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:

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

9.2. Estimativa de Parâmetros

A estimativa de parâmetros na regressão linear envolve encontrar os coeficientes β que melhor descrevem a relação entre as variáveis independentes X e a variável dependente y. O método mais comum para essa estimativa é o método dos mínimos quadrados ordinários (OLS - Ordinary Least Squares).

Modelo de Regressão Linear Simples

Para a regressão linear simples, o modelo é:

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

Queremos encontrar os valores de β_0 e β_1 que minimizam a soma dos quadrados dos resíduos (SSE):

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

Onde y_i são os valores observados da variável dependente e $\hat{y}_i$ são os valores preditos pelo modelo:

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

Estimadores dos Coeficientes

Os estimadores dos coeficientes β_0 e β_1 são obtidos resolvendo-se o sistema de equações normais derivadas da minimização da SSE.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Derivação dos Estimadores

Vamos derivar as fórmulas para os estimadores usando o método dos mínimos quadrados. Primeiro reescrevemos a SSE:

$$\text{SSE} = \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$$

Para minimizar a SSE, derivamos em relação a β_0 e β_1 e igualamos a zero:

$$\frac{\partial \text{SSE}}{\partial \beta_0} = 0$$

$$\frac{\partial \text{SSE}}{\partial \beta_1} = 0$$

Resolvendo essas duas equações simultaneamente, obtemos as fórmulas para $\hat{\beta}_0$ e $\hat{\beta}_1$.

Modelo de Regressão Linear Múltipla

Para a regressão linear múltipla, o modelo é:

$$y = X\beta + \varepsilon$$

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).
- β é o vetor dos coeficientes do modelo ((p+1) x 1).
- ε é o vetor dos termos de erro (n x 1).

Estimadores dos Coeficientes

Os coeficientes são estimados minimizando a soma dos quadrados dos resíduos em notação matricial:

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

Para encontrar os estimadores, derivamos a SSE em relação a β e igualamos a zero:

$$\frac{\partial \text{SSE}}{\partial \beta} = 0$$

Resolvendo para β, obtemos:

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

Propriedades dos Estimadores

Os estimadores obtidos pelo método dos mínimos quadrados possuem algumas propriedades importantes:

- **Não Viesados:** Os estimadores são não viesados, ou seja, $E(\hat{\beta}) = \beta$.
- **Variância Mínima:** Entre todos os estimadores lineares não viesados, os estimadores de mínimos quadrados têm a menor variância.
- **Distribuição:** Se os erros ε são normalmente distribuídos, então os estimadores $\hat{\beta}$ também são normalmente distribuídos.

9.2.1. Estimativa de Máxima Verossimilhança

Para encontrar os parâmetros que maximizam a verossimilhança, utilizamos a técnica de máxima verossimilhança:

$$\theta_{\text{ML}} \in \arg\max_{\theta} p(\mathscr{Y} \mid \mathscr{X}, \theta)$$

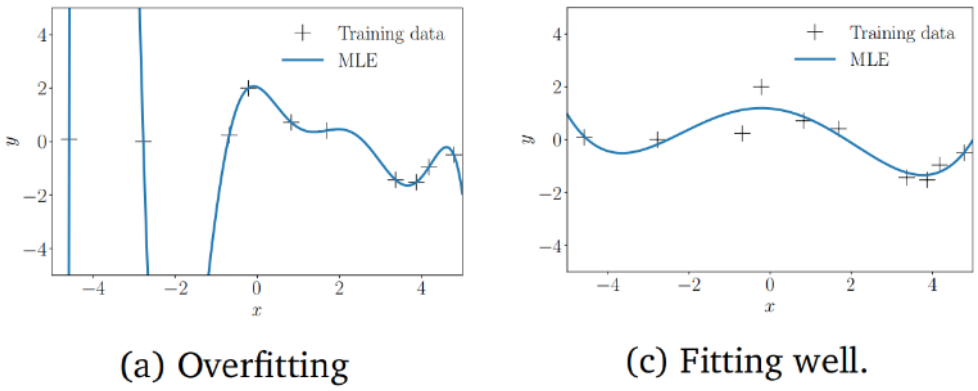
A solução analítica para a estimativa de máxima verossimilhança é:

$$\theta_{\text{ML}} = (X^\top X)^{-1} X^\top y$$

onde X é a matriz de design que contém os vetores de entrada x_n.

9.2.2. Overfitting em Regressão Linear

Overfitting ocorre quando o modelo se ajusta muito bem aos dados de treinamento, mas não generaliza bem para novos dados. Isso geralmente acontece quando o modelo é muito complexo.



(a) Overfitting (c) Fitting well.

Figura 9.1: Exemplo de modelos com e sem overfitting respectivamente

Técnicas para Evitar Overfitting

- **Regularização:** Adicionar um termo de penalização na função de custo como na estimativa MAP.
- **Validação Cruzada:** Utilizar técnicas de validação cruzada para avaliar o desempenho do modelo em diferentes subconjuntos dos dados.
- **Seleção de Modelo:** Escolher o modelo mais simples que ainda explica bem os dados utilizando critérios como o erro quadrático médio (MSE) ou o erro quadrático médio da raiz (RMSE).

9.2.3. Estimativa Máxima a Posteriori (MAP)

A estimativa de máxima verossimilhança é propensa a overfitting. Para mitigar esse efeito, podemos impor uma distribuição priori $p(\theta)$ sobre os parâmetros. A distribuição priori

codifica os valores plausíveis dos parâmetros antes de observar os dados. A fórmula da posteriori é dada por:

$$p(\theta|\mathscr{X},\mathscr{Y}) = \frac{p(\mathscr{Y}|\mathscr{X},\theta)p(\theta)}{p(\mathscr{Y}|\mathscr{X})}$$

O vetor de parâmetros θ_{MAP} que maximiza a posteriori é a estimativa MAP. Para encontrar a estimativa MAP, minimizamos a função:

$$\theta_{\text{MAP}} \in \arg\min_{\theta} \{-\log p(\mathscr{Y}|\mathscr{X},\theta) - \log p(\theta)\}$$

9.2.4. Estimativa MAP como Regularização

Além de impor uma distribuição priori sobre os parâmetros, é possível mitigar o overfitting penalizando a magnitude dos parâmetros por meio da regularização. Na regularização de mínimos quadrados, consideramos a função de perda:

$$\|\mathscr{Y} - \Phi\theta\|_2^2 + \lambda\|\theta\|_2^2$$

O termo de regularização λ controla a rigidez da regularização. A minimização desta função de perda fornece a solução:

$$\theta_{\text{RLS}} = (\Phi^\top\Phi + \lambda I)^{-1}\Phi^\top\mathscr{Y}$$

9.3. Regressão Linear Bayesiana

Na regressão linear Bayesiana, não buscamos uma estimativa pontual dos parâmetros, mas sim a distribuição posterior completa dos parâmetros. A posteriori dos parâmetros é dada por:

$$p(\theta|\mathscr{X},\mathscr{Y}) \propto p(\mathscr{Y}|\mathscr{X},\theta)p(\theta)$$

Onde S_n e m_n são calculados como:

$$S_n = (S_0^{-1} + \sigma^{-2}X^TX)^{-1}$$

$$m_n = S_n(S_0^{-1}m_0 + \sigma^{-2}X^Ty)$$

9.3.1. Modelo

Na regressão linear Bayesiana, consideramos o modelo:

$$p(\theta) = \mathscr{N}(m_0, S_0)$$

$$p(y \mid x,\theta) = \mathscr{N}(y \mid \phi^\top(x)\theta, \sigma^2)$$

Isso transforma o vetor de parâmetros em uma variável aleatória e permite escrever o modelo probabilístico completo:

$$p(y, \theta \mid x) = p(y \mid x, \theta) p(\theta)$$

9.3.2. Previsões Anteriores

Para fazer previsões em x^*, integramos θ e obtemos:

$$p(y_* \mid x_*) = \int p(y_* \mid x_*, \theta) p(\theta) d\theta = \mathbb{E}_\theta [p(y_* \mid x_*, \theta)]$$

Com uma priori Gaussiana em θ, a distribuição preditiva é Gaussiana:

$$p(y_* \mid x_*) = \mathcal{N}(\phi^\top(x_*) m_0, \phi^\top(x_*) S_0 \phi(x_*) + \sigma^2)$$

9.3.3. Distribuição Posterior

Dado um conjunto de treinamento, a posteriori sobre os parâmetros é:

$$p(\theta \mid \mathcal{X}, \mathcal{Y}) = \frac{p(\mathcal{Y} \mid \mathcal{X}, \theta) p(\theta)}{p(\mathcal{Y} \mid \mathcal{X})}$$

Com uma priori Gaussiana conjugada, a posteriori é:

$$p(\theta \mid \mathcal{X}, \mathcal{Y}) = \mathcal{N}(\theta \mid m_N, S_N)$$

$$S_N = \left(S_0^{-1} + \sigma^{-2} \Phi^\top \Phi \right)^{-1}$$

$$m_N = S_N \left(S_0^{-1} m_0 + \sigma^{-2} \Phi^\top \mathcal{Y} \right)$$

9.3.4. Previsões Posteriores

Para prever valores sem ruído, usamos a média e a variância da distribuição posterior:

$$\mathbb{E}[f(x_*) \mid \mathcal{X}, \mathcal{Y}] = \mathbb{E}_\theta \left[\phi^\top(x_*) \theta \mid \mathcal{X}, \mathcal{Y} \right] = \phi^\top(x_*) m_N$$

$$\mathrm{Var}_\theta [f(x_*) \mid \mathcal{X}, \mathcal{Y}] = \phi^\top(x_*) \mathrm{Var}_\theta [\theta \mid \mathcal{X}, \mathcal{Y}] \phi(x_*) = \phi^\top(x_*) S_N \phi(x_*)$$

A média preditiva coincide com a previsão usando a estimativa MAP, e a variância preditiva reflete a incerteza posterior dos parâmetros.

9.3.5. Calculando a Probabilidade Marginal

A probabilidade marginal é importante para a seleção de modelos Bayesianos. A probabilidade marginal é dada por:

$$p(\mathscr{Y} \mid \mathscr{X}) = \int p(\mathscr{Y} \mid \mathscr{X}, \theta) p(\theta) d\theta$$

Com a priori Gaussiana conjugada, a probabilidade marginal é Gaussiana. A média e a variância da probabilidade marginal são:

$$\mathbb{E}[\mathscr{Y} \mid \mathscr{X}] = \mathbb{E}_{\theta,\varepsilon}[\mathscr{X}\theta + \varepsilon] = \mathscr{X}\mathbb{E}_{\theta}[\theta] = \mathscr{X} m_0$$

$$\text{Cov}[\mathscr{Y} \mid \mathscr{X}] = \text{Cov}_{\theta,\varepsilon}[\mathscr{X}\theta + \varepsilon] = \text{Cov}_{\theta}[\mathscr{X}\theta] + \sigma^2 I = \mathscr{X}\,\text{Cov}_{\theta}[\theta]\mathscr{X}^{\top} + \sigma^2 I = \mathscr{X} S_0 \mathscr{X}^{\top} + \sigma^2 I$$

9.4. Máxima Probabilidade como Projeção Ortogonal

A interpretação da estimativa de máxima verossimilhança na regressão linear como uma projeção ortogonal oferece uma compreensão geométrica poderosa do problema de ajuste de linha.

Fundamentos da Regressão Linear em Notação Matricial

O modelo de regressão linear pode ser escrito como:

$$y = X\beta + \varepsilon$$

Onde:

- y é um vetor $n \times 1$ das observações da variável dependente.
- X é uma matriz $n \times (p+1)$ das observações das variáveis independentes incluindo uma coluna de 1s para o intercepto.
- β é um vetor $(p+1) \times 1$ dos coeficientes do modelo.
- ε é um vetor $n \times 1$ dos termos de erro.

Interpretação Geométrica: Projeção Ortogonal

A interpretação geométrica considera os vetores no espaço euclidiano $\mathbb{R}^n$. A matriz X define um subespaço S de $\mathbb{R}^n$. O vetor y pode ser decomposto em duas componentes ortogonais: uma componente $\hat{y}$ no subespaço S e uma componente de erro ε ortogonal a S:

$$y = \hat{y} + \varepsilon$$

Onde $\hat{y}$ é a projeção ortogonal de y no subespaço S gerado por X:

$$\hat{y} = X\hat{\beta}$$

Projeção Ortogonal

A projeção ortogonal $\mathbf{P}$ do vetor y no subespaço gerado por X é dada por:

$$\mathbf{P} = X(X^T X)^{-1} X^T$$

Portanto:

$$\hat{y} = \mathbf{P}y$$

Propriedades da Projeção Ortogonal

Algumas propriedades importantes da projeção ortogonal incluem:

- $\mathbf{P}$ é uma matriz idempotente: $\mathbf{PP} = \mathbf{P}$.
- $\mathbf{P}$ é uma matriz simétrica: $\mathbf{P}^T = \mathbf{P}$.
- O vetor de resíduos $\mathbf{e} = y - \hat{y}$ é ortogonal ao subespaço gerado por X.

9.5. Conclusão

Neste trabalho, exploramos a formulação, estimativa e avaliação de modelos de regressão linear, tanto no contexto simples quanto no múltiplo. Abordamos as suposições fundamentais que garantem a validade dos modelos, como linearidade, independência, homocedasticidade e normalidade dos erros, e discutimos a função de custo baseada na soma dos quadrados dos resíduos (SSE).

Através do método dos mínimos quadrados ordinários (OLS), derivamos as fórmulas para os estimadores dos coeficientes, demonstrando a importância da matriz de design X e a matriz dos coeficientes β. Discutimos também as propriedades desejáveis dos estimadores, como a não-viesabilidade e a variância mínima.

Além disso, exploramos as técnicas para evitar overfitting, como a regularização e a validação cruzada, essenciais para garantir a generalização do modelo. Analisamos a abordagem bayesiana na regressão linear, que oferece uma compreensão probabilística mais robusta e integra a incerteza nas previsões.

A interpretação geométrica da máxima verossimilhança como projeção ortogonal no espaço euclidiano proporciona uma visão intuitiva e poderosa sobre o ajuste dos modelos de regressão linear, destacando a decomposição ortogonal do vetor de respostas e as propriedades das projeções.

Em síntese, este estudo reforça a importância da regressão linear como uma ferramenta fundamental em análise de dados e modelagem estatística. A compreensão profunda dos princípios, métodos e técnicas associadas à regressão linear é crucial para a aplicação eficaz em diversos contextos, desde pesquisa acadêmica até soluções práticas em engenharia e ciências sociais.

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.