Machine learning: estudos de regressão

Regressão linear, regressão polinomial

REGRESSÃO LINEAR

Variável Dependente (ou Resposta):

A variável dependente é a principal variável de interesse em um estudo ou análise. É a variável que você está tentando entender, explicar ou prever. Em um modelo estatístico, a variável dependente é aquela que você está tentando "explicar" ou "prever" com base em outras informações. Ela é muitas vezes representada pela letra "Y" na equação de um modelo de regressão.

Por exemplo, se estivermos estudando o impacto do tempo de estudo (variável independente) no desempenho em um exame (variável dependente), a variável dependente seria o resultado do exame. Queremos entender como o tempo de estudo afeta o desempenho no exame.

Variáveis Independentes (ou Explicativas ou Preditivas):

As variáveis independentes, por outro lado, são aquelas que você acredita que podem ter influência sobre a variável dependente. São as variáveis que são usadas para explicar ou prever a variação na variável dependente. No contexto de um modelo de regressão, as variáveis independentes são representadas por "X" na equação do modelo.

No exemplo mencionado anteriormente, o tempo de estudo seria uma variável independente, pois estamos considerando que o tempo de estudo pode influenciar o desempenho no exame. Podemos também ter outras variáveis independentes, como a quantidade de sono na noite anterior ao exame, o número de aulas assistidas, entre outras. Todas essas variáveis independentes são usadas para explicar o desempenho no exame (variável dependente.

Regressão linear

Y = a + bX

- onde Y é a variável dependente
- X é a variável independente
- a é o intercepto
- b é o coeficiente da variável independente.

Tempo de Estudo (X)	Desempenho no ENEM (Y)
1 hora	60 pontos
2 horas	70 pontos
3 horas	80 pontos
4 horas	90 pontos
5 horas	100 pontos

Nesse caso devemos começar desta forma:

```
Primeiro, calcule a média de X (tempo de estudo) e Y (desempenho no
exame):

Média de X = (1 + 2 + 3 + 4 + 5) / 5 = 3 horas
Média de Y = (60 + 70 + 80 + 90 + 100) / 5 = 80 pontos
```

Agora, vamos calcular os valores de “a” e “b”. Usamos as fórmulas:

b (coeficiente da variável independente) = Σ((X—$\bar{X}$) * (Y—$\bar{Y}$)) / Σ((X—$\bar{X}$)²) a (intercepto) = $\bar{Y}$—b * $\bar{X}$

- Σ é a soma.
- $\bar{X}$ é a média de X (3 horas).
- $\bar{Y}$ é a média de Y (80 pontos).

Cálculo de b

```
b = [(1 - 3) * (60 - 80) + (2 - 3) * (70 - 80) + (3 - 3) * (80 - 80)
+ (4 - 3) * (90 - 80) + (5 - 3) * (100 - 80)] / [(1 - 3)^2 + (2 - 3)^2
+ (3 - 3)^2 + (4 - 3)^2 + (5 - 3)^2]

b = [(-40) + (-10) + (0) + 10 + 40] / [4 + 1 + 0 + 1 + 4] = 0.5
```

cálculo de a

```
a = 80 - 0.5 * 3 = 80 - 1.5 = 78.5
```

Portanto, a equação de regressão linear simples para este exemplo é:

```
Y = 78.5 + 0.5X
```

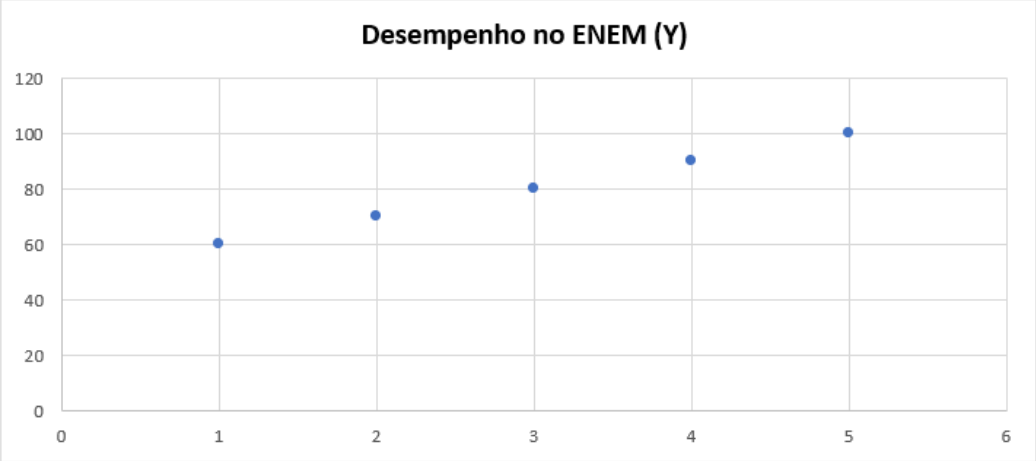
Agora você pode usar essa equação para prever o desempenho no exame com base no tempo de estudo. Por exemplo, se alguém estuda por 6 horas (X = 6), a previsão do desempenho no exame seria:

```
Y = 78.5 + 0.5 * 6 = 78.5 + 3 = 81.5 pontos.
```

Criação de gráfico no Excel

- Selecione os dados de tempo de estudo (X) e desempenho no exame (Y) na planilha.
- Clique com o botão direito nos pontos de dados e escolha a opção “Inserir gráfico de dispersão”.

Tempo de Estudo (X) em horas	Desempenho no ENEM (Y)
1	60
2	70
3	80
4	90
5	100



REGRESSÃO POLIMONIAL

Vamos realizar uma regressão polinomial para entender como a quantidade de açúcar afeta o crescimento das plantas.

Como escolher

Regressão Linear

- Use a regressão linear quando você acredita que a relação entre as variáveis é aproximadamente linear, ou seja, os pontos de dados parecem seguir uma tendência linear.
- Uma regressão linear é apropriada quando você espera uma mudança constante na variável dependente (Y) em resposta a uma mudança constante na variável independente (X).
- É útil para relacionamentos simples, onde uma variável depende de forma direta da outra sem curvas acentuadas.

Regressão Polinomial

- Use a regressão polinomial quando você observa que a relação entre as variáveis não é linear e os pontos de dados seguem uma tendência curvilínea. Por exemplo, uma curva em forma de U, ou uma relação cúbica.
- A regressão polinomial é apropriada quando você espera uma mudança não linear na variável dependente em resposta a uma mudança na variável independente.
- Pode ser útil para modelar situações em que a relação entre as variáveis tem curvas ou reviravoltas.

Para determinar qual tipo de regressão é mais adequado, é importante visualizar seus dados primeiro, geralmente por meio de um gráfico de dispersão. Isso permitirá que você avalie a forma dos dados e identifique qualquer padrão que possa sugerir uma relação linear ou não linear.

```
Quantidade de Açúcar (X) | Crescimento da Planta (Y)
-------------------------|-------------------------
            1 grama      |          5 centímetros
            2 gramas     |         12 centímetros
            3 gramas     |         25 centímetros
            4 gramas     |         42 centímetros
            5 gramas     |         65 centímetros
```

Aqui está uma visualização gráfica

Quantidade de Açúcar em gramas (X)	Crescimento da Planta em centimetros (Y)
1	5
2	12
3	25
4	42
5	65

Vamos usar um polinômio de segundo grau (grau 2), o que significa que nossa equação de regressão será um polinômio de segundo grau:

Y=a0+a1X+a2X2

Para fazer isso, você pode usar software estatístico, como Python com a biblioteca NumPy. Você também pode fazer isso manualmente, mas é mais trabalhoso. Vou usar Python para encontrar os coeficientes do polinômio:

```
import numpy as np

# Dados
X = np.array([1, 2, 3, 4, 5])
Y = np.array([5, 12, 25, 42, 65])

# Ajustando o polinômio de segundo grau
coefficients = np.polyfit(X, Y, 2)

# Obtendo os coeficientes a0, a1 e a2
a2, a1, a0 = coefficients

# Exibindo os coeficientes
print(f"Coeficiente a2: {a2}")
print(f"Coeficiente a1: {a1}")
print(f"Coeficiente a0: {a0}")
```

Aqui está como você pode realizar uma regressão polinomial manualmente:

Calcule as somas e produtos das variáveis:

- Soma de X (ΣX)
- Soma de Y (ΣY)
- Soma de X^2 (ΣX^2)
- Soma de X*Y (ΣXY)
- Soma de X^3 (ΣX^3), se você estiver ajustando um polinômio de grau 3.
- Soma de X^4 (ΣX^4), se você estiver ajustando um polinômio de grau 4.

Use as fórmulas para calcular os coeficientes do polinômio para a0, a1 e a2.

Com os coeficientes calculados, você pode criar a equação do polinômio, como mencionado anteriormente, e usá-la para fazer previsões.