Professor: Cayan Portela
e-mail: cayan.saavedra@ceub.edu.br

Otimização Aplicada á Ciência de Dados (2023/2)

Lista 1

---

1. Um modelo de modelo de regressão linear simples (1) pode ser generalizado em sua forma matricial (2) possibilitando mais de uma covariável. Dessa forma, podemos expressar os modelos da seguinte maneira:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{1}$$

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{\epsilon} \tag{2}$$

em que ${}_n\boldsymbol{Y}_1$ é o vetor coluna da variável de interrese; ${}_n\boldsymbol{X}_p$ é a matriz de *input*; ${}_p\boldsymbol{\beta}_1$ é a matriz de parâmetros (a serem estimados) e ${}_n\boldsymbol{\epsilon}_1$ é o vetor de erros (que representa a diferença entre o valor real e o valor predito). A representação matricial do modelo é dada por:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{12} \\ \vdots & \vdots \\ 1 & x_{1n} \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Os dados insurance (disponíveis no nosso repositório do github) possuem as seguintes informações referente aos benificários de um certo plano de saúde:

- age: idade do titular.
- sex: gênero do titular (masculino, feminino).
- bmi: índice de massa corporal ($\frac{kg}{altura^2}$, idealmente entre 18.5 e 24.9)
- children: número de dependentes cobertos pelo plano.
- smkoer: fumante(sim ou não)
- region: região de residência.
- charges: despesas médicas pagas pelo plano de saúde.

Estime um modelo de regressão linear para predição de despesas médicas de usuários, considerando as seguinte informações: idade, gênero, número de dependentes e fumante/não-fumante.

(a) Qual o objetivo em estimar o vetor de parâmetros $\hat{\boldsymbol{\beta}}$?

(b) Encontre o vetor de parâmetros $\hat{\boldsymbol{\beta}}$, utilizando as seguintes abordagens (indique os cálculos/pseudo-códigos utilizados partindo da equação (2)):

   i. Solução analítica matricial.
   ii. Decomposisção QR
   iii. Gradiente descendente.

(c) Calcule o Erro Quadrático Médio na base de teste.

2. Considerando o seguinte modelo de regressão logística, ajuste um modelo de classificação para os dados bank_customer (disponível no github).

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i + \epsilon_i \tag{3}$$

Utilize como covariáveis as seguintes informações: gender, age, credit_card, credit_score.

- Considerando a variável target, qual o objetivo dessa modelagem?
- Indique explicitamente a função de perda a ser maximizada, e sua derivação (como foi obtida).
- Encontre o vetor de parâmetros $\hat{\beta}$ utilizando as seguintes abordagens:
  - Newton-Raphson
  - Gradiente Descendente

3. Utilizando o método de Newton-Raphson, encontre $\sqrt{10}$ com uma precisão de 10 casas decimais. (Dica: $\sqrt{10}$ é a raiz de qual função?)

4. Utilizando o método de Newton-Raphson, encontre a raíz de $f(x) = x^4 - 2x^3 - 2$ que está entre 1 e 2.

5. Calcule os 10 primeiros passos do algoritmo de Newton-Raphson para $3x^{1/3} = 0$ utilizando $x_0 = 0.1$

6. Utilize o método de Newton-Raphson para encontrar a raíz da equação $x^4 - 5x^3 + 9x + 3 = 0$ no intervalo [4, 6]