Exercício

Previsão com Regressão Linear – parte 1

No exercício de hoje você deverá fazer uma regressão linear para criar um modelo de aprendizado de máquina capaz de prever o pesso de uma pessoa a partir de sua altura (problema fictício). Os dados serão criados com os comandos a seguir, onde X é a altura e Y é o peso de cada pessoa:

```
import numpy as np
# Exemplo de dados
X = np.array([1.47, 1.50, 1.52, 1.55, 1.57, 1.60, 1.63, 1.65, 1.68, 1.70,
1.73, 1.75, 1.78, 1.80, 1.83])
y = np.array([52.21, 53.12, 54.48, 55.84, 57.20, 58.57, 59.93, 61.29,
63.11, 64.47, 66.28, 68.10, 69.92, 72.19, 74.46])
```

Você não vai utilizar bibliotecas já existentes para fazer a regressão linear, mas vai implementar o cálculo dos coeficientes.

Passo a passo para calcular a regressão linear de forma manual

A equação da regressão linear é:

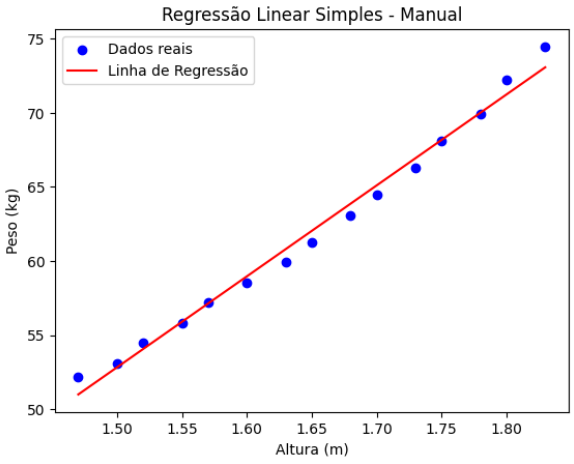
$$y = \beta_0 + \beta_1 x$$

Os coeficientes β_0 (intercepto) e β_1 (inclinação) podem ser calculados pelas fórmulas:

$$\beta_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$\beta_0 = \frac{\sum y - \beta_1 \sum x}{n}$$

Crie uma função para prever os valores utilizando os coeficientes e trace a reta de regressão em um gráfico como o mostrado abaixo:



Calcule o RSE e o R^2 para o mnodelo no conjunto de teste. Deverá ser entregue um relatório em PDF, mostrando o gráfico e o valor de RSE e R^2.

Erro Padrão Residual (RSE)

O **RSE** é uma medida da precisão do modelo. Ele representa o desvio padrão dos resíduos (diferença entre os valores reais e os valores preditos). Quanto menor o RSE, melhor o modelo se ajusta aos dados.

A fórmula para calcular o **RSE** é:

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

Onde:

- y_i são os valores reais,
- $\hat{y}_i$ são os valores preditos pelo modelo,
- n é o número de observações.

Coeficiente de Determinação (R^2)

O R^2 mede a proporção da variabilidade da variável dependente explicada pelo modelo. É dado por:

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

Onde:

- y_i são os valores reais,
- $\hat{y}_i$ são os valores preditos pelo modelo,
- n é o número de observações.
- $\overline{y}$ é a média dos valores reais.

Previsão com Regressão Linear – parte 2

Agora vamos fazer a mesma coisa mas com um problema real. Vamos utilizar o dataset **tips**, que contém dados sobre gorjetas e pode ser usado para prever o valor da gorjeta com base em variáveis como total da conta. O banco de dados pode ser carregado como abaixo:

```
import seaborn as sns
from sklearn.model_selection import train_test_split

# Carregar o dataset de gorjetas
df = sns.load_dataset('tips')

# Mostrar as primeiras linhas do DataFrame
print(df.head())

# Selecionar a variável independente (total_bill) e a variável dependente
(tip)
X = df[['total_bill']] # Total da conta
y = df['tip'] # Gorjeta

# Dividir os dados em treino e teste
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)
```

Da mesma forma, calcule os coeficientes da regressão, e plote um gráfico mostrando os dados de treinamento, os dados de teste e a previsão para os dados de teste. Calcule o RSE e o R^2 para o modelo no conjunto de teste. Deverá ser entregue no mesmo relatório anterior, o gráfico e o valor de RSE e R^2 para este problema real. E por fim, preveja quanto o garçon irá ganhar de gorjeta se o total da conta for de 80U$.