MAE 5905: Introdução à Ciência de Dados

Lista 4. Primeiro Semestre de 2020. Entregar 28/05/2020.

1. Considere o conjunto de dados **Weekly** do pacote **ISLR**, contendo 1.089 retornos semanais de ações de 1990 a 2010.

(a) Calcule algumas medidas numéricas dos dados, como média, variância, quantis etc. Faça alguns gráficos para sumarizar os dados (use, por exemplo, o pacote **astsa**).

(b) Use o conjunto todo de dados e ajuste uma regressão logística, com **Direction** (up and down) como variável resposta e variável defasada **Lag1** como preditora. Comente os resultados.

(c) repita (b), agora tendo como preditores **Lag1** e **Lag2**. Comente.

(d) Ajuste uma regressão logística usando como período de treinamento os dados de 1990 a 2008, com **Lag2** como preditor. Obtenha a matriz de confusão e a taxa de erro de classificação para o período de teste, 2009-2010.

(e) repita (d) usando KNN, com K=1.

(f) Qual método fornece os melhores resultados?

2. Considere o conjunto de dados **Auto** do pacote ISLR.

(a) Crie uma variável binária, **mpg1**, que é igual a 1 se **mpg** for maior que sua mediana, e **mpg1** igual a zero, se **mpg** for menor que sua mediana. (Use a função data.frame () para criar um conjunto de dados contendo **mpg1** e as outras variáveis do conjunto **Auto**).

(b) Faça gráficos para investigar a associação entre **mpg1** e as outras variáveis (e.g., *draftsman display, boxplots*). Divida os dados em conjunto de treinamento e de teste.

(c) Use análise discriminante linear de Fisher para prever **mpg1** usando os preditores que você acha que sejam mais associadas com ela, usando o item (b). Qual a taxa de erros do conjunto teste?

(d) Use KNN, com vários valores de K, e determine a taxa de erros do conjunto teste. Qual valor de K é melhor nesse caso?

(e) Qual classificador você julga que é melhor?

3. Dadas as observações abaixo ($n = 7,\ p = 2$), faça um gráfico no plano $(X1, X_2)$.

(a) Use o classificador de margem máxima e dê a regra de classificação. Obtenha a equação do hiperplano separador.

(b) No gráfico, indique a margem e os vetores suporte para o CMM.

(c) Calcule o valor da margem. O que acontece com o classificador se modificarmos um pouco a sétima observação?

Obs.	X_1	X_2	Y
1	3	4	Azul
2	2	2	Azul
3	4	4	Azul
4	1	4	Azul
5	2	1	Vermelho
6	4	3	Vermelho
7	4	1	Vermelho

4. Simule um conjunto de dados com $n = 500$ e $p = 2$, tal que as observações pertençam a duas classes com uma fronteira de decisão não linear. Por exemplo, você pode usar:

$>$ x1=runif(500)-0.5
$>$ x2=runif(500)-0.5
$> y = 1 * (x1 \wedge 2 - x2 \wedge 2 > 0)$.

(a) Faça um gráfico das observações, com símbolos (ou cores) de acordo com cada classe.

(b) Separe os dados em conjunto de treinamento e de teste. Obtenha o classificador de margem máxima, tendo X_1 e X_2 com preditores. Obtenha as previsões para o conjunto de teste e a acurácia do classificador.

(c) Obtenha o classificador de margem flexível, tendo X_1 e X_2 com preditores. Obtenha as previsões para o conjunto de teste e a taxa de erros de classificação.

(d) Obtenha o classificador de margem não linear, usando um kernel apropriado. Calcule a taxa de erros de classificação.

(e) Compare os dois classificadores.