UNIVERSIDADE FEDERAL DO CEARÁ

DEPARTAMENTO DE COMPUTAÇÃO

Aprendizagem de Máquina

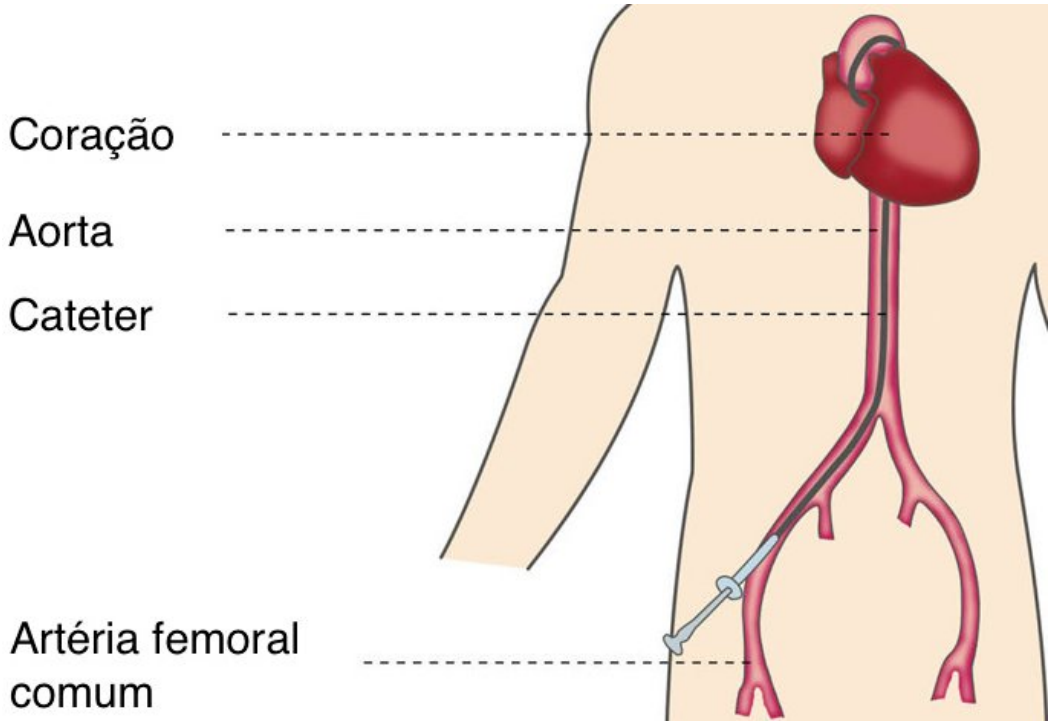
César Lincoln Cavalcante Mattos

2024

Agenda

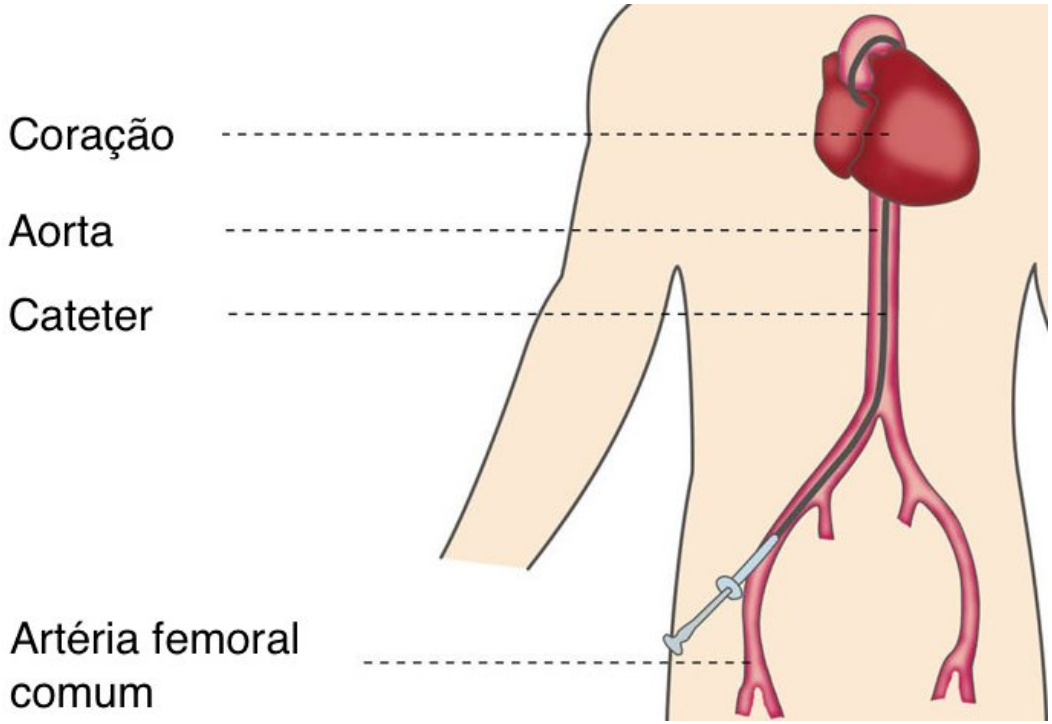
Regressão linear simples

- **Cataterismo cardíaco**: procedimento de inserção de um cateter (usualmente pela artéria femural) para diagnosticar problemas de obstrução no coração.



Regressão linear simples

- **Cataterismo cardíaco**: procedimento de inserção de um cateter (usualmente pela artéria femural) para diagnosticar problemas de obstrução no coração.



- **Problema**: Dada a **altura** de um paciente, qual o **comprimento** do cateter necessário para alcançar seu coração?

Regressão linear simples

- Considere a tabela a seguir relacionando alturas de jovens pacientes e comprimentos do cateter correspondente:

Altura (m)	**Comprimento (cm)**
1.087	37
1.613	50
0.953	34
1.003	36
1.156	43
0.978	28
1.092	37
0.572	20
0.940	34
0.597	30
0.838	38
1.473	47

Regressão linear simples

- Considere a tabela a seguir relacionando alturas de jovens pacientes e comprimentos do cateter correspondente:

Altura (m)	**Comprimento (cm)**
1.087	37
1.613	50
0.953	34
1.003	36
1.156	43
0.978	28
1.092	37
0.572	20
0.940	34
0.597	30
0.838	38
1.473	47

- **Problema**: Dado uma altura **não presente** na tabela, qual deverá ser o comprimento do cateter?

Regressão linear simples

- A coluna **Altura** é a **entrada** do nosso modelo.
- A coluna **Comprimento** é a **saída** do nosso modelo.
- Nosso **conjunto de dados** é formado por 12 alturas e 12 comprimentos correspondentes.

Regressão linear simples

- A coluna **Altura** é a **entrada** do nosso modelo.
- A coluna **Comprimento** é a **saída** do nosso modelo.
- Nosso **conjunto de dados** é formado por 12 alturas e 12 comprimentos correspondentes.
- Matematicamente, temos:

$$\mathcal{D} = \{(x_1, y_1), \cdots, (x_2, y_2)\} = \{(x_i, y_i)\}_{i=1}^{12},$$

em que x_i é a i-ésima entrada e y_i é a i-ésima saída.

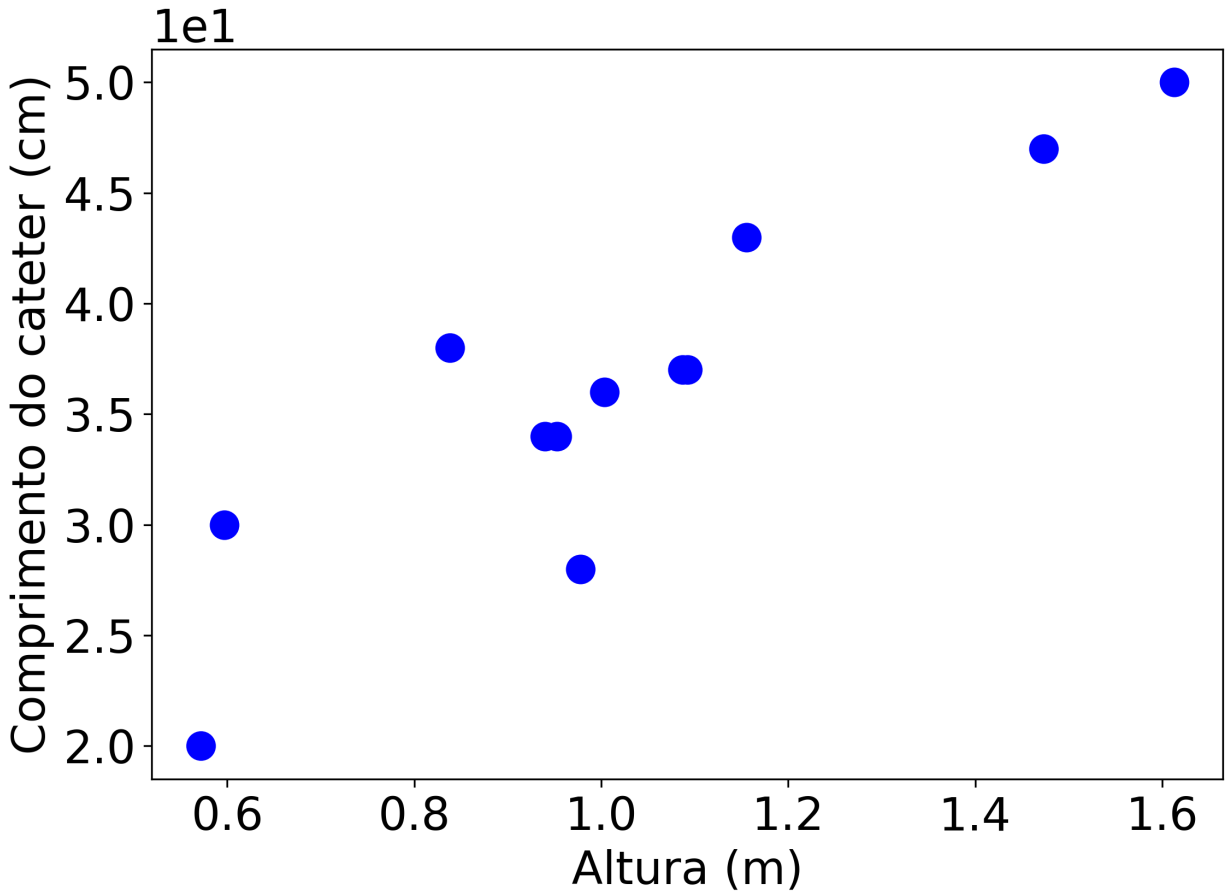
Regressão linear simples

- A coluna **Altura** é a **entrada** do nosso modelo.
- A coluna **Comprimento** é a **saída** do nosso modelo.
- Nosso **conjunto de dados** é formado por 12 alturas e 12 comprimentos correspondentes.
- Matematicamente, temos:

$$\mathcal{D} = \{(x_1, y_1), \cdots, (x_2, y_2)\} = \{(x_i, y_i)\}_{i=1}^{12},$$

  em que x_i é a i-ésima entrada e y_i é a i-ésima saída.
- **Objetivo**: Encontrar uma relação entre x_i e y_i que forneça uma predição $\hat{y}_i$ o mais próximo possível da saída real y_i.

Regressão linear simples



Regressão linear simples

Terminologia

- **Atributo (feature)**: Uma dada característica de um padrão.

Regressão linear simples

Terminologia

- **Atributo (feature)**: Uma dada característica de um padrão.
- **Padrão (pattern)**: Um vetor de atributos que representa um exemplo.

Regressão linear simples

Terminologia

- **Atributo (feature)**: Uma dada característica de um padrão.
- **Padrão (pattern)**: Um vetor de atributos que representa um exemplo.
- **Modelo**: Uma função que expressa a relação entre um padrão de entrada e sua saída correspondente.

Regressão linear simples

Terminologia

- **Atributo (feature)**: Uma dada característica de um padrão.
- **Padrão (pattern)**: Um vetor de atributos que representa um exemplo.
- **Modelo**: Uma função que expressa a relação entre um padrão de entrada e sua saída correspondente.
- **Função custo (ou função objetivo)**: Indica o quão mal (ou o quão bem) um modelo aproxima os dados disponíveis.

Regressão linear simples

Terminologia

- **Atributo (feature)**: Uma dada característica de um padrão.
- **Padrão (pattern)**: Um vetor de atributos que representa um exemplo.
- **Modelo**: Uma função que expressa a relação entre um padrão de entrada e sua saída correspondente.
- **Função custo (ou função objetivo)**: Indica o quão mal (ou o quão bem) um modelo aproxima os dados disponíveis.
- **Parâmetros**: Variáveis que caracterizam o modelo proposto.

Regressão linear simples

Terminologia

- **Atributo (feature)**: Uma dada característica de um padrão.
- **Padrão (pattern)**: Um vetor de atributos que representa um exemplo.
- **Modelo**: Uma função que expressa a relação entre um padrão de entrada e sua saída correspondente.
- **Função custo (ou função objetivo)**: Indica o quão mal (ou o quão bem) um modelo aproxima os dados disponíveis.
- **Parâmetros**: Variáveis que caracterizam o modelo proposto.
- **Risco empírico**: Estimativa do risco (custo) obtida a partir dos dados disponíveis.

Regressão linear simples

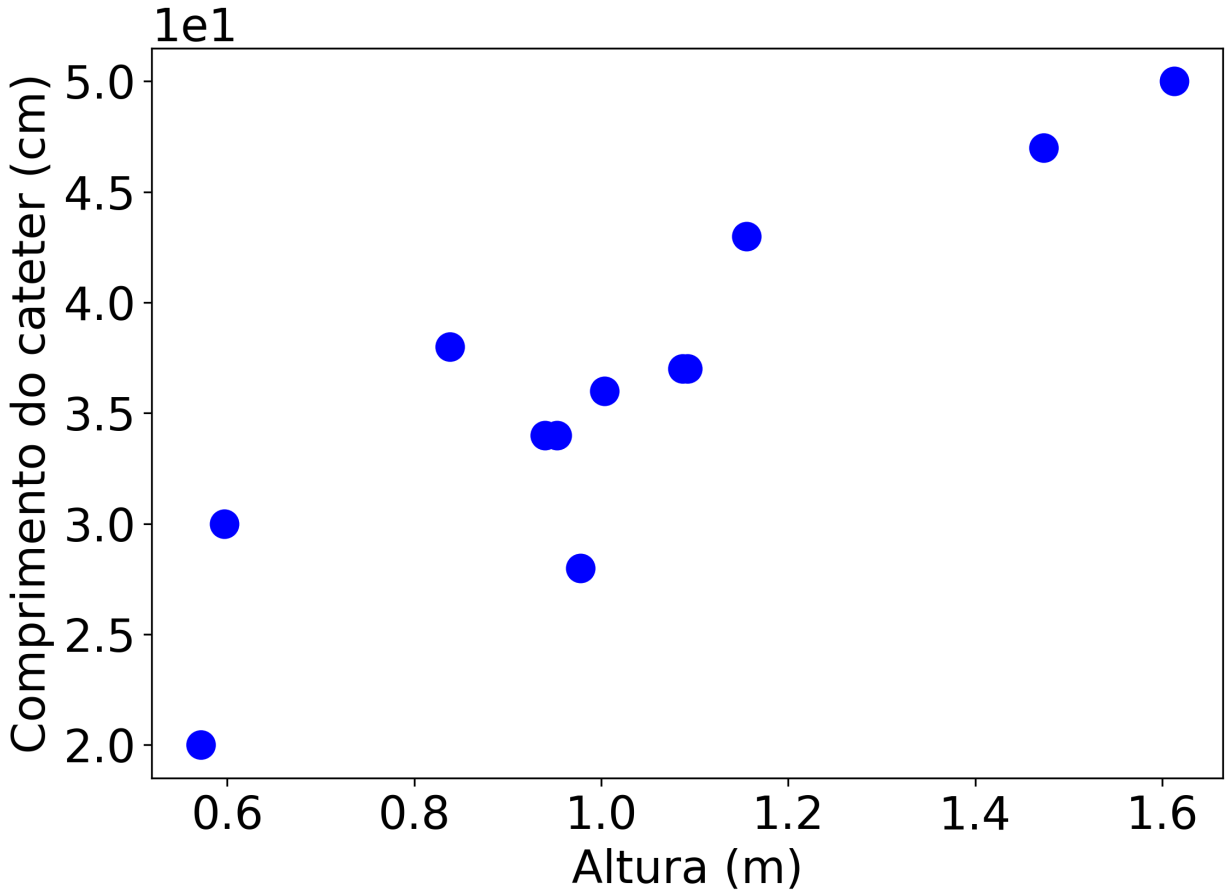
Terminologia

- **Atributo (feature)**: Uma dada característica de um padrão.
- **Padrão (pattern)**: Um vetor de atributos que representa um exemplo.
- **Modelo**: Uma função que expressa a relação entre um padrão de entrada e sua saída correspondente.
- **Função custo (ou função objetivo)**: Indica o quão mal (ou o quão bem) um modelo aproxima os dados disponíveis.
- **Parâmetros**: Variáveis que caracterizam o modelo proposto.
- **Risco empírico**: Estimativa do risco (custo) obtida a partir dos dados disponíveis.
- **Otimização (ou treinamento, aprendizagem)**: Algoritmo de obtenção dos parâmetros do modelo que minimizem uma função custo (ou maximizem uma função objetivo).

Regressão linear simples

- Considere uma relação linear entre x_i (entrada do modelo) e $\hat{y}_i$ (saída do modelo):

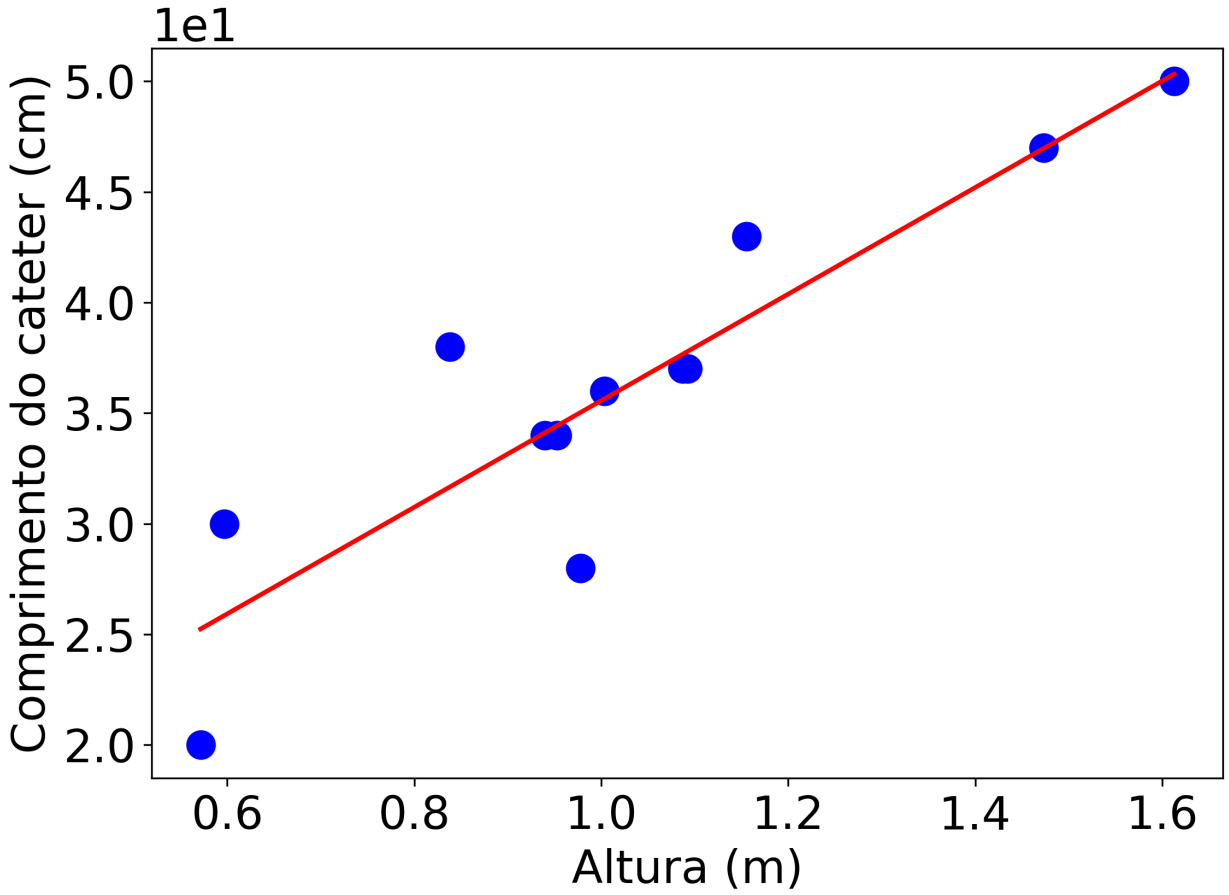
$$\hat{y}_i = w_0 + w_1 x_i.$$



Regressão linear simples

- Considere uma relação linear entre x_i (entrada do modelo) e $\hat{y}_i$ (saída do modelo):

$$\hat{y}_i = w_0 + w_1 x_i.$$



Regressão linear simples

- Escolhemos uma função custo quadrática para os erros obtidos pelo modelo:

$$\mathcal{J}(w_0, w_1) = \frac{1}{2N} \sum_{i=1}^{N} e_i^2,$$

$$e_i = y_i - \hat{y}_i.$$

Regressão linear simples

- Escolhemos uma função custo quadrática para os erros obtidos pelo modelo:

$$\mathcal{J}(w_0, w_1) = \frac{1}{2N} \sum_{i=1}^{N} e_i^2,$$

$$e_i = y_i - \hat{y}_i.$$

- Esse custo é chamado **Erro Quadrático Médio (MSE, Mean Squared Error)**.

Regressão linear simples

- Desejamos minimizar a função custo em relação aos parâmetros do modelo:

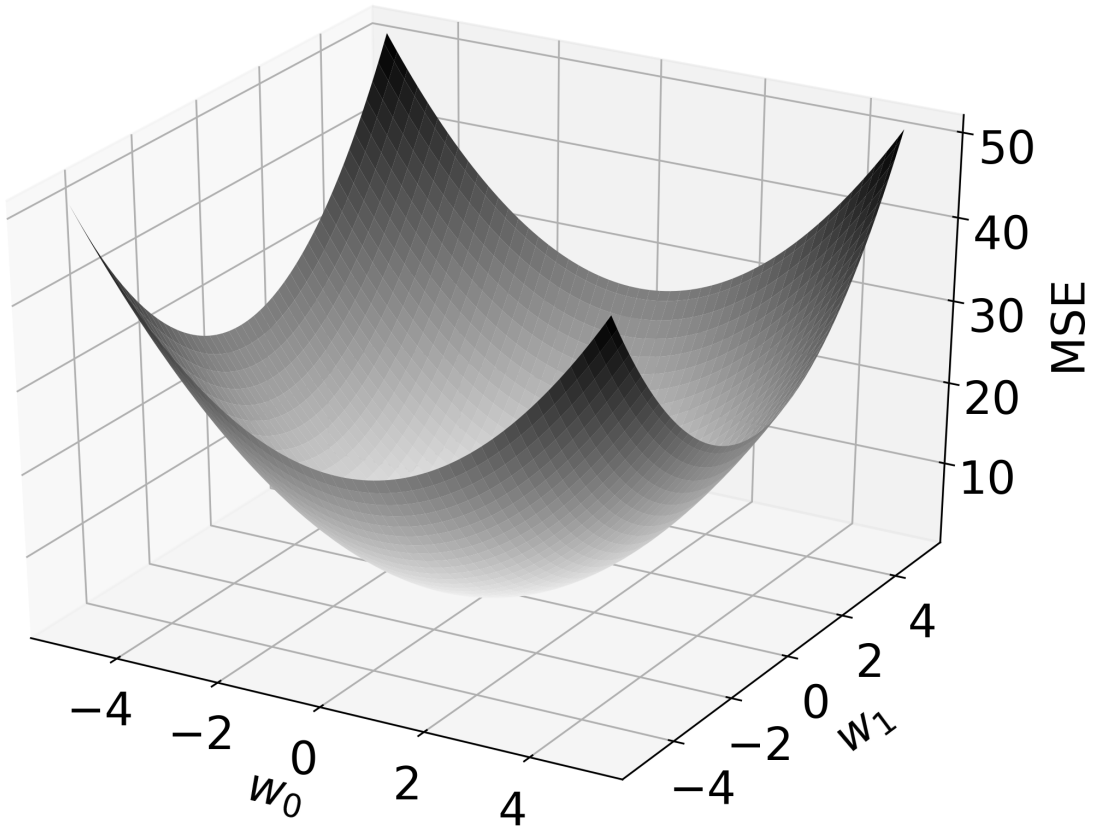
$$\min_{w_0, w_1} \mathcal{J}(w_0, w_1)$$

$$\min_{w_0, w_1} \frac{1}{2N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

$$\min_{w_0, w_1} \frac{1}{2N} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i)^2$$

Regressão linear simples

$$\min_{w_0, w_1} \frac{1}{2N} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i)^2$$



Regressão linear simples

- Escolhemos valores iniciais para w_0 e w_1.
- Movimentamos os parâmetros na direção que diminui a função custo $\mathcal{J}(w_0, w_1)$:

$$w_0 \leftarrow w_0 - \alpha \frac{\partial \mathcal{J}}{\partial w_0},$$
$$w_1 \leftarrow w_1 - \alpha \frac{\partial \mathcal{J}}{\partial w_1}$$

- $\alpha > 0$ é um **passo de aprendizado**.

Regressão linear simples

$$\frac{\partial \mathcal{J}}{\partial w_0} = \frac{\partial}{\partial w_0} \frac{1}{2N} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i)^2$$

$$\frac{\partial \mathcal{J}}{\partial w_0} = \frac{1}{N} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i)(-1)$$

$$\frac{\partial \mathcal{J}}{\partial w_0} = -\frac{1}{N} \sum_{i=1}^{N} e_i$$

Regressão linear simples

$$\frac{\partial \mathcal{J}}{\partial w_0} = \frac{\partial}{\partial w_0} \frac{1}{2N} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i)^2$$

$$\frac{\partial \mathcal{J}}{\partial w_0} = \frac{1}{N} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i)(-1)$$

$$\frac{\partial \mathcal{J}}{\partial w_0} = -\frac{1}{N} \sum_{i=1}^{N} e_i$$

- Logo:

$$w_0 \leftarrow w_0 + \alpha \frac{1}{N} \sum_{i=1}^{N} e_i$$

Regressão linear simples

$$\frac{\partial \mathcal{J}}{\partial w_1} = \frac{\partial}{\partial w_1} \frac{1}{2N} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i)^2$$

$$\frac{\partial \mathcal{J}}{\partial w_1} = \frac{1}{N} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i)(-x_i)$$

$$\frac{\partial \mathcal{J}}{\partial w_1} = -\frac{1}{N} \sum_{i=1}^{N} e_i x_i$$

Regressão linear simples

$$\frac{\partial \mathcal{J}}{\partial w_1} = \frac{\partial}{\partial w_1} \frac{1}{2N} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i)^2$$

$$\frac{\partial \mathcal{J}}{\partial w_1} = \frac{1}{N} \sum_{i=1}^{N} (y_i - w_0 - w_1 x_i)(-x_i)$$

$$\frac{\partial \mathcal{J}}{\partial w_1} = -\frac{1}{N} \sum_{i=1}^{N} e_i x_i$$

- Logo:

$$w_1 \leftarrow w_1 + \alpha \frac{1}{N} \sum_{i=1}^{N} e_i x_i$$

Regressão linear simples

Gradiente Descendente (GD, gradient descent)

1. Escolha um valor α positivo e pequeno.
2. Inicialize os parâmetros do modelo na iteração $t = 0$.
3. Repita por diversas iterações (épocas):
   1. $t \leftarrow t + 1$;
   2. Calcule os erros do modelo:

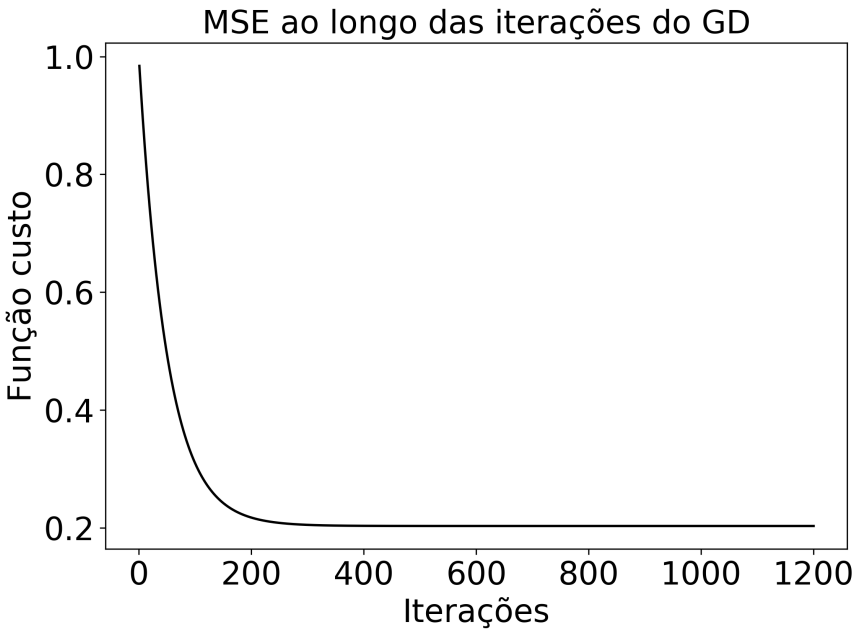
$$\hat{y}_i(t-1) = w_0(t-1) + w_1(t-1)x_i, \quad \forall i,$$
$$e_i(t-1) = y_i - \hat{y}_i(t-1), \quad \forall i.$$

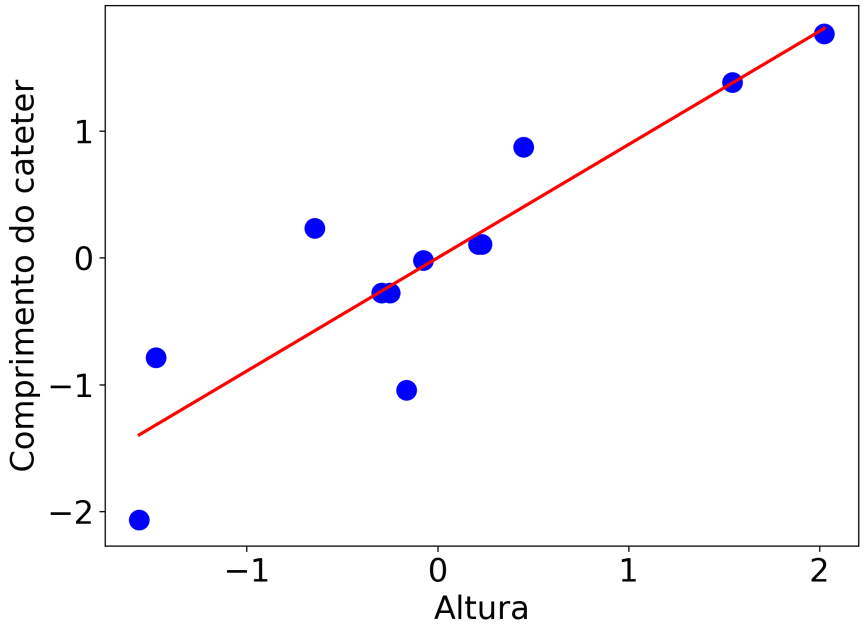
   3. Atualize os parâmetros:

$$w_0(t) = w_0(t-1) + \alpha \frac{1}{N} \sum_{i=1}^{N} e_i(t-1)$$

$$w_1(t) = w_1(t-1) + \alpha \frac{1}{N} \sum_{i=1}^{N} e_i(t-1)x_i$$

Regressão linear simples - Otimização via GD





Regressão linear simples

Gradiente Descendente Estocástico (SGD, stochastic gradient descent)

1. Escolha um valor α positivo e pequeno.
2. Inicialize os parâmetros do modelo na iteração $t = 0$.
3. Repita por diversos ciclos (épocas):
   1. Permute aleatoriamente a ordem dos dados.
   2. Para cada padrão de entrada, $i = 1, \ldots, N$, repita:
      1. Faça $t \leftarrow t + 1$.
      2. Calcule os erros do modelo:

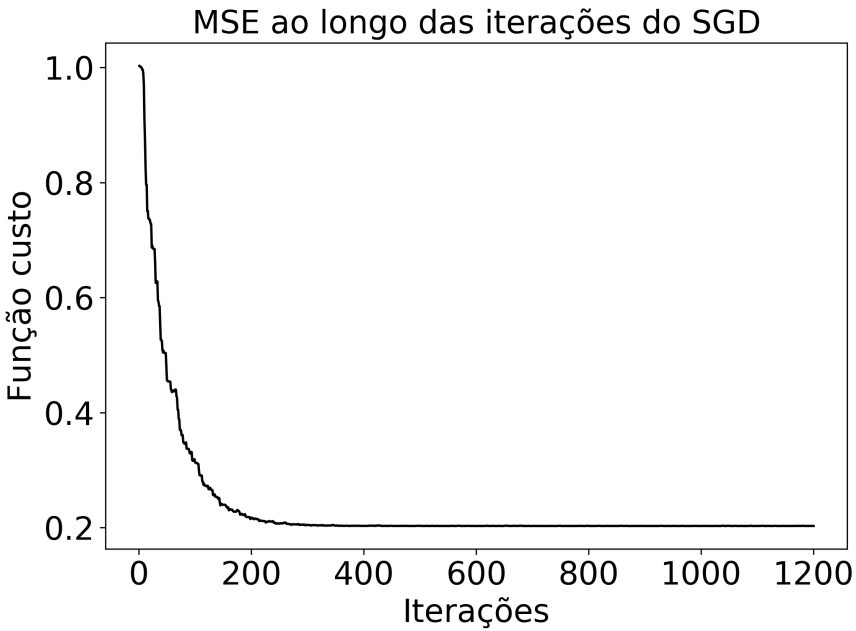
$$\hat{y}_i(t-1) = w_0(t-1) + w_1(t-1)x_i,$$
$$e_i(t-1) = y_i - \hat{y}_i(t-1).$$

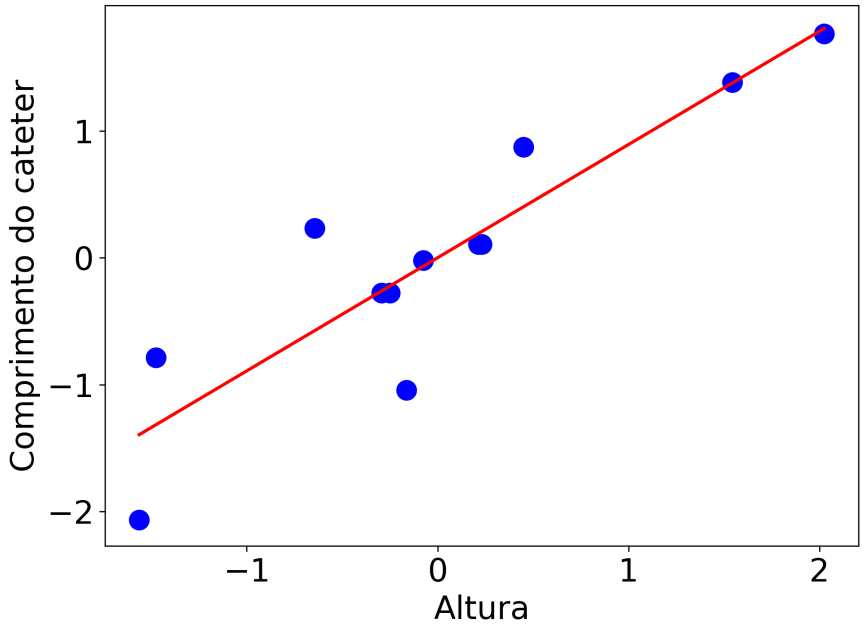
      3. Atualize os parâmetros:

$$w_0(t) = w_0(t-1) + \alpha e_i(t-1)$$
$$w_1(t) = w_1(t-1) + \alpha e_i(t-1)x_i$$

- Também chamado de **algoritmo LMS** (*Least Mean Squares*).

Regressão linear simples - Otimização via SGD





Regressão linear múltipla

- Podemos reconsiderar o problema inserindo o peso do paciente:

Altura (m)	**Peso (Kg)**	**Comprimento (cm)**
1.087	18.141	37
1.613	42.404	50
0.953	16.100	34
1.003	13.605	36
1.156	23.583	43
0.978	7.710	28
1.092	17.460	37
0.572	3.855	20
0.940	14.966	34
0.597	4.308	30
0.838	9.524	38
1.473	35.828	47

Regressão linear múltipla

- Podemos reconsiderar o problema inserindo o peso do paciente:

Altura (m)	**Peso (Kg)**	**Comprimento (cm)**
1.087	18.141	37
1.613	42.404	50
0.953	16.100	34
1.003	13.605	36
1.156	23.583	43
0.978	7.710	28
1.092	17.460	37
0.572	3.855	20
0.940	14.966	34
0.597	4.308	30
0.838	9.524	38
1.473	35.828	47

- Novo modelo linear **múltiplo**:

$$\hat{y}_i = w_0 + w_1 x_{i1} + w_2 x_{i2}$$

- x_{i1} é a i-ésima **Altura** e x_{i2} é o i-ésimo **Peso**.

Regressão linear múltipla

- Caso façamos $x_{i0} = 1$, temos:

$$\hat{y}_i = w_0 x_{i0} + w_1 x_{i1} + w_2 x_{i2}$$
$$\hat{y}_i = \boldsymbol{w}^\top \boldsymbol{x}_i.$$

- Note que:

$$\boldsymbol{w} = [w_0, w_1, w_2]^\top,$$
$$\boldsymbol{x}_i = [1, x_{i1}, x_{i2}]^\top.$$

Regressão linear múltipla

Gradiente Descendente

- Regra de atualização:

$$\boldsymbol{w}(t) = \boldsymbol{w}(t-1) + \alpha \frac{1}{N} \sum_{i=1}^{N} e_i(t-1)\boldsymbol{x}_i$$

Gradiente Descendente Estocástico

- Regra de atualização:

$$\boldsymbol{w}(t) = \boldsymbol{w}(t-1) + \alpha e_i(t-1)\boldsymbol{x}_i$$

Diferença probabilística entre o GD e o SGD

- O erro quadrático $\mathcal{J}$ é uma **variável aleatória**:

$$\min_{\boldsymbol{w}} \frac{1}{2}\mathbb{E}\{\mathcal{J}(\boldsymbol{w})\}, \quad \mathcal{J}(\boldsymbol{w}) = e^2 = (y - \hat{y})^2,$$

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \frac{\alpha}{2}\frac{\partial}{\partial \boldsymbol{w}}\mathbb{E}\{\mathcal{J}\}$$

Diferença probabilística entre o GD e o SGD

- O erro quadrático $\mathcal{J}$ é uma **variável aleatória**:

$$\min_{\boldsymbol{w}} \frac{1}{2}\mathbb{E}\{\mathcal{J}(\boldsymbol{w})\}, \quad \mathcal{J}(\boldsymbol{w}) = e^2 = (y - \hat{y})^2,$$

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \frac{\alpha}{2}\frac{\partial}{\partial \boldsymbol{w}}\mathbb{E}\{\mathcal{J}\}$$

- A **média amostral** resulta no algoritmo GD:

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \frac{\alpha}{2}\frac{\partial}{\partial \boldsymbol{w}}\left(\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right)$$

Diferença probabilística entre o GD e o SGD

- O erro quadrático $\mathcal{J}$ é uma **variável aleatória**:

$$\min_{\boldsymbol{w}} \frac{1}{2}\mathbb{E}\{\mathcal{J}(\boldsymbol{w})\}, \quad \mathcal{J}(\boldsymbol{w}) = e^2 = (y - \hat{y})^2,$$

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \frac{\alpha}{2}\frac{\partial}{\partial \boldsymbol{w}}\mathbb{E}\{\mathcal{J}\}$$

- A **média amostral** resulta no algoritmo GD:

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \frac{\alpha}{2}\frac{\partial}{\partial \boldsymbol{w}}\left(\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right)$$

- Uma **aproximação estocástica** resulta no algoritmo SGD:

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \frac{\alpha}{2}\frac{\partial}{\partial \boldsymbol{w}}(y_i - \hat{y}_i)^2$$

Agenda

Regressão linear analítica

- Reunimos todos os padrões de entrada $\boldsymbol{x}_i$ em uma matriz $\boldsymbol{X}$:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \cdots & \boldsymbol{x}_N \end{bmatrix}^\top \in \mathbb{R}^{N \times (D+1)}.$$

  - N é o número de observações/amostras/vetores/padrões.
  - D é a dimensão da entrada (excluindo o termo $x_{i0} = 1$).

- Agrupamos as saídas disponíveis em um vetor $\boldsymbol{y}$:

$$\boldsymbol{y} = \begin{bmatrix} y_1, y_2, \cdots, y_N \end{bmatrix}^\top \in \mathbb{R}^N.$$

- Agrupamos as saídas do modelo em um vetor $\hat{\boldsymbol{y}}$:

$$\hat{\boldsymbol{y}} = \begin{bmatrix} \hat{y}_1, \hat{y}_2, \cdots, \hat{y}_N \end{bmatrix}^\top \in \mathbb{R}^N.$$

- Agrupamos os parâmetros em um vetor $\boldsymbol{w}$:

$$\boldsymbol{w} = \begin{bmatrix} w_0, w_1, \cdots, w_D \end{bmatrix}^\top \in \mathbb{R}^{D+1}.$$

Regressão linear analítica

- Dados do problema do cateterismo cardíaco em formato matricial:

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1.087 & 18.141 \\ 1 & 1.613 & 42.404 \\ 1 & 0.953 & 16.100 \\ 1 & 1.003 & 13.605 \\ 1 & 1.156 & 23.583 \\ 1 & 0.978 & 7.710 \\ 1 & 1.092 & 17.460 \\ 1 & 0.572 & 3.855 \\ 1 & 0.940 & 14.966 \\ 1 & 0.597 & 4.308 \\ 1 & 0.838 & 9.524 \\ 1 & 1.473 & 35.828 \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} 37 \\ 50 \\ 34 \\ 36 \\ 43 \\ 28 \\ 37 \\ 20 \\ 34 \\ 30 \\ 38 \\ 47 \end{bmatrix}$$

Regressão linear analítica

- Reformulamos nosso modelo linear na forma matricial:

$$\hat{y}_i = \boldsymbol{w}^\top \boldsymbol{x}_i,$$
$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{w}.$$

Regressão linear analítica

- Reformulamos nosso modelo linear na forma matricial:

$$\hat{y}_i = \boldsymbol{w}^\top \boldsymbol{x}_i,$$
$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{w}.$$

- Reformulamos também a função custo:

$$\mathcal{J}(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top(\boldsymbol{y} - \hat{\boldsymbol{y}})$$
$$\mathcal{J}(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})$$

Regressão linear analítica

- O mínimo de $\mathcal{J}(\boldsymbol{w})$ ocorrerá em $\frac{\partial \mathcal{J}(\boldsymbol{w})}{\partial \boldsymbol{w}} = 0$:

$$\begin{aligned}
\frac{\partial \mathcal{J}(\boldsymbol{w})}{\partial \boldsymbol{w}} &= \frac{1}{2} 2(-\boldsymbol{X}^\top)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \\
-\boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} &= 0 \\
\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} &= \boldsymbol{X}^\top \boldsymbol{y} \\
\boldsymbol{w} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.
\end{aligned}$$

Regressão linear analítica

- O mínimo de $\mathcal{J}(\boldsymbol{w})$ ocorrerá em $\frac{\partial \mathcal{J}(\boldsymbol{w})}{\partial \boldsymbol{w}} = 0$:

$$\begin{aligned}
\frac{\partial \mathcal{J}(\boldsymbol{w})}{\partial \boldsymbol{w}} &= \frac{1}{2} 2(-\boldsymbol{X}^\top)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \\
-\boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} &= 0 \\
\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{w} &= \boldsymbol{X}^\top \boldsymbol{y} \\
\boldsymbol{w} &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}.
\end{aligned}$$

- $\boldsymbol{X}^+ = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$, em que $\boldsymbol{X}^+ \boldsymbol{X} = \boldsymbol{I}$, é chamada de **inversa de Moore-Penrose** ou **pseudo-inversa**.

Regressão linear analítica

Método dos mínimos quadrados ordinários (OLS, *ordinary least squares*)

- O vetor de parâmetros $\boldsymbol{w}$ que minimiza $\mathcal{J}(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top(\boldsymbol{y} - \hat{\boldsymbol{y}})$ é dado por

$$\boldsymbol{w} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y},$$

em quem $\boldsymbol{X}$ é a matriz de vetores de entrada (um por linha) e $\boldsymbol{y}$ é o vetor de saídas desejadas.

Regressão linear analítica

- OLS equivale ao método de Newton aplicado na função de custo quadrática $\mathcal{J}$:

$$\boldsymbol{w} = \boldsymbol{w}_0 - \left(\frac{\partial^2 \mathcal{J}(\boldsymbol{w}_0)}{\partial \boldsymbol{w}_0^2}\right)^{-1} \frac{\partial \mathcal{J}(\boldsymbol{w}_0)}{\partial \boldsymbol{w}_0},$$

$$\frac{\partial \mathcal{J}(\boldsymbol{w}_0)}{\partial \boldsymbol{w}_0} = -\boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_0),$$

$$\frac{\partial^2 \mathcal{J}(\boldsymbol{w}_0)}{\partial \boldsymbol{w}_0^2} = \boldsymbol{X}^\top \boldsymbol{X},$$

$$\boldsymbol{w} = \boldsymbol{w}_0 - (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \left(-\boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_0)\right)$$

- Podemos escolher $\boldsymbol{w}_0 = \boldsymbol{0}$ para obter o mínimo global.

$$\boldsymbol{w} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

Regressão linear analítica

- De onde vem a função custo $\mathcal{J}(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top(\boldsymbol{y} - \hat{\boldsymbol{y}})$?

Regressão linear analítica

- De onde vem a função custo $\mathcal{J}(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top(\boldsymbol{y} - \hat{\boldsymbol{y}})$?
- Considerando um ruído independente $\epsilon \sim \mathcal{N}(0, \sigma^2)$:

$$
\begin{aligned}
y_i &= \hat{y}_i + \epsilon = \boldsymbol{w}^\top \boldsymbol{x}_i + \epsilon, \\
p(y_i|\boldsymbol{x}_i, \boldsymbol{w}) &= \mathcal{N}(y_i|\boldsymbol{w}^\top \boldsymbol{x}_i, \sigma^2), \\
p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) &= \prod_{i=1}^{N} \mathcal{N}(y_i|\boldsymbol{w}^\top \boldsymbol{x}_i, \sigma^2) \\
\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) &= \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2}{2\sigma^2}\right) \\
\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) &= \underbrace{-\frac{N}{2}\log(2\pi\sigma^2)}_{\text{const. em relação a } \boldsymbol{w}} - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2
\end{aligned}
$$

Regressão linear analítica

- Queremos maximizar $\mathcal{L}(\boldsymbol{w}) = \log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})$, o que equivale a minimizar $\mathcal{J}(\boldsymbol{w}) = -\mathcal{L}(\boldsymbol{w})$.
- Ignorando os termos constantes:

$$\mathcal{J}(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2$$

$$\mathcal{J}(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})$$

$$\mathcal{J}(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top(\boldsymbol{y} - \hat{\boldsymbol{y}})$$

Regressão linear analítica

- Queremos maximizar $\mathcal{L}(\boldsymbol{w}) = \log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})$, o que equivale a minimizar $\mathcal{J}(\boldsymbol{w}) = -\mathcal{L}(\boldsymbol{w})$.
- Ignorando os termos constantes:

$$\mathcal{J}(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2$$

$$\mathcal{J}(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})$$

$$\mathcal{J}(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{y} - \hat{\boldsymbol{y}})^\top(\boldsymbol{y} - \hat{\boldsymbol{y}})$$

Solução de máxima verossimilhança

A solução obtida via OLS (e aproximada via GD e SGD), chamada de solução de **máxima verossimilhança (maximum likelihood)**, é ótima quando o ruído é Gaussiano:

$$\boldsymbol{w}_{\text{OLS}} = \boldsymbol{w}_{\text{ML}} = \arg\max \log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})$$

Agenda

Tópicos adicionais

- Viés de indução (inductive bias).

Tópicos adicionais

- Viés de indução (inductive bias).
- Estimação de parâmetros via solução de máximo a posteriori (MAP, maximum a posteriori):

$$\boldsymbol{w}_{\text{MAP}} = \arg\max[\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) + \log p(\boldsymbol{w})].$$

Tópicos adicionais

- Viés de indução (inductive bias).
- Estimação de parâmetros via solução de máximo a posteriori (MAP, maximum a posteriori):

$$\boldsymbol{w}_{\text{MAP}} = \arg\max[\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) + \log p(\boldsymbol{w})].$$

- (algumas) Alternativas ao OLS:
  - → Weighted LS - ruído heteroscedástico:

$$\boldsymbol{w} = (\boldsymbol{X}^\top \boldsymbol{\Lambda} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Lambda} \boldsymbol{y}, \quad \boldsymbol{\Lambda} = \text{diag}(1/\sigma^2(\boldsymbol{x}_i)).$$

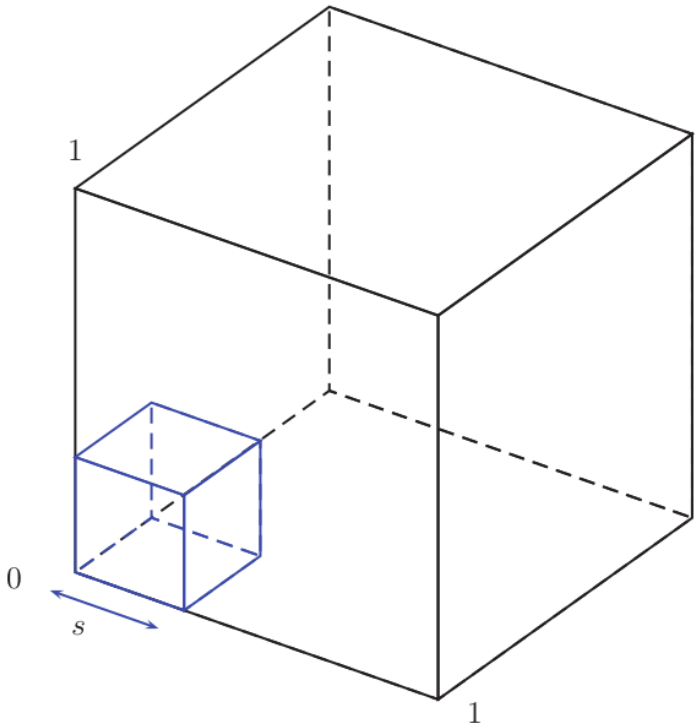
  - → Generalized LS - ruído correlacionado:

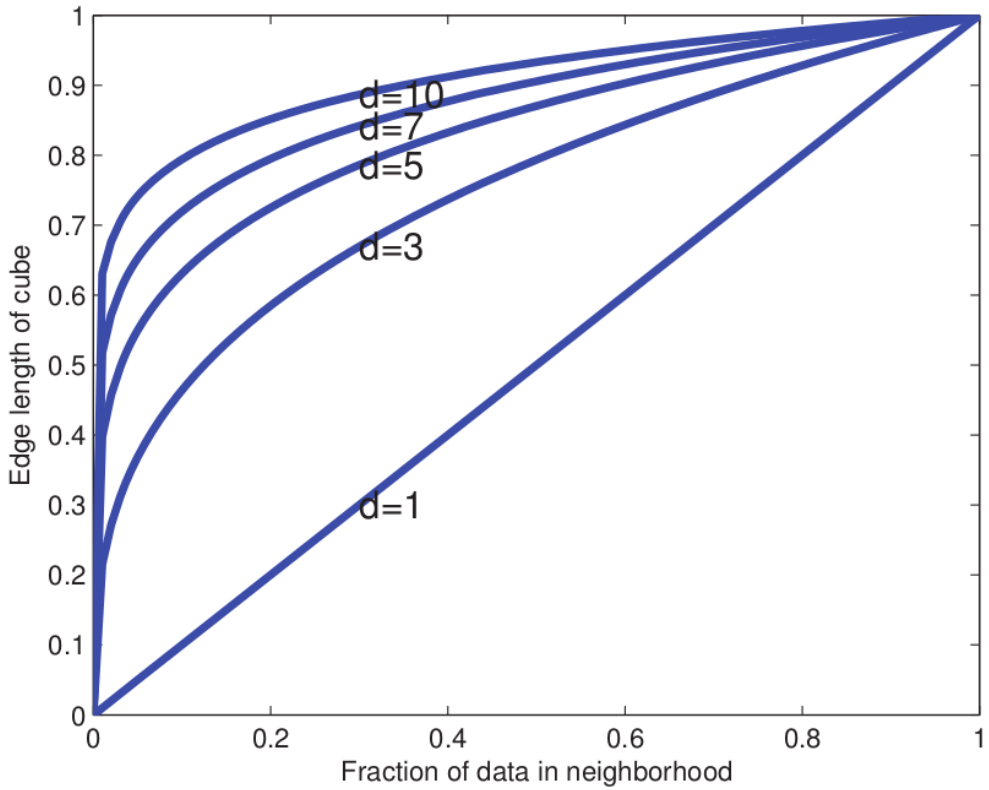
$$\boldsymbol{w} = (\boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\Omega} \boldsymbol{y}, \quad \boldsymbol{\Omega} = \text{cov}(\boldsymbol{\epsilon}|\boldsymbol{X}).$$

  - → Iterative Reweighted LS: WLS iterativo para ruído não Gaussiano.

Tópicos adicionais

- Maldição da dimensionalidade (curse of dimensionality).





Agenda

Referências bibliográficas

- **Cap. 9** - DEISENROTH, M. *et al.* **Mathematics for machine learning**. 2019.
- **Caps. 1 e 7** - MURPHY, Kevin P. **Machine learning: a probabilistic perspective**, 2012.
- **Caps. 1 e 11** - MURPHY, Kevin P. **Probabilistic Machine Learning: An Introduction**, 2021.
- **Caps. 2 e 3** - HAYKIN, Simon. **Neural Networks and Learning Machines**, 3ed., 2010.
- **Cap. 3** - BISHOP, Christopher M. **Pattern recognition and machine learning**, 2006.