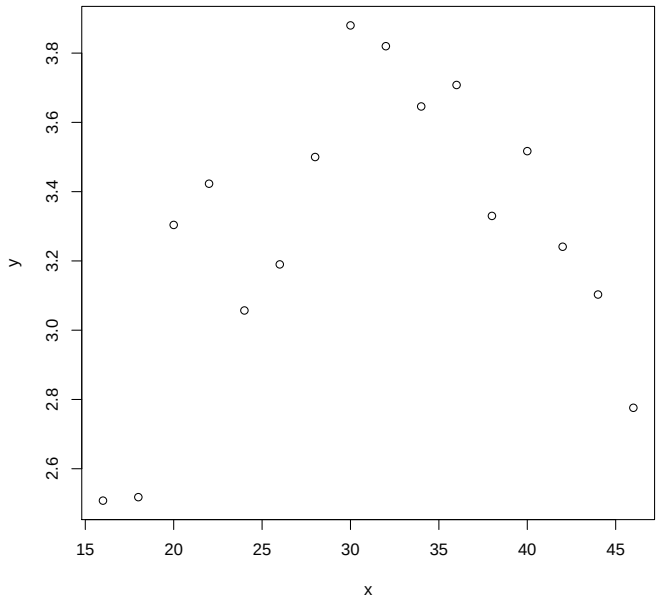
Regressão Linear Múltipla

Prof. Dr. Leandro Balby Marinho

Aprendizagem de Máquina

Regressão Polinomial

Em muitos casos o grafo de dispersão sugere uma relação não linear entre x e y.

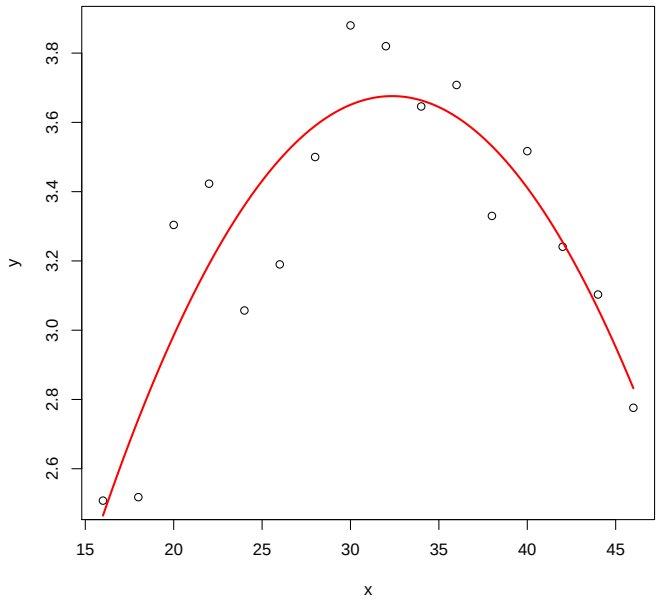


A equação do modelo quadrático, por exemplo, é dada por

$$y_i = w_0 + w_1 x_i + w_2 x_i^2 + \epsilon_i$$

Regressão Polinomial

Em muitos casos o grafo de dispersão sugere uma relação não linear entre x e y.



A equação do modelo quadrático, por exemplo, é dada por

$$y_i = w_0 + w_1 x_i + w_2 x_i^2 + \epsilon_i$$

Regressão Polinomial

Modelo:

$$y_i = w_0 + w_1 x_i + w_2 x_i^2 + \ldots + w_p x_i^p + \epsilon_i$$

- atributo 1 $=$ 1 (constante) parâmetro 1 $= w_0$
- atributo 2 $= x$ parâmetro 2 $= w_1$
- atributo 3 $= x^2$ parâmetro 3 $= w_2$
- $\ldots$
- atributo $p+1 = x^p$ parâmetro $p+1 = w_d$

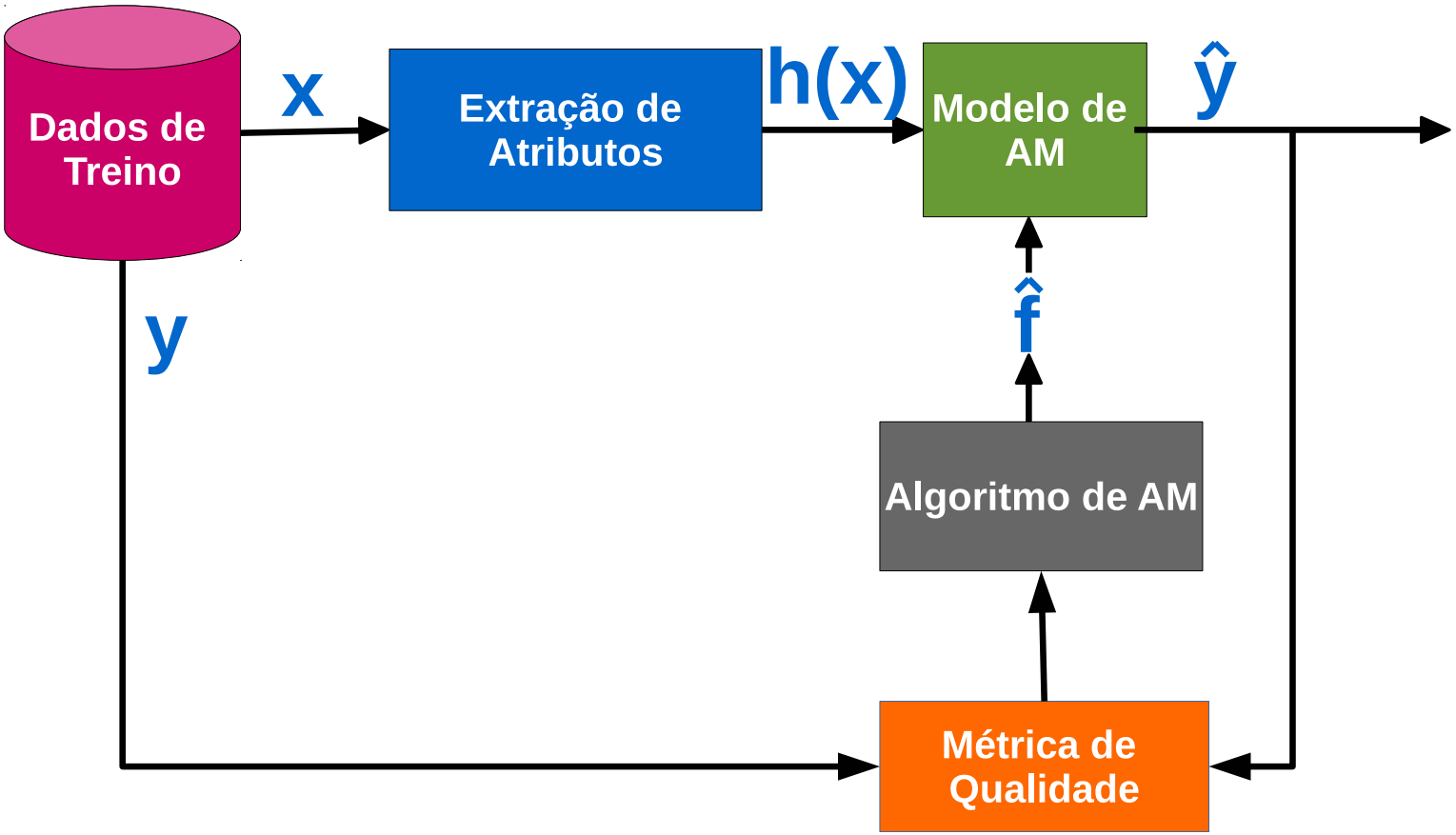
Atributos como funções

Modelo:

$$y_i = w_0 h_0(x_i) + w_1 h_1(x_i) + w_2 h_2(x_i) + \ldots + w_p h_p(x_i) + \epsilon_i$$
$$= \sum_{j=0}^{D} w_j h_j(x_i) + \epsilon_i$$

- atributo 1 = $h_0(x)$ . . . geralmente 1 (constante)
- atributo 2 = $h_1(x)$ . . . e.g., x
- atributo 3 = $h_2(x)$ . . . e.g., x^2
- . . .
- atributo p+1 = $h_p(x)$ . . . e.g., x^p

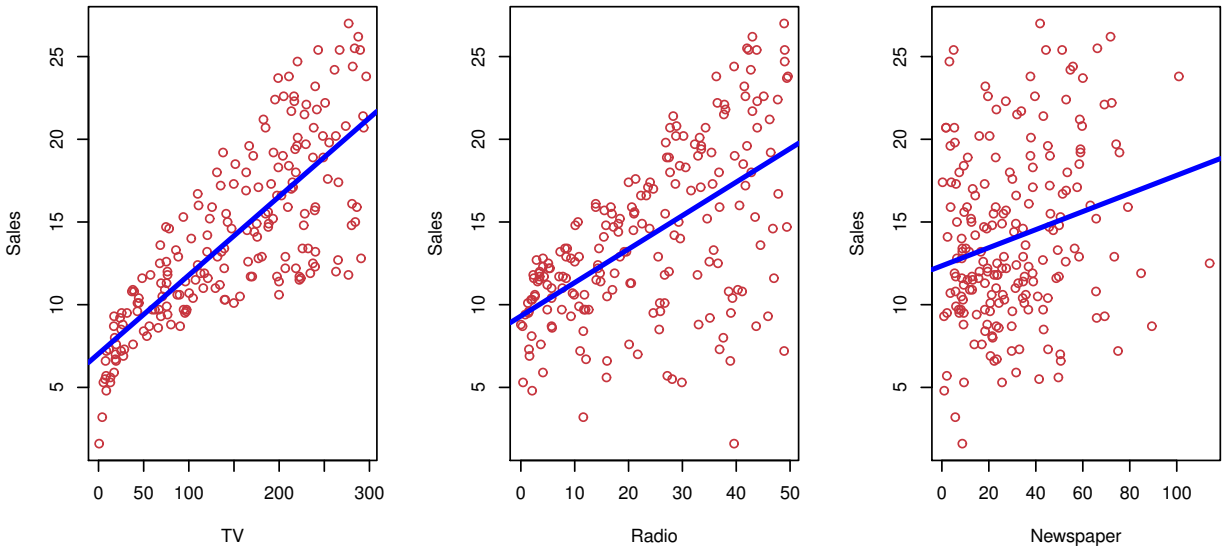
Atributos como funções



Roteiro

Regressão Múltipla

Como usar as outras variáveis disponíveis no modelo de regressão?



Notação

Saída: y (escalar)
Entradas: $\mathrm{x} = (\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_d) \in \mathbb{R}^d$

- $\mathrm{x}^{(i)} = i$-ésima observação (vetor)
- $\mathrm{x}_j^{(i)} = j$-ésima entrada da i-ésima observação (escalar)
- $h_j(\mathrm{x}) = j$-ésimo atributo (escalar)
- # observações (x, y): N
- # entradas x_j: d
- # atributos $h_j(\mathrm{x})$: D

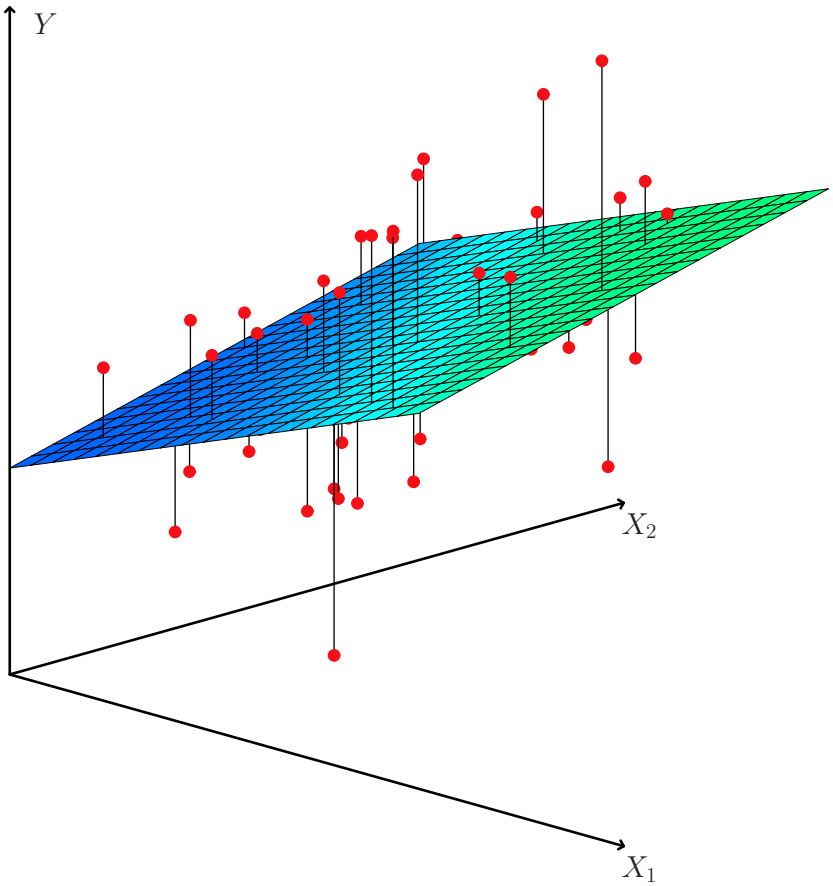
Regressão Linear Múltipla

Modelo:

$$y^{(i)} = w_0 + w_1 x_1^{(i)} + w_2 x_2^{(i)} + \ldots + w_d x_d^{(i)} + \epsilon_i$$

- variável 1 = 1
- variável 2 = x_1 . . . e.g., investimento em TV
- variável 3 = x_2 . . . e.g., investimento em Rádio
- . . .
- variável $d + 1$ = x_d . . . e.g., investimento em redes sociais

Modelo de Regressão como Hiperplano



Modelo de Regressão como uma Curva D-dimensional

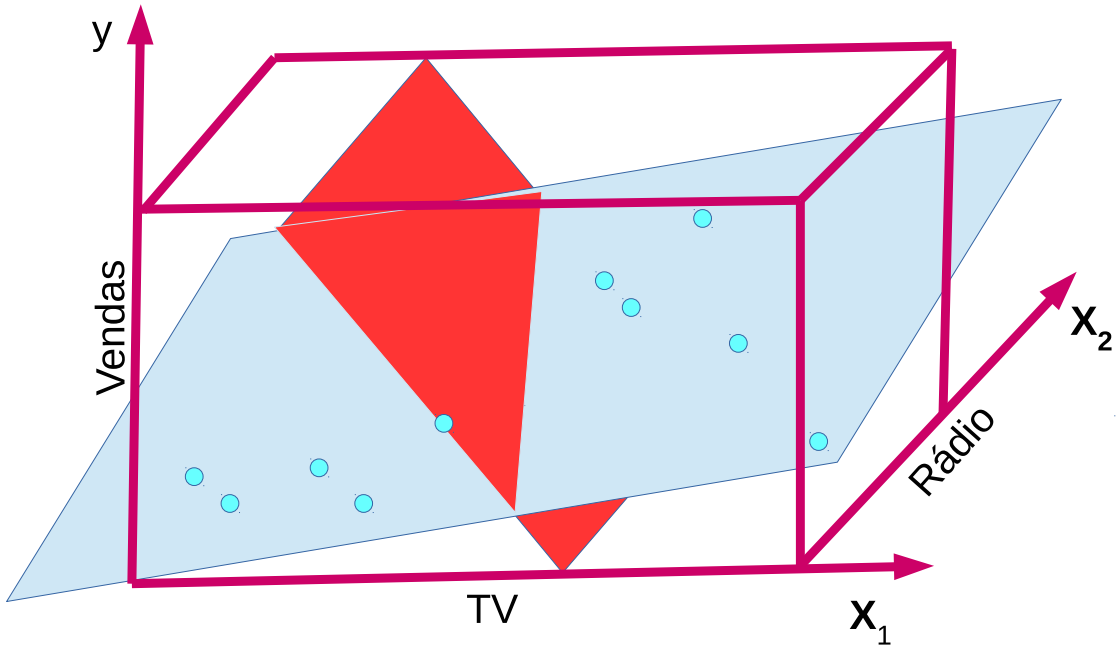
Modelo:

$$y^{(i)} = w_0 + w_1 h_0(\mathrm{x}^{(i)}) + w_2 h_1(\mathrm{x}^{(i)}) + \ldots + w_D h_D(\mathrm{x}^{(i)}) + \epsilon_i$$
$$= \sum_{j=0}^{d} w_j h_j(\mathrm{x}^{(i)}) + \epsilon_i$$

- atributo 1 = $h_0(\mathrm{x})$ . . . e.g., 1
- atributo 2 = $h_1(\mathrm{x})$ . . . e.g., x_1 =investimento em TV
- atributo 3 = $h_2(\mathrm{x})$ . . . e.g., $\log(\mathrm{x}_2)\mathrm{x}_1$
- . . .
- atributo $D+1$=$h_D(\mathrm{x})$ . . . alguma outra função de $\mathrm{x}_1, \ldots, \mathrm{x}_D$

Interpretando os coeficientes: dois atributos

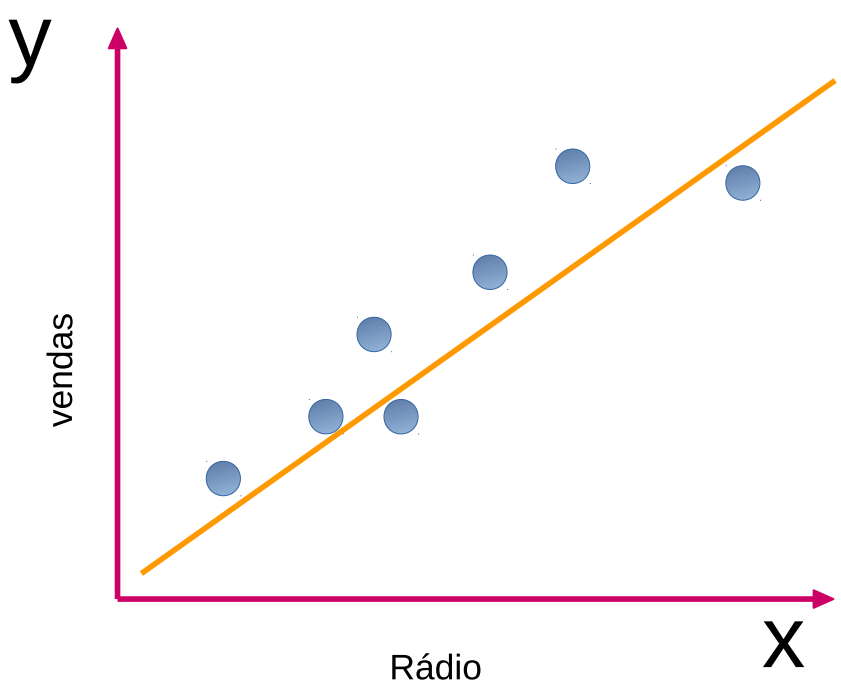
$$\hat{y} = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2$$



Fixando x_1 a intepretação é a mesma da regressão linear simples.

Interpretando os coeficientes: dois atributos

$$\hat{y} = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2$$



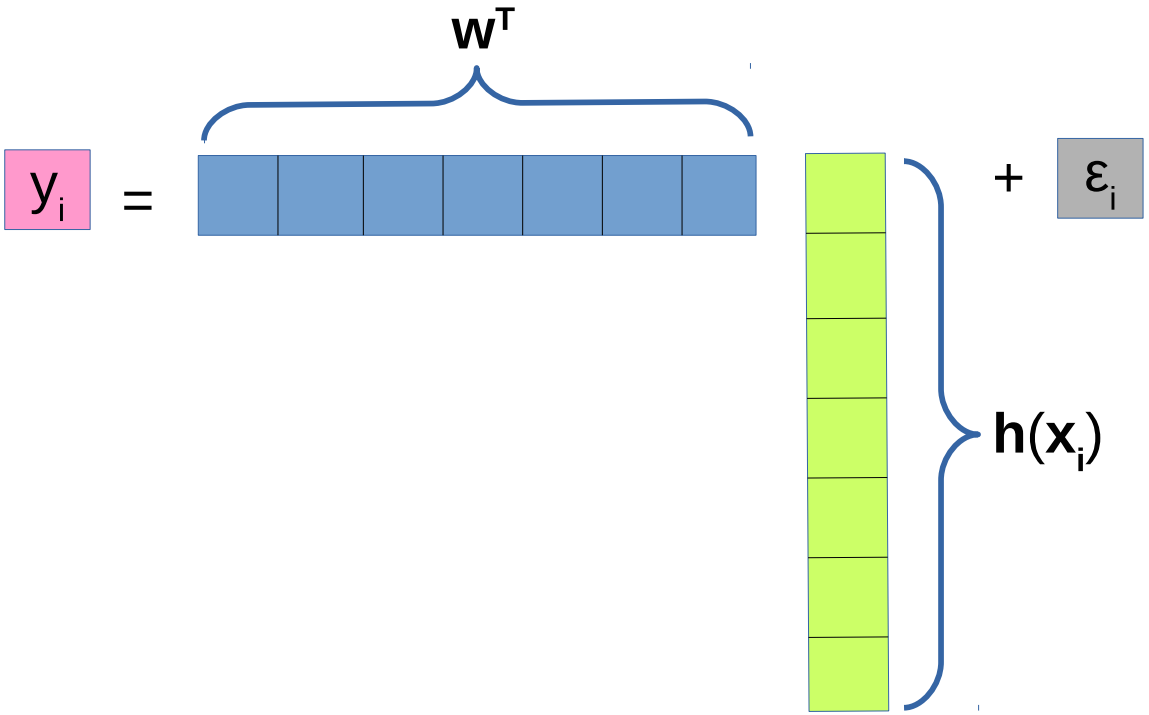
Fixando x_1 a intepretação é a mesma da regressão linear simples.

Interpretando os coeficientes: múltiplos atributos

$$\hat{y} = \hat{w}_0 + \hat{w}_1 x_1 + \ldots + \hat{w}_j x_j + \ldots + \hat{w}_d x_d$$

Fixando todas as variáveis menos uma a intepretação é a mesma da regressão linear simples.

Usando notação de vetores: uma observação

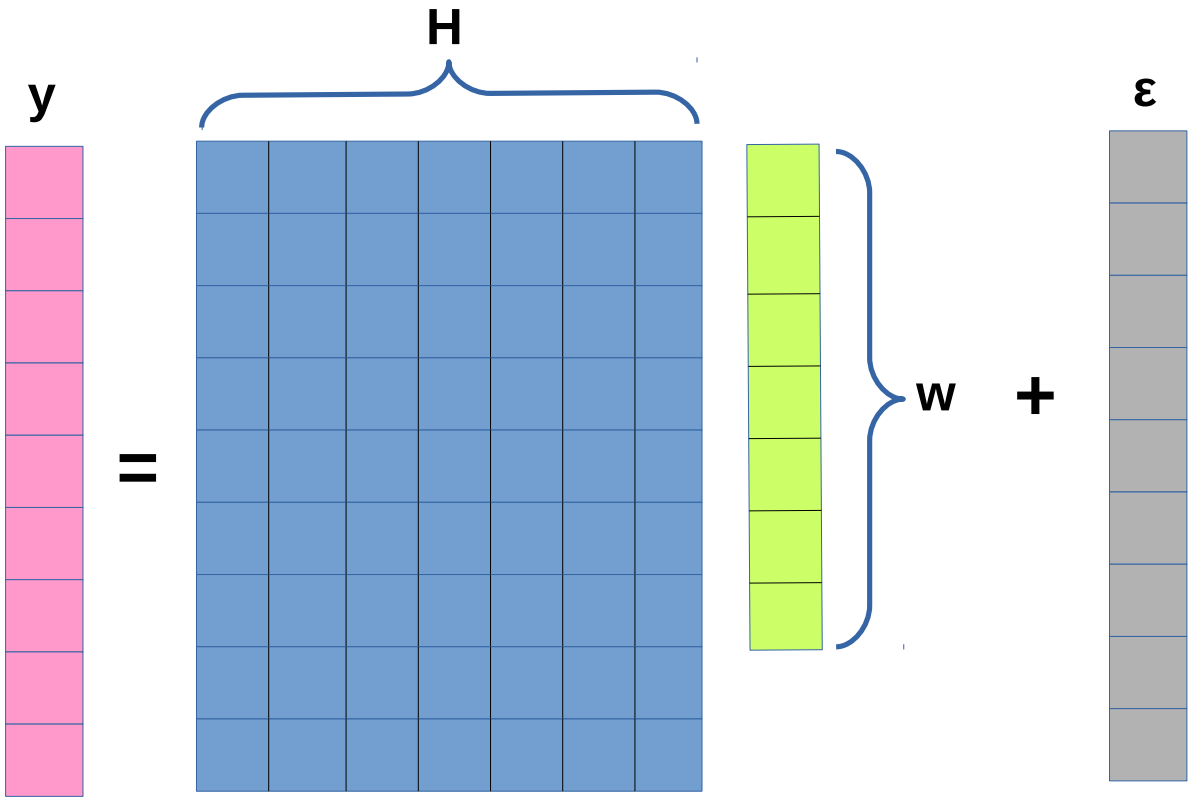


Usando notação de vetores: uma observação

Para a observação i:

$$y^{(i)} = \sum_{j=0}^{D} w_j h_j(\mathbf{x}^{(i)}) + \epsilon^{(i)}$$
$$= \mathbf{w}^T \mathbf{h}(\mathbf{x}^{(i)}) + \epsilon^{(i)}$$

Usando notação de matrizes: todas as observações



$$y = Hw + \epsilon$$

Custo de uma curva D-dimensional

$$\begin{aligned} \text{RSS}(\mathbf{w}) &= \sum_{i=1}^{N} (y_i - \text{h}(\mathbf{x}^{(i)})^T \mathbf{w})^2 \\ &= (\mathbf{y} - \mathbf{H}\mathbf{w})^T (\mathbf{y} - \mathbf{H}\mathbf{w}) \end{aligned}$$

Gradiente do RSS

$$\begin{aligned}\nabla \text{RSS}(w) &= \nabla[(y - Hw)^T (y - Hw)] \\ &= -2H^T (y - Hw)\end{aligned}$$

Calculando parâmetros de forma fechada

$$\nabla \text{RSS}(\mathbf{w}) = -2\mathbf{H}^T(\mathbf{y} - \mathbf{H}\mathbf{w}) = 0$$

Resolvendo para $\mathbf{w}$:

$$\begin{aligned}
-2\mathbf{H}^T\mathbf{y} + 2\mathbf{H}^T\mathbf{H}\hat{\mathbf{w}} &= 0 \\
-\mathbf{H}^T\mathbf{y} + \mathbf{H}^T\mathbf{H}\hat{\mathbf{w}} &= 0 \quad \text{(divide ambos os lados por 2)} \\
\mathbf{H}^T\mathbf{H}\hat{\mathbf{w}} &= \mathbf{H}^T\mathbf{y} \\
(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{H}\hat{\mathbf{w}} &= (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y} \\
\hat{\mathbf{w}} &= (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y}
\end{aligned}$$

Custo da inversão de matrizes (quando inversível): $O(D^3)$

Equações Normais como Sistemas de Equações Lineares

- As equação normais podem ser dadas por:

$$\mathrm{H}^T \mathrm{H} \hat{\mathrm{w}} = \mathrm{H}^T \mathrm{y}$$

- A equação acima pode ser representada por um sistema de equações lineares da forma:

$$\underbrace{\mathrm{H}^T \mathrm{H}}_{A} \underbrace{\mathrm{w}}_{x} = \underbrace{\mathrm{H}^T \mathrm{y}}_{b}$$

Vários métodos de resolução:

- Eliminação Gaussiana
- Fatoração de Cholesky
- Fatoração QR

Exemplo 1

Use um modelo de regressão linear múltipla para estimar o valor de y para $x_1 = 3$ e $x_2 = 4$ considerando os dados abaixo.

x_1	x_2	y
1	2	3
2	3	2
4	1	7
5	5	1

Exemplo 1 cont.

Modelo regressão múltipla:

$$y_i = w_0 + w_1 x_1 + w_2 x_2 + \epsilon_i$$

$$\mathrm{H} = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 1 & 4 & 1 \\ 1 & 5 & 5 \end{bmatrix}, \quad \mathrm{y} = \begin{bmatrix} 3 \\ 2 \\ 7 \\ 1 \end{bmatrix}$$

$$\mathrm{H}^T\mathrm{H} = \begin{bmatrix} 4 & 12 & 11 \\ 12 & 46 & 37 \\ 11 & 37 & 39 \end{bmatrix}, \quad \mathrm{H}^T\mathrm{y} = \begin{bmatrix} 13 \\ 40 \\ 24 \end{bmatrix}$$

Exemplo 1 cont.

Estimando os parâmetros por Eliminação Gaussiana:

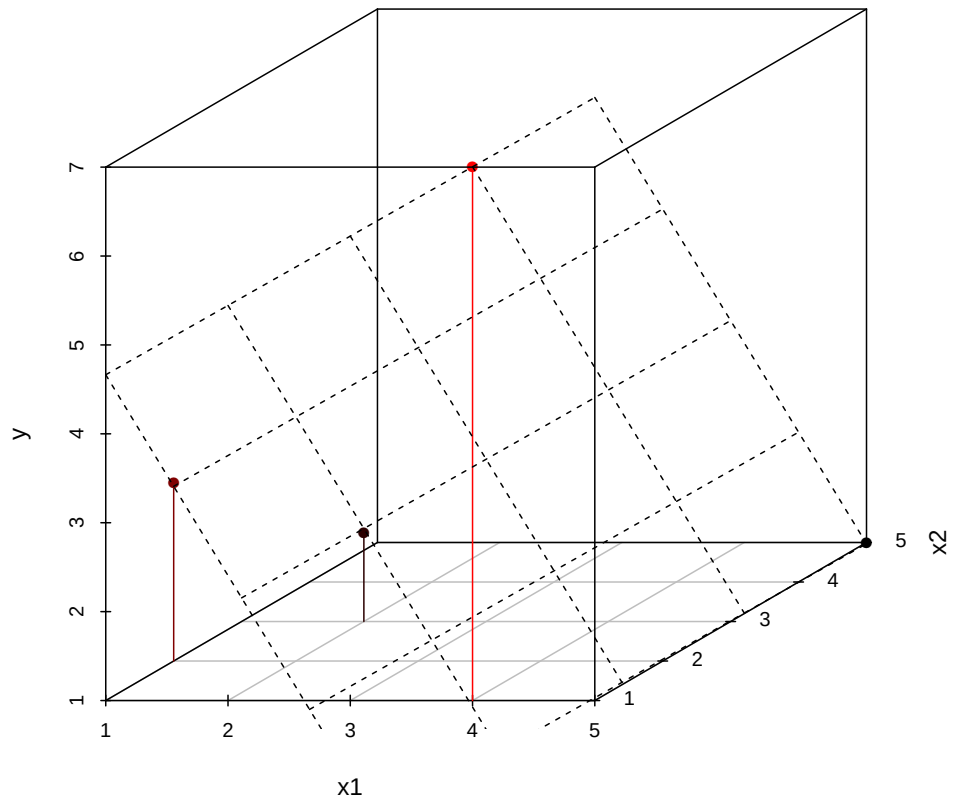
$$\left[\begin{array}{ccc|c} 4 & 12 & 11 & 13 \\ 12 & 46 & 37 & 40 \\ 11 & 37 & 39 & 24 \end{array}\right] \rightarrow \left[\begin{array}{ccc|c} 1 & 3 & 2.75 & 3.25 \\ 0 & 10 & 4 & 1 \\ 0 & 4 & 8.75 & -11.75 \end{array}\right]$$

$$\rightarrow \left[\begin{array}{ccc|c} 4 & 12 & 11 & 13 \\ 0 & 1 & 0.4 & 0.1 \\ 0 & 0 & 7.15 & -12.15 \end{array}\right]$$

$$\mathbf{w} \approx \begin{bmatrix} 5.583 \\ 0.779 \\ -1.699 \end{bmatrix}$$

Exemplo 1 cont.

$$\hat{y}(x_1 = 3, x_2 = 4) = 5.583 + 0.779x_1 - 1.699x_2 = 1.124$$



Gradiente Descendente

Gradient-Descent

1 **while** not converged

2 $$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \alpha \underbrace{\nabla \text{RSS}(\mathbf{w}^{(t)})}_{-2\mathbf{H}^T(\mathbf{y}-\mathbf{H}\mathbf{w})}$$

Gradiente Descendente

Gradient-Descent

1 **while** not converged

2 $$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + 2\alpha \mathbf{H}^T (\mathbf{y} - \underbrace{\mathbf{H}\mathbf{w}^{(t)}}_{\hat{y}})$$

Derivada parcial de um parâmetro

$$\begin{aligned}\text{RSS}(\mathbf{w}) &= \sum_{i=1}^{N}(y_i - \mathrm{h}(\mathrm{x}^{(i)})^T\mathbf{w})^2 \\ &= \sum_{i=1}^{N}(y^{(i)} - w_0 h_0(\mathrm{x}^{(i)}) - w_1 h_1(\mathrm{x}^{(i)}) - \ldots - w_D h_D(\mathrm{x}^{(i)}))\end{aligned}$$

Derivada parcial em relação a w_j

$$\begin{aligned}&= \sum_{i=1}^{N} 2(y^{(i)} - w_0 h_0(\mathrm{x}^{(i)}) - w_1 h_1(\mathrm{x}^{(i)}) - \ldots - w_D h_D(\mathrm{x}^{(i)})(-h_j(\mathrm{x}^{(i)})) \\ &= -2\sum_{i=1}^{N} h_j(\mathrm{x}^{(i)})(y^{(i)} - \mathrm{h}(\mathrm{x}^{(i)})^T\mathbf{w})\end{aligned}$$

Algoritmo do Gradiente Descendente

GradientDescent(α, ϵ)

1 initialize $\mathrm{w}, t = 1$
2 **while** $||\nabla \mathrm{RSS}(\mathrm{w}^{(t)})|| \geq \epsilon$
3 **for** $j = 0, \ldots, D$
4 $\mathrm{partial}[j] = -2 \sum_{i=1}^{N} h_j(\mathrm{x}^{(i)})(y^{(i)} - \mathrm{h}(\mathrm{x}^{(i)})^T \mathrm{w})$
5 $\mathrm{w}_j^{(t+1)} = \mathrm{w}_j^{(t)} - \alpha \cdot \mathrm{partial}[j]$
6 $t = t + 1$
7 **return** w

Gradiente Descendente vs. Equações Normais

- Gradiente Descendente
  - Precisa escolher α.
  - Pode precisar de muitas iterações.
  - Relativamente eficiente para D grande.

- Equações Normais
  - Não precisa escolher α.
  - Não precisa iterar.
  - Métodos de resolução de sistemas de equações lineare podem ser caros (e.g. fatoração de Cholesky $\in O(D^3)$).
  - Lento para D muito grande.

Referências


Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning with Applications in R. Springer, 2013.


Emily Fox and Carlos Guestrin. Machine Learning Specialization. Curso online disponível em `https://www.coursera.org/specializations/machine-learnin` Último acesso: 04/09/2017.