MAE 5905: Introdução à Ciência de Dados

Lista 3. Primeiro Semestre de 2023. Entregar 06/06/2023.

1. Para o conjunto de dados Iris, use somente o comprimento de pétalas (X_1) e comprimento de sépalas (X_2) como preditores e a variável resposta Y= espécie(Setosa, Versicolor, Virgínica). Construa uma árvore para classificação. Escreva com detalhes as regiões no plano e faça o gráfico da árvore e das regiões, usando o pacote tree. Obtenha a taxa de erro de classificação.

2. Considere o conjunto de dados **rehabcardio**, sendo preditores X_1 =HDL, X_2=LDL, X_3 =Trigl, X_4=Glicose e X_5=Peso e resposta Y=Diabete (presente=1, ausente=0). Utilize um subconjunto em que as amostras têm todas as medidas completas. Construa árvores usando bagging e floresta aleatória. Usando a taxa de erro de classificação, escolha o melhor classificador.

3. Considere as variáveis Altura e Idade da Tabela 12.1 do Capítulo 12 (Análise de Agrupamentos):

(a) Obtenha os agrupamentos usando o método hierárquico, até um ponto que você considere adequado, usando a distância Euclidiana e o método do centróide. Obtenha o dendrograma correspondente.

(b) Refaça o item (a) usando a distância L_1 (Manhattan).

(c) Use o algoritmo K-médias, com $K = 3$ para obter os grupos para os mesmos dados. Comente o resultado. Qual é o centróide de cada grupo?

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.