de decisão numa *random forest:*

Bias	**Variância**
Diminuir	Diminuir
Aumentar	Aumentar
Permanecer inalterado	Permanecer inalterado