Capítulo 9

Regressão Linear

João Gabriel Junqueira da Silva, Manoel Veríssimo dos Santos Neto e Rafael Rodrigues Silva

Abstract

This report presents a theoretical summary of Chapter 9 - Linear Regression from the book, focusing on the essential mathematical foundations for data science. The chapter addresses the linear regression algorithm, a widely used statistical technique to model the relationship between a dependent variable and one or more independent variables. It explores the mathematical foundation and the main techniques associated with the algorithm, highlighting everything from problem formulation to parameter estimation. In simple linear regression, we seek to minimize the sum of squared residuals to find the coefficients that best fit the data. In multiple linear regression, we use matrix notation to handle multiple predictors, deriving the estimators through normal equations.

Resumo

Este relatório apresenta um resumo teórico do Capítulo 9 - Regressão Linear do livro Mathematics for Machine Learning, focando nas bases matemáticas essenciais para a ciência de dados. O capítulo aborda o algoritmo da regressão linear é uma técnica estatística amplamente utilizada para modelar a relação entre uma variável dependente e uma ou mais variáveis independentes. É explorado a fundamentação matemática e as principais técnicas associadas ao algoritmo, destacando desde a formulação do problema até as estimativas de parâmetros. Na regressão linear simples, buscamos minimizar a soma dos quadrados dos resíduos para encontrar os coeficientes que melhor ajustam os dados. Na regressão múltipla, utilizamos a notação matricial para lidar com múltiplos preditores, derivando os estimadores por meio das equações normais.

9.1. Formulação do Problema

A formulação do problema de regressão linear envolve definir a relação entre uma variável dependente y e uma ou mais variáveis independentes X. O objetivo é encontrar os coeficientes de um modelo linear que melhor se ajuste aos dados observados [Deisenroth et al. 2020] e [Montgomery et al. 2012].

9.1.1. Modelo de Regressão Linear Simples

Na regressão linear simples, modelamos a relação entre uma única variável independente x e a variável dependente y usando a equação:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Onde:

- **y** é a variável dependente (resposta).
- **x** é a variável independente (preditor).
- $\boldsymbol{\beta_0}$ é o intercepto da reta.
- $\boldsymbol{\beta_1}$ é o coeficiente angular (slope).
- $\boldsymbol{\varepsilon}$ é o termo de erro aleatório, que segue uma distribuição normal $\boldsymbol{\varepsilon \sim N(0, \sigma^2)}$.

9.1.2. Modelo de Regressão Linear Múltipla

Na regressão linear múltipla, modelamos a relação entre várias variáveis independentes $\boldsymbol{x_1, x_2, \ldots, x_p}$ e a variável dependente y usando a equação [Seber and Lee 2012]:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

Ou, em notação matricial:

$$\boldsymbol{y = X\beta + \varepsilon}$$

Onde:

- **y** é o vetor das observações da variável dependente (n x 1).
- **X** é a matriz das observações das variáveis independentes (n x (p+1)), com a primeira coluna sendo composta por 1s (para o intercepto).
- $\boldsymbol{\beta}$ é o vetor dos coeficientes do modelo ((p+1) x 1).
- $\boldsymbol{\varepsilon}$ é o vetor dos termos de erro (n x 1).

9.1.3. Suposições do Modelo de Regressão Linear

Para que o modelo de regressão linear produza estimativas válidas e úteis, algumas suposições devem ser satisfeitas:

- **Linearidade:** A relação entre a variável dependente e as variáveis independentes é linear.

- **Independência:** As observações são independentes entre si.
- **Homocedasticidade:** A variância dos erros é constante para todos os valores das variáveis independentes.
- **Normalidade dos Erros:** Os erros do modelo são normalmente distribuídos com média zero.

9.1.4. Função de Custo

A função de custo na regressão linear é a soma dos quadrados dos resíduos (SSE):

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Onde:

- $\boldsymbol{y_i}$ são os valores observados.
- $\boldsymbol{\hat{y}_i}$ são os valores preditos pelo modelo.

9.1.5. Solução pelo Método dos Mínimos Quadrados

Para encontrar os coeficientes que minimizam a soma dos quadrados dos resíduos, usamos o método dos mínimos quadrados. Em notação matricial, a solução é obtida resolvendo a equação normal:

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

Essa equação nos fornece os estimadores dos coeficientes que melhor se ajustam aos dados, minimizando a soma dos quadrados dos resíduos.

9.1.6. Interpretação dos Coeficientes

- $\boldsymbol{\beta_0}$ (intercepto): Representa o valor esperado de y quando todas as variáveis independentes são iguais a zero.
- $\boldsymbol{\beta_j}$ (coeficiente de $\boldsymbol{x_j}$): Representa a mudança esperada em $\boldsymbol{y}$ para uma unidade de mudança em $\boldsymbol{x_j}$, mantendo todas as outras variáveis constantes.

9.1.7. Avaliação do Modelo

- $\boldsymbol{R^2}$ **(coeficiente de determinação):** Mede a proporção da variância total em y explicada pelo modelo. É calculado como:

$$\boldsymbol{R^2} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

- **Erro padrão dos resíduos (SE):** Fornece uma medida da precisão das previsões do modelo. É calculado como:

$$\boldsymbol{SE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p-1}}$$

Referências

[Deisenroth et al. 2020] Deisenroth, M. P., Faisal, A. A., and Ong, C. S. (2020). *Mathematics for Machine Learning*. Cambridge University Press.

[Montgomery et al. 2012] Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. John Wiley & Sons, Hoboken, NJ, 5th edition.

[Seber and Lee 2012] Seber, G. A. and Lee, A. J. (2012). *Linear Regression Analysis*. John Wiley & Sons, Hoboken, NJ, 2nd edition.