

## MAE 5905: Introdução à Ciência de Dados

Lista 2. Primeiro Semestre de 2023. Entregar 12/05/2023.

1. (a) Use a função `rnorm()` (simula valores de uma distribuição normal) do R para gerar um preditor  $X$  com  $n = 100$  observações, bem como um erro  $\varepsilon$  também de comprimento 100.

(b) Simule um vetor de resposta  $Y$ , de comprimento  $n = 100$  de acordo com o modelo

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon,$$

na qual os parâmetros  $\beta_i$  são constantes, de sua escolha.

(c) Considere o modelo de (b), agora com os  $\beta_i$  e  $\varepsilon$  desconhecidos,  $X$  como em (a)  $Y$  como em (b). Qual seria o melhor modelo usando  $R^2$  ajustado e BIC?

(d) Para o modelo como em (c), obtenha os estimadores ridge e lasso. Use VC para selecionar o valor ótimo de  $\lambda$ .

2. 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?

3. 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?