Lista 1: Conceitos Fundamentais

April 8, 2022

1 Regressão Linear e o Método dos Mínimos Quadrados

1. Considere um modelo linear simples:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j,$$

e a métrica de erro dada pelo erro quadrático médio:

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2,$$

demonstre que esse modelo possui solução geral, dada pela equação normal:

$$\hat{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

2. O método dos mínimos quadrados pode ser ampliado para uma função linear nos parâmetros :

$$f(X) = \beta_0 + \sum_{j=1}^{p} \sum_{i=1}^{k} \phi_i(X_j) \beta_j.$$

Demonstre que, sob as mesmas considerações, a equação normal pode ser escrita em termos da pseudo-inversa de Moore-Penrose:

$$\hat{\beta} = (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \boldsymbol{y}$$

3. Defina: underfitting e overfitting

4. Uma das possíveis soluções para o overfitting é utilizar a regularização da função erro:

$$RSS(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda \sum_{j=1}^{p} \beta_j^n,$$

O caso $n = 2$ é conhecido como regressão Ridge. Obtenha a equação normal para esse caso. Sob quais condições é possível determinar a equação normal de uma equação regularizada? O caso $n = 1$ é conhecido como Regressão Lasso, é possível obter uma equação normal?

5. O método dos mínimos quadrados pode ser interpretado por um prisma geométrico. Demonstre.

6. O problema da regressão linear pode ser formulado em termos de outras funções erro, por exemplo:

   - Soma dos erros absolutos: $MAE(\beta) = \sum_{i=1}^{N} |y_i - f(x_i)|$
   - Entropia Cruzada: $CEE(\beta) = -\sum_{i=1}^{N} (y_i \log(f(x_i)) + (1 - y_i) log(1 - f(x_i)))$

   A solução para os coeficientes é unívoca? Elabore um argumento justificando a escolha no erro quadrático médio como medida do erro para o cálculo dos coeficientes.

2 Construção de um Modelo de Regressão Linear

1. Esse problema envolve o uso do dataset "Auto" e o módulo "linearmodels" da biblioteca sklearn:

   - Use o método LinearRegression() para produzir um modelo linear usando a variável "mpg" como variável resposta e o *horsepower* como variável preditora. Faça uma análise do resultado obtido, conceitual e estatística.
   - Faça um plot da resposta vs o preditor;
   - Faça um plot diagnóstico do predito vs a resposta.

2. Agora, utilize o mesmo módulo para formular um modelo de regressão multilinear com o mesmo dataset.

   - Produza um plot bivariado incluindo todas as variáveis no dataset.
   - Produza uma matriz de correlação das variáveis e avalie quais variáveis são correlacionadas entre si e quais são mais ou menos correlacionadas com o alvo.
   - Existe multicolinearidade nos dados?
   - Repita todos os passos do exercício anterior e construa um modelo de regressão linear.
   - Em alguns casos, transformações lineares nos dados podem trazer uma melhora substancial no resultado. Repita os passos anteriores usando:
     - $\log(X)$
     - $\sqrt{x}$
     - X^2
     - A Transformação de Box-Cox

3. Repita o exercício anterior usando o dataset Carseats