MAE 5905: Introdução à Ciência de Dados

Primeiro Semestre de 2026. Lista 3. Entregar 25/05/2026.

1. Para o conjunto de dados **Iris**, use somente o comprimento de pétalas (X_1) e o comprimento de sépalas (X_2) como preditores e a variável resposta Y=espécie (Setosa, Versicolor, Virgínica). Cosntrua uma árvore para classificação. Escreva com detalhes as regiões no plano e faça o gráfico da árvore e das regiões, usando um pacote de sua preferência. 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. O conjunto de dados **Auto** do pacote **ISLR** contém as seguintes variáveis:

mpg: miles per gallon
cylinders: Number of cylinders between 4 and 8
displacement: Engine displacement (cubic inches)
horsepower: Engine horsepower
weight: Vehicle weight (lbs.)
acceleration: Time to accelerate from 0 to 60 mph (sec.)
year: Model year (modulo 100)
origin: Origin of car (1. American, 2. European, 3. Japanese)
name: Vehicle name

(a) Divida os dados em conjunto de treinamento(S) e conjunto de teste (T).

(b) Ajuste um modelo aos dados de S tendo **horsepower** como preditor e **mpg** como resposta. Obtenha os EMQ de treinamento e faça o diagnóstico do modelo. O que você nota no gráfico dos resíduos contra valores ajustados? Obtenha o EQM de teste.

(c) Agora inclua (**horsepower**)^2 no modelo e proceda como no item (b). Qual modelo você escolheria? Justifique.

(d) Ajuste um modelo de regressão **ridge** aos dados de S, tendo **mpg** com resposta e **displacement**, **horsepower**, **weight** e **acceleration** como preditores, com λ escolhido por VC. Obtenha o EQM de teste.

(e) Ajuste um modelo de regressão **lasso** e proceda como em (d). Quais coeficientes foram zerados?

(f) Comente sobre os resultados obtidos em (d) e (e), baseados no R^2 e EQM.