Regressão Linear Múltipla (Parte 1)

Nas aula anteriores, aprendemos a fazer regressão linear com uma **única variável**:

Área da casa [m²] (x)	Custo em R$ (y)
32	51.000
149	265.000
78	110.000
...	...
220	315.000

Variável 1 (característica 1): Área da casa

Modelo Utilizado:

$$f_{w,b} = wx + b$$

Onde estamos e para onde vamos?

Agora estudaremos regressão linear com **múltiplas variáveis**:

Área da casa [m²] (x_1)	Número de quartos (x_2)	Andares (x_3)	Idade [anos] (x_4)	Custo em R$ (y)
32	1	1	2	51.000
149	3	2	10	265.000
78	2	1	30	110.000
...	...	...	...	...
220	4	2	5	315.000

Variável 1 (característica 1): Área da casa
Variável 2 (característica 2): Número de quartos
Variável 3 (característica 3): Número de andares
Variável 4 (característica 4): idade da construção

Levando em conta mais características, é esperado que o modelo seja capaz de prever melhor o valor de uma casa?

Regressão Linear Múltipla

Área da casa [m²] (x_1)	Número de quartos (x_2)	Andares (x_3)	Idade [anos] (x_4)	Custo em R$ (y)
32	1	1	2	51.000
149	3	2	10	265.000
78	2	1	30	110.000
...	...	...	...	...
220	4	2	5	315.000

Notação:

x_j = j-ésima característica $\rightarrow$ $j = 1, 2, \cdots, 4$

n = número total de características $\rightarrow$ $n = 4$

$\overrightarrow{x}^{(i)}$ = características do i-ésimo exemplo de treinamento $\rightarrow$ $\overrightarrow{x}^{(2)} = \begin{bmatrix} 149 & 3 & 2 & 10 \end{bmatrix}$

$\overrightarrow{x}_j^{(i)}$ = valor da característica j do i-ésimo exemplo de treinamento $\rightarrow$ $\overrightarrow{x}_2^{(2)} = 3$

Observação:

Para simplificar a notação, vamos tratar a sobre-barra como um elemento opcional de notação, tal que $\overrightarrow{x}_j^{(2)} = x_j^{(2)}$, por exemplo. $\rightarrow$ (serve apenas para enfatizar que trata-se de um vetor).

Pergunta

Seja o conjunto de dados abaixo. Quanto vale $\vec{x}_4^{(3)}$?

Área da casa [m²] (x_1)	Número de quartos (x_2)	Andares (x_3)	Idade [anos] (x_4)	Custo em R$ (y)
32	1	1	2	51.000
149	3	2	10	265.000
78	2	1	30	110.000
...	...	...	...	...
220	4	2	5	315.000

Antes, na regressão linear com uma única variável, tínhamos o seguinte modelo:

$$f_{w,b}(x) = wx + b$$

Agora, na regressão linear com múltiplas variáveis, teremos:

$$f_{w,b}(x) = w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + b$$

Exemplo:

$$f_{w,b}(x) = 0.1x_1 + 4x_2 + 10x_3 - 2x_4 + 80 \quad \rightarrow \quad \text{valor da casa em milhares de R\$}$$

x_1: Área da casa
x_2: Número de quartos
x_3: Número de andares
x_4: idade da construção

Olhando com detalhes o modelo

Área da casa [m²] (x_1)	Número de quartos (x_2)	Andares (x_3)	Idade [anos] (x_4)	Custo em R$ (y)
32	1	1	2	51.000
149	3	2	10	265.000
78	2	1	30	110.000
...	...	...	...	...
220	4	2	5	315.000

O que significam os coeficientes do modelo?

$$f_{w,b}(x) = 0.1x_1 + 4x_2 + 10x_3 - 2x_4 + 80 \quad \rightarrow \quad \text{valor da casa em milhares de R\$}$$

- Note que cada m^2 adiciona R$ 100 ao valor da casa
- Cada quarto adiciona R$ 4000 ao valor da casa
- Cada andar adiciona R$ 10000 ao valor da casa
- Cada ano **reduz** em R$ 2000 o valor da casa
- R$ 80000 seria o “valor base” de qualquer casa

Olhando com detalhes o modelo

Área da casa [m²] (x_1)	Número de quartos (x_2)	Andares (x_3)	Idade [anos] (x_4)	Custo em R$ (y)
32	1	1	2	51.000
149	3	2	10	265.000
78	2	1	30	110.000
...	...	...	...	...
220	4	2	5	315.000

O modelo abaixo é bom para estimar o preço das casas?

$$f_{w,b}(x) = 0.1x_1 + 4x_2 + 10x_3 - 2x_4 + 80 \quad \rightarrow \quad \text{valor da casa em milhares de R\$}$$

- Para a primeira casa do conjunto de dados, temos:

$$f_{w,b}(x) = 0.1 \cdot 32 + 4 \cdot 1 + 10 \cdot 1 - 2 \cdot 2 + 80 = 93.2$$

- Para as demais casas, temos

$$f_{w,b}(x) = 0.1 \cdot 149 + 4 \cdot 3 + 10 \cdot 2 - 2 \cdot 10 + 80 = 106.9$$

$$f_{w,b}(x) = 0.1 \cdot 78 + 4 \cdot 2 + 10 \cdot 1 - 2 \cdot 30 + 80 = 45.8$$

$$f_{w,b}(x) = 0.1 \cdot 220 + 4 \cdot 4 + 10 \cdot 2 - 2 \cdot 5 + 80 = 128.0$$

Olhando com detalhes o modelo

Área da casa [m²] (x_1)	Número de quartos (x_2)	Andares (x_3)	Idade [anos] (x_4)	Custo em R$ (y)
32	1	1	2	51.000
149	3	2	10	265.000
78	2	1	30	110.000
...	...	...	...	...
220	4	2	5	315.000

Conclusão

Observando como o modelo em tela se comporta para os dados que temos, parece que um modelo mais assertivo poderia ter sido obtido.

Pergunta:

Como obter um modelo mais preciso?

Um modelo com n características é dado por

$$f_{w,b}(x) = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b$$

Parâmetros do modelo:

$$\overrightarrow{w} = \begin{bmatrix} w_1 & w_2 & \cdots & w_n \end{bmatrix}$$

$b \quad \rightarrow \quad b$ não é um vetor, mas sim um escalar!

Características presentes no modelo:

$$\overrightarrow{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$$

Utilizando essa notação, note que podemos reescrever $f_{w,b}(x)$ na seguinte forma compacta:

$$f_{w,b}(x) = \overrightarrow{w} \cdot \overrightarrow{x} + b$$

onde $\cdot$ denota o produto escalar, tal que

$$\overrightarrow{w} \cdot \overrightarrow{x} = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$$

Notação sem vetorização

Modelo:

$$f_{w_1,w_2,\cdots,w_n,b}(x_1,x_2,\cdots,x_n) = w_1x_1 + w_2x_2 + \cdots + w_nx_n + b$$

Parâmetros:

$$w_1, w_2, \cdots, w_n \quad \text{e} \quad b$$

Função custo:

$$J(w_1, w_2, \cdots, w_n, b)$$

Método do Gradiente consiste em repetir até convergir:

$$w_j = w_j - \alpha \frac{d}{dw_j} J(w_1, w_2, \cdots, w_n, b)$$

$$b = b - \alpha \frac{d}{db} J(w_1, w_2, \cdots, w_n, b)$$

Implementando a Regressão Linear Múltipla na prática

Notação com vetorização

Modelo:

$$f_{\overrightarrow{w},b}(\overrightarrow{x}) = \overrightarrow{w} \cdot \overrightarrow{x} + b$$

Parâmetros:

$$\overrightarrow{w} \quad \text{e} \quad b$$

Função custo:

$$J(\overrightarrow{w}, b)$$

Método do Gradiente consiste em repetir até convergir:

$$w_j = w_j - \alpha \frac{d}{dw_j} J(\overrightarrow{w}, b)$$

$$b = b - \alpha \frac{d}{db} J(\overrightarrow{w}, b)$$

Perguntas:

- Agora que temos n características, quanto vale $\frac{d}{dw_j} J(\overrightarrow{w}, b)$ para cada w_j?
- Agora que temos n características, quanto vale $\frac{d}{db} J(\overrightarrow{w}, b)$?

É possível mostrar que

$$\frac{d}{dw_j} J(\overrightarrow{w}, b) = \frac{1}{m} \sum_{i=1}^{m} \left(f_{\overrightarrow{w},b} \left(\overrightarrow{x}^{(i)} \right) - y^{(i)} \right) \overrightarrow{x}_j^{(i)}$$

$$\frac{d}{db} J(\overrightarrow{w}, b) = \frac{1}{m} \sum_{i=1}^{m} \left(f_{\overrightarrow{w},b} \left(\overrightarrow{x}^{(i)} \right) - y^{(i)} \right)$$

Portanto, o Método do Gradiente aplicado ao contexto de Regressão Linear múltipla consiste em...

Repetir até convergir:

$$w_1 = w_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(f_{\overrightarrow{w},b} \left(\overrightarrow{x}^{(i)} \right) - y^{(i)} \right) \overrightarrow{x}_1^{(i)}$$

$$\vdots$$

$$w_n = w_n - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(f_{\overrightarrow{w},b} \left(\overrightarrow{x}^{(i)} \right) - y^{(i)} \right) \overrightarrow{x}_n^{(i)}$$

$$b = b - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(f_{\overrightarrow{w},b} \left(\overrightarrow{x}^{(i)} \right) - y^{(i)} \right)$$

De olho no código!

De olho no código!

Vamos agora ver como implementar na prática o **Método do Gradiente para Regressão linear Múltipla**

Acesse o Python Notebook usando o QR code ou o link abaixo:

https://colab.research.google.com/github/xaximpvp2/master/blob/main/codigo_aula6_regressao_linear_multipla.ipynb

Atividade de aula

Parte 1

Rode todo o código. Responda às questões nele contidas e complete-o, se necessário.

Parte 2

1. Quais foram os valores obtidos para $\overrightarrow{w}$ e b?
2. Esses valores são os melhores possíveis?
3. O que pode estar acontecendo?
4. Rode novamente o Método do Gradiente inicializando $\overrightarrow{w}$ e b num local mais próximo do ótimo. Quais foram os valores agora obtidos para $\overrightarrow{w}$ e b?