



Universidade de Fortaleza - UNIFOR

---

# Sistemas Inteligentes - T951

**Msc. Prof. Paulo Cirillo Souza Barbosa**

Centro de Ciências Tecnológicas - CCT

Universidade de Fortaleza

Fortaleza, Ceará, Brasil

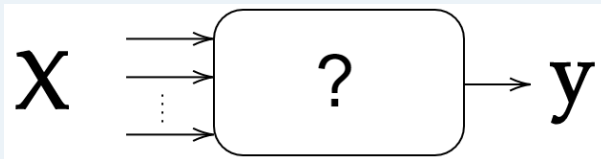
10 de setembro de 2023



- 1 Aprendizado Supervisionado (Algoritmos Lineares).
  - 1.1 Regressão.
- 2 Regressão Linear Simples/ Múltipla.
- 3 Estimação através do método dos Mínimos Quadrados Ordinário.
- 4 Equações normais dos mínimos quadrados
  - 4.1 Forma Escalar



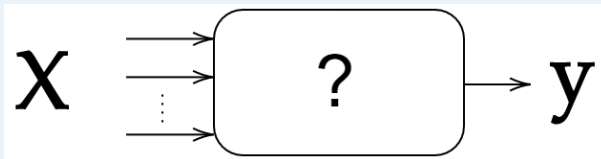
## Regressão × Classificação



- Tem-se uma saída  $y$  que pode ser:
  - 1 Quantitativa
  - 2 Qualitativa (classes, categorias, fatores, variáveis discretas)



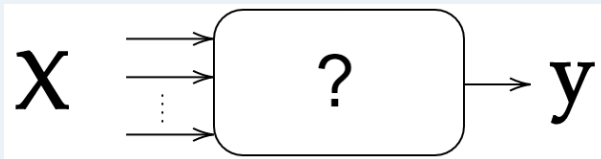
## Regressão × Classificação



- Tem-se uma saída  $y$  que pode ser:
  - 1 Quantitativa
  - 2 Qualitativa (classes, categorias, fatores, variáveis discretas)
- Para ambos casos, o que é necessário para prever a saída do sistema?



## Regressão × Classificação



- Tem-se uma saída  $y$  que pode ser:
  - 1 Quantitativa
  - 2 Qualitativa (classes, categorias, fatores, variáveis discretas)
- Para ambos casos, o que é necessário para prever a saída do sistema?
  - 1 **Ex1:** Dado alguma observação atmosférica do presente e passado, deseja-se prever o **nível** de ozônio amanhã!



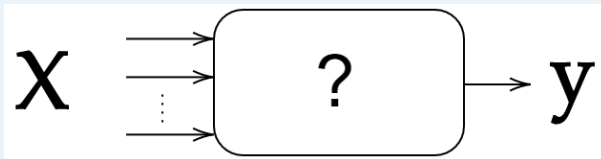
## Regressão × Classificação



- Tem-se uma saída  $y$  que pode ser:
  - 1 Quantitativa
  - 2 Qualitativa (classes, categorias, fatores, variáveis discretas)
- Para ambos casos, o que é necessário para prever a saída do sistema?
  - 1 **Ex1:** Dado alguma observação atmosférica do presente e passado, deseja-se prever o **nível** de ozônio amanhã!
  - 2 **Ex1:** Dado valores de intensidade de pixels em escala de cinza de imagens digitalizadas de dígitos manuscritos, deseja-se prever qual **rótulo da classe**.



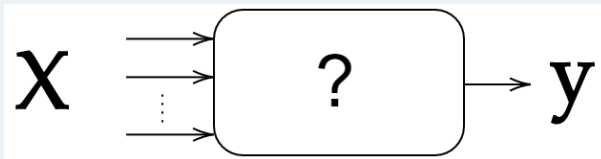
## Regressão × Classificação



- Tem-se uma saída  $y$  que pode ser:
  - 1 Quantitativa
  - 2 Qualitativa (classes, categorias, fatores, variáveis discretas)
- Para ambos casos, o que é necessário para prever a saída do sistema?
  - 1 **Ex1:** Dado alguma observação atmosférica do presente e passado, deseja-se prever o **nível** de ozônio amanhã! (**REGRESSÃO**)
  - 2 **Ex1:** Dado valores de intensidade de *pixels* em escala de cinza de imagens digitalizadas de dígitos manuscritos, deseja-se prever qual **rótulo da classe**. (**CLASSIFICAÇÃO**)



## Regressão × Classificação

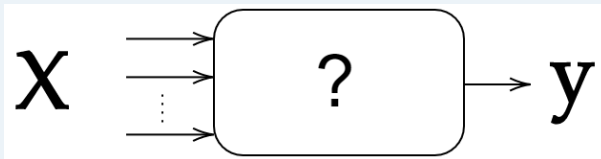


- Tratando de regressões (inicialmente pelo menos):
  - 1 Necessita-se dos dados de treinamento.





## Regressão × Classificação



- Tratando de regressões (inicialmente pelo menos):
  - 1 Necessita-se dos dados de treinamento.
  - 2 Observam-se as saídas para um conjunto de amostras.



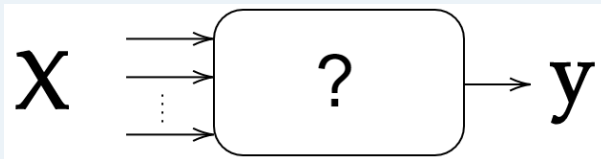
## Regressão × Classificação



- Tratando de regressões (inicialmente pelo menos):
  - 1 Necessita-se dos dados de treinamento.
  - 2 Observam-se as saídas para um conjunto de amostras.
  - 3 Utiliza-se estes dados para construir o modelo preditivo.



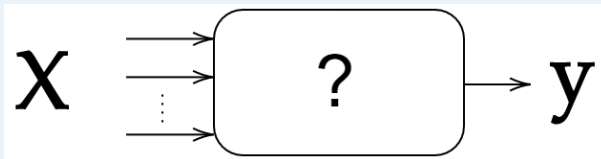
## Regressão × Classificação



- Tratando de regressões (inicialmente pelo menos):
  - 1 Necessita-se dos dados de treinamento.
  - 2 Observam-se as saídas para um conjunto de amostras.
  - 3 Utiliza-se estes dados para construir o modelo preditivo.
  - 4 Este no que lhe concerne, pode realizar predições para amostras **desconhecidas**.



## Regressão × Classificação



- Tratando de regressões (inicialmente pelo menos):
  - 1 Necessita-se dos dados de treinamento.
  - 2 Observam-se as saídas para um conjunto de amostras.
  - 3 Utiliza-se estes dados para construir o modelo preditivo.
  - 4 Este no que lhe concerne, pode realizar previsões para amostras **desconhecidas**.
  - 5 **Obs:** Hipótese de correlações entre entrada saída. **Spurious Correlations**
- Este trata-se de um caso de **aprendizado supervisionado**?



## Regressão × Classificação



- Tratando de regressões (inicialmente pelo menos):
  - 1 Necessita-se dos dados de treinamento.
  - 2 Observam-se as saídas para um conjunto de amostras.
  - 3 Utiliza-se estes dados para construir o modelo preditivo.
  - 4 Este no que lhe concerne, pode realizar previsões para amostras **desconhecidas**.
  - 5 **Obs:** Hipótese de correlações entre entrada saída. **Spurious Correlations**
- Este trata-se de um caso de **aprendizado supervisionado**?
  - 1 Sim! Pois, há a presença das previsões conhecidas no conjunto, para guiar o processo de treinamento do modelo.



## Regressão Linear Simples/ Múltipla

- Técnica cujo objetivo principal reside justamente na investigação das relações entre variáveis e na modelagem matemática.



## Regressão Linear Simples/ Múltipla

- Técnica cujo objetivo principal reside justamente na investigação das relações entre variáveis e na modelagem matemática.
  - ① Pode ser usada na construção de um modelo que expressa o resultado de uma variável como função de uma ou mais variáveis.



## Regressão Linear Simples/ Múltipla

- Técnica cujo objetivo principal reside justamente na investigação das relações entre variáveis e na modelagem matemática.
  - ① Pode ser usada na construção de um modelo que expressa o resultado de uma variável como função de uma ou mais variáveis.
  - ② Tal modelo pode então, ser usado para prever o resultado de uma variável em função de outra.





## Regressão Linear Simples/ Múltipla

- Técnica cujo objetivo principal reside justamente na investigação das relações entre variáveis e na modelagem matemática.
  - ① Pode ser usada na construção de um modelo que expressa o resultado de uma variável como função de uma ou mais variáveis.
  - ② Tal modelo pode então, ser usado para prever o resultado de uma variável em função de outra.
- Assume-se que exista uma única variável dependente  $y \in \mathbb{R}$ , relacionada com  $p$  variáveis independentes, ou regressoras  $x_1, x_2, \dots, x_p \in \mathbb{R}$ .



## Regressão Linear Simples/ Múltipla

- Técnica cujo objetivo principal reside justamente na investigação das relações entre variáveis e na modelagem matemática.
  - ① Pode ser usada na construção de um modelo que expressa o resultado de uma variável como função de uma ou mais variáveis.
  - ② Tal modelo pode então, ser usado para prever o resultado de uma variável em função de outra.
- Assume-se que exista uma única variável dependente  $y \in \mathbb{R}$ , relacionada com  $p$  variáveis independentes, ou regressoras  $x_1, x_2, \dots, x_p \in \mathbb{R}$ .
  - ①  $y$  trata-se de uma variável aleatória.
  - ② As variáveis regressoras são medidas com erro desprezível (controladas pelo experimentador).
- O Modelo **teórico** que relaciona  $y$  e as variáveis regressoras, pode ser escrito como

$$y = f(x_1, x_2, \dots, x_p | \beta) + \epsilon = f(\mathbf{x} | \beta) + \epsilon$$



## Regressão Linear Simples/ Múltipla

$$y = f(x_1, x_2, \dots, x_p | \beta) + \epsilon = f(\mathbf{x} | \beta) + \epsilon$$

- $f(\cdot | \cdot)$  é denominada a função de regressão.
- $\mathbf{x} \in \mathbb{R}^p$  é o vetor de variáveis regressoras.
- $\beta \in \mathbb{R}^p$  é o vetor de parâmetros da função regressora.
- $\epsilon$  denota o erro aleatório (ruído) presentes na medição de  $y$ .
- O que implica esse modelo ser considerado **teórico**?



## Regressão Linear Simples/ Múltipla

$$y = f(x_1, x_2, \dots, x_p | \beta) + \epsilon = f(\mathbf{x} | \beta) + \epsilon$$

- $f(\cdot | \cdot)$  é denominada a função de regressão.
- $\mathbf{x} \in \mathbb{R}^p$  é o vetor de variáveis regressoras.
- $\beta \in \mathbb{R}^p$  é o vetor de parâmetros da função regressora.
- $\epsilon$  denota o erro aleatório (ruído) presentes na medição de  $y$ .
- O que implica esse modelo ser considerado **teórico**?
  - 1  $f(\cdot | \cdot)$  e a componente aleatória são DESCONHECIDAS!
- A forma funcional da equação de regressão  $f(\cdot)$  é realizada com base em informação **à priori**:
  - 1 Conhecimento prévio.
  - 2 Experimentação com diferentes formas funcionais.



## Regressão Linear Simples/ Múltipla

- Independente da forma funcional da equação de regressão, os seus parâmetros (coeficientes) precisam ser estimados.
- O que é necessário para fazer esta estimação??



## Regressão Linear Simples/ Múltipla

- Independente da forma funcional da equação de regressão, os seus parâmetros (coeficientes) precisam ser estimados.
- O que é necessário para fazer esta estimação??
- **DADOS:** um conjunto de  $N$  valores de  $y$  e suas variáveis regressoras  $\{x_1, x_2, \dots, x_p\}$ :

$$(y_i, x_{i1}, x_{i2}, \dots, x_{ip}), i = 1, \dots, N$$

$$(\mathbf{x}_i, y_i), i = 1, \dots, N$$

- A estimação de  $\beta$  é simbolizada por  $\hat{\beta}$ , que é utilizado na seguinte equação para novas predições:

$$\hat{y} = \hat{f}(x_1, x_2, \dots, x_p | \hat{\beta}) = \hat{f}(\mathbf{x} | \hat{\beta})$$



## Regressão Linear Simples/ Múltipla

- A análise de regressão é dita linear quando se assume que a relação matemática entre variáveis de interesse é uma função linear dos seus parâmetros.



## Regressão Linear Simples/ Múltipla

- A análise de regressão é dita linear quando se assume que a relação matemática entre variáveis de interesse é uma função linear dos seus parâmetros.
- Neste caso, o modelo passa a ser chamado de *regressão linear simples/múltipla*.





## Regressão Linear Simples/ Múltipla

- A análise de regressão é dita linear quando se assume que a relação matemática entre variáveis de interesse é uma função linear dos seus parâmetros.
- Neste caso, o modelo passa a ser chamado de *regressão linear simples/múltipla*.
- Quando o problema de regressão linear envolve apenas uma única variável regressora  $x$ , tem-se uma *regressão linear simples*.



## Regressão Linear Simples/ Múltipla

- A análise de regressão é dita linear quando se assume que a relação matemática entre variáveis de interesse é uma função linear dos seus parâmetros.
- Neste caso, o modelo passa a ser chamado de *regressão linear simples/múltipla*.
- Quando o problema de regressão linear envolve apenas uma única variável regressora  $x$ , tem-se uma *regressão linear simples*.
- Neste caso, a relação matemática entre uma única variável de entrada  $x$  e uma variável de saída  $y$  é definida por uma reta, como

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$



## Regressão Linear Simples/ Múltipla

- A análise de regressão é dita linear quando se assume que a relação matemática entre variáveis de interesse é uma função linear dos seus parâmetros.
- Neste caso, o modelo passa a ser chamado de *regressão linear simples/múltipla*.
- Quando o problema de regressão linear envolve apenas uma única variável regressora  $x$ , tem-se uma *regressão linear simples*.
- Neste caso, a relação matemática entre uma única variável de entrada  $x$  e uma variável de saída  $y$  é definida por uma reta, como

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

- Nesta, o  $\beta_0$  é o intercepto (*intercept*), e  $\beta_1$  é a inclinação (*slope*) da reta.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$



## Regressão Linear Simples/ Múltipla

- Para o caso de uma regressão linear múltipla, o problema pode ser escrito como

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon =$$



## Regressão Linear Simples/ Múltipla

- Para o caso de uma regressão linear múltipla, o problema pode ser escrito como

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon = \boldsymbol{\beta}^T \mathbf{x} + \epsilon$$

- Neste caso, o vetor de parâmetros  $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$
- Como consequência disto, a primeira componente do vetor  $\mathbf{x} \in \mathbb{R}^{p+1}$  é igual a 1. Qual a interpretação geométrica disto?



## Regressão Linear Simples/ Múltipla

- Para o caso de uma regressão linear múltipla, o problema pode ser escrito como

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon = \boldsymbol{\beta}^T \mathbf{x} + \epsilon$$

- Neste caso, o vetor de parâmetros  $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$
- Como consequência disto, a primeira componente do vetor  $\mathbf{x} \in \mathbb{R}^{p+1}$  é igual a 1. Qual a interpretação geométrica disto? Quais os valores das demais componentes?



## Regressão Linear Simples/ Múltipla

- Para o caso de uma regressão linear múltipla, o problema pode ser escrito como

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon = \boldsymbol{\beta}^T \mathbf{x} + \epsilon$$

- Neste caso, o vetor de parâmetros  $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$
- Como consequência disto, a primeira componente do vetor  $\mathbf{x} \in \mathbb{R}^{p+1}$  é igual a 1. Qual a interpretação geométrica disto? Quais os valores das demais componentes?
- **Observações importantes:**
  - 1 Estes modelos de regressão linear simples ou múltipla, são utilizados como **funções aproximadoras** e a equação de regressão é ajustada ao conjunto de pares  $(\mathbf{x}_i, y_i), i = 1, \cdots N$ .
  - 2 E a verdadeira equação que relaciona os conjuntos de pares?



## Regressão Linear Simples/ Múltipla

- Para o caso de uma regressão linear múltipla, o problema pode ser escrito como

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon = \boldsymbol{\beta}^T \mathbf{x} + \epsilon$$

- Neste caso, o vetor de parâmetros  $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$
- Como consequência disto, a primeira componente do vetor  $\mathbf{x} \in \mathbb{R}^{p+1}$  é igual a 1. Qual a interpretação geométrica disto? Quais os valores das demais componentes?
- **Observações importantes:**
  - 1 Estes modelos de regressão linear simples ou múltipla, são utilizados como **funções aproximadoras** e a equação de regressão é ajustada ao conjunto de pares  $(\mathbf{x}_i, y_i), i = 1, \cdots N$ .
  - 2 E a verdadeira equação que relaciona os conjuntos de pares? É **DESCONHECIDA**.





## Regressão Linear Simples/ Múltipla - Implementação exemplo 1 - O QUE É O BIAS?



## Regressão Linear Simples/ Múltipla - Implementação exemplo 1 - O QUE É O BIAS?

```
1 import numpy as np
2 import matplotlib.pyplot as plt
3 plt.figure(0)
4 x = np.linspace(-10,10,100)
5 for i in range(10):
6     y = np.random.randn()*x
7     plt.plot(x,y)
8 plt.grid(True)
9 plt.title("Sem intercepto")
10
11 plt.figure(1)
12 for i in range(10):
13     y = np.random.randn()*x + np.random.randn()
14     plt.plot(x,y)
15 plt.grid(True)
16 plt.title("Com intercepto")
17 plt.show()
```



## Regressão Linear Simples/ Múltipla

- De maneira similar ao modelo de regressão linear simples, o modelo estimado que realiza novas previsões é:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots \hat{\beta}_n x_n = \beta_0 + \sum_{i=1}^p \beta_i x_i = \hat{\beta}^T \mathbf{x}$$



## Regressão Linear Simples/ Múltipla

- De maneira similar ao modelo de regressão linear simples, o modelo estimado que realiza novas predições é:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots \hat{\beta}_n x_n = \beta_0 + \sum_{i=1}^p \beta_i x_i = \hat{\beta}^T \mathbf{x}$$

- Em comparação ao modelo linear simples, este produz um hiperplano que ajusta as variáveis regressoras.
- Qual interpretação geométrica para este ajuste em ambos os modelos?
- Quais vantagens em se utilizar este modelo linear?
  - 1 Simplicidade do modelo.
  - 2 Interpretação direta ao parâmetro  $\beta_j$ . Permite a identificação direta de quais variáveis regressoras influenciam mais a variável de resposta.



### Estimação através do método dos Mínimos Quadrados Ordinário.

- A estimação do modelo, consequentemente dos coeficientes de  $\hat{\beta}$ , é feita através do método dos Mínimos Quadrados Ordinário (MQO) ou do inglês, *Ordinary Least Squares (OLS)*
- **Formalização (revisão) do problema.** Para estimação do modelo (vetor de parâmetros  $\hat{\beta}$ ), precisa-se:
  - ①  $N$  observações (amostras) do par entrada-saída  $(\mathbf{x}_i, y_i), i = 1, \dots, N$ .



### Estimação através do método dos Mínimos Quadrados Ordinário.

- A estimação do modelo, consequentemente dos coeficientes de  $\hat{\beta}$ , é feita através do método dos Mínimos Quadrados Ordinário (MQO) ou do inglês, *Ordinary Least Squares (OLS)*
- **Formalização (revisão) do problema.** Para estimação do modelo (vetor de parâmetros  $\hat{\beta}$ ), precisa-se:
  - ①  $N$  observações (amostras) do par entrada-saída  $(\mathbf{x}_i, y_i), i = 1, \dots, N$ .
  - ② Em outras palavras, a  $i$ -ésima observação tem como resposta um escalar  $y_i$  a partir do vetor de variáveis correspondentes  $\mathbf{x}_i = [x_{i1} \ x_{i2} \ \dots \ x_{ip}]^T$ , em que  $x_{ij}$  representa a  $j$ -ésima variável regressora da  $i$ -ésima amostra  $j - i = \{1, \dots, N\}$  e  $j = \{1, \dots, p\}$



### Estimação através do método dos Mínimos Quadrados Ordinário.

- A estimação do modelo, consequentemente dos coeficientes de  $\hat{\beta}$ , é feita através do método dos Mínimos Quadrados Ordinário (MQO) ou do inglês, *Ordinary Least Squares (OLS)*
- **Formalização (revisão) do problema.** Para estimação do modelo (vetor de parâmetros  $\hat{\beta}$ ), precisa-se:
  - ①  $N$  observações (amostras) do par entrada-saída  $(\mathbf{x}_i, y_i), i = 1, \dots, N$ .
  - ② Em outras palavras, a  $i$ -ésima observação tem como resposta um escalar  $y_i$  a partir do vetor de variáveis correspondentes  $\mathbf{x}_i = [x_{i1} \ x_{i2} \ \dots \ x_{ip}]^T$ , em que  $x_{ij}$  representa a  $j$ -ésima variável regressora da  $i$ -ésima amostra  $j - i = \{1, \dots, N\}$  e  $j = \{1, \dots, p\}$
  - ③ Assume-se que existam ( **muito** ) mais observações do que incógnitas ( $N \gg p$ ).



### Estimação através do método dos Mínimos Quadrados Ordinário.

- A estimação do modelo, consequentemente dos coeficientes de  $\hat{\beta}$ , é feita através do método dos Mínimos Quadrados Ordinário (MQO) ou do inglês, *Ordinary Least Squares (OLS)*
- **Formalização (revisão) do problema.** Para estimação do modelo (vetor de parâmetros  $\hat{\beta}$ ), precisa-se:
  - 1  $N$  observações (amostras) do par entrada-saída  $(\mathbf{x}_i, y_i), i = 1, \dots, N$ .
  - 2 Em outras palavras, a  $i$ -ésima observação tem como resposta um escalar  $y_i$  a partir do vetor de variáveis correspondentes  $\mathbf{x}_i = [x_{i1} \ x_{i2} \ \dots \ x_{ip}]^T$ , em que  $x_{ij}$  representa a  $j$ -ésima variável regressora da  $i$ -ésima amostra  $j - i = \{1, \dots, N\}$  e  $j = \{1, \dots, p\}$
  - 3 Assume-se que existam ( **muito** ) mais observações do que incógnitas ( $N \gg p$ ).
  - 4 Assume-se que o ruído tem média 0 e variância  $\sigma^2$





### Estimação através do método dos Mínimos Quadrados Ordinário.

- A estimação do modelo, consequentemente dos coeficientes de  $\hat{\beta}$ , é feita através do método dos Mínimos Quadrados Ordinário (MQO) ou do inglês, *Ordinary Least Squares (OLS)*
- **Formalização (revisão) do problema.** Para estimação do modelo (vetor de parâmetros  $\hat{\beta}$ ), precisa-se:
  - 1  $N$  observações (amostras) do par entrada-saída  $(\mathbf{x}_i, y_i), i = 1, \dots, N$ .
  - 2 Em outras palavras, a  $i$ -ésima observação tem como resposta um escalar  $y_i$  a partir do vetor de variáveis correspondentes  $\mathbf{x}_i = [x_{i1} \ x_{i2} \ \dots \ x_{ip}]^T$ , em que  $x_{ij}$  representa a  $j$ -ésima variável regressora da  $i$ -ésima amostra  $j - i = \{1, \dots, N\}$  e  $j = \{1, \dots, p\}$
  - 3 Assume-se que existam ( **muito** ) mais observações do que incógnitas ( $N \gg p$ ).
  - 4 Assume-se que o ruído tem média 0 e variância  $\sigma^2$
  - 5 Assume-se que as observações em  $\varepsilon$  são não-correlacionadas.



### Estimação através do método dos Mínimos Quadrados Ordinário.

- A estimação do modelo, consequentemente dos coeficientes de  $\hat{\beta}$ , é feita através do método dos Mínimos Quadrados Ordinário (MQO) ou do inglês, *Ordinary Least Squares (OLS)*
- **Formalização (revisão) do problema.** Para estimação do modelo (vetor de parâmetros  $\hat{\beta}$ ), precisa-se:
  - 1  $N$  observações (amostras) do par entrada-saída  $(\mathbf{x}_i, y_i), i = 1, \dots, N$ .
  - 2 Em outras palavras, a  $i$ -ésima observação tem como resposta um escalar  $y_i$  a partir do vetor de variáveis correspondentes  $\mathbf{x}_i = [x_{i1} \ x_{i2} \ \dots \ x_{ip}]^T$ , em que  $x_{ij}$  representa a  $j$ -ésima variável regressora da  $i$ -ésima amostra  $j - i = \{1, \dots, N\}$  e  $j = \{1, \dots, p\}$
  - 3 Assume-se que existam ( **muito** ) mais observações do que incógnitas ( $N \gg p$ ).
  - 4 Assume-se que o ruído tem média 0 e variância  $\sigma^2$
  - 5 Assume-se que as observações em  $\varepsilon$  são não-correlacionadas.
- **Obs:** A estimação mostrada, é consolidadas para o caso geral (múltiplas variáveis regressoras)



## Estimação através do método dos Mínimos Quadrados Ordinário.

- O problema pode ser expandido, para abranger todas as observações:

$$\beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \epsilon_1 = y_1$$

$$\beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \epsilon_2 = y_2$$

$$\vdots$$

$$\beta_0 + \beta_1 x_{N1} + \beta_2 x_{N2} + \cdots + \beta_p x_{Np} + \epsilon_N = y_N$$

- Em notação matricial, o sistema pode ser descrito como:



## Estimação através do método dos Mínimos Quadrados Ordinário.

- O problema pode ser expandido, para abranger todas as observações:

$$\beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \epsilon_1 = y_1$$

$$\beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \epsilon_2 = y_2$$

$$\vdots$$

$$\beta_0 + \beta_1 x_{N1} + \beta_2 x_{N2} + \cdots + \beta_p x_{Np} + \epsilon_N = y_N$$

- Em notação matricial, o sistema pode ser descrito como:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Nesta, é possível verificar que a variável de resposta é uma função linear das variáveis regressoras.



## Estimação através do método dos Mínimos Quadrados Ordinário.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Em que  $\mathbf{y}$  e  $\boldsymbol{\epsilon} \in \mathbb{R}^N$ ,  $\mathbf{X} \in \mathbb{R}^{N \times (p+1)}$  e  $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}_{N \times 1}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix}_{N \times (p+1)}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}_{(p+1) \times 1}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}_{N \times 1}$$



### Estimação através do método dos Mínimos Quadrados Ordinário.

- A técnica de estimação dos coeficientes de regressão  $\beta$  corresponde a minimização dos **desvios**  $\epsilon_i$  entre os valores observados de  $y_i$  e o hiperplano de regressão. Ou seja, fazer com que a soma dos quadrados dos desvios seja mínima.
- Para uma observação este desvio é:

$$y = \beta_0 + \sum_{j=1}^p \beta_j x_j + \epsilon$$



## Estimação através do método dos Mínimos Quadrados Ordinário.

- A técnica de estimação dos coeficientes de regressão  $\beta$  corresponde a minimização dos **desvios**  $\epsilon_i$  entre os valores observados de  $y_i$  e o hiperplano de regressão. Ou seja, fazer com que a soma dos quadrados dos desvios seja mínima.
- Para uma observação este desvio é:

$$y = \beta_0 + \sum_{j=1}^p \beta_j x_j + \epsilon \qquad \left( y - \left[ \beta_0 + \sum_{j=1}^p \beta_j x_j \right] \right)^2 = \epsilon^2$$

- Normalmente esta equação é conhecida como **função custo**. Então, para todas as observações de treinamento, a estimação dos coeficientes pelo método dos MQO é realizar a minimização da função custo  $J(\cdot)$ :

$$J(\beta_0, \beta_1, \dots, \beta_p) = \sum_{i=1}^N \epsilon^2$$



## Estimação através do método dos Mínimos Quadrados Ordinário.

$$\begin{aligned} J(\beta_0, \beta_1, \dots, \beta_p) &= \sum_{i=1}^n \epsilon^2 \\ &= \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_j \right)^2 \end{aligned}$$

- Ou em sua forma vetorial:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$$

$$J(\boldsymbol{\beta}) = \|\boldsymbol{\epsilon}\|_2^2$$





## Estimação através do método dos Mínimos Quadrados Ordinário.

$$\begin{aligned} J(\beta_0, \beta_1, \dots, \beta_p) &= \sum_{i=1}^n \epsilon^2 \\ &= \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_j \right)^2 \end{aligned}$$

- Ou em sua forma vetorial:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$$

$$J(\boldsymbol{\beta}) = \|\boldsymbol{\epsilon}\|_2^2 = \boldsymbol{\epsilon}^T \boldsymbol{\epsilon}$$



## Estimação através do método dos Mínimos Quadrados Ordinário.

$$\begin{aligned} J(\beta_0, \beta_1, \dots, \beta_p) &= \sum_{i=1}^n \epsilon^2 \\ &= \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_j \right)^2 \end{aligned}$$

- Ou em sua forma vetorial:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$$

$$J(\boldsymbol{\beta}) = \|\boldsymbol{\epsilon}\|_2^2 = \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

**Obs:** Qual a interpretação da minimização desta equação em forma vetorial?



## Equações normais dos mínimos quadrados (Forma Escalar)

- A função  $J(\beta_0, \beta_1, \dots, \beta_p)$  deve ser minimizada individualmente em relação a cada um dos parâmetros.
- Portanto a derivada parcial de  $J(\beta_0, \beta_1, \dots, \beta_p)$  deve ser tomada em relação a cada parâmetro  $\beta_j, j = 1, \dots, p$  e igualada a zero.

$$\frac{\partial J}{\partial \beta_0} = -2 \sum_{i=1}^N \left( y_i - \hat{\beta}_0 - \sum_{j=1}^p \hat{\beta}_j x_{ij} \right) = 0$$

$$\frac{\partial J}{\partial \beta_j} = -2 \sum_{i=1}^N x_{ij} \left( y_i - \hat{\beta}_0 - \sum_{j=1}^p \hat{\beta}_j x_{ij} \right) = 0, j = 1, 2, \dots, p$$



## Equações normais dos mínimos quadrados (Forma Escalar)

- Com a resolução das equações anteriores, pode-se montar um sistema conhecido como **equações normais de mínimos quadrados**, em sua forma escalar:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^N x_{i1} + \hat{\beta}_2 \sum_{i=1}^N x_{i2} + \cdots + \hat{\beta}_p \sum_{i=1}^N x_{ip} = \sum_{i=1}^N y_i,$$

$$N\hat{\beta}_0 \sum_{i=1}^N x_{i1} + \hat{\beta}_1 \sum_{i=1}^N x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^N x_{i1}x_{i2} + \cdots + \hat{\beta}_p \sum_{i=1}^N x_{i1}x_{ip} = \sum_{i=1}^N x_{i1}y_i,$$

$$\vdots$$
$$\vdots$$
$$\vdots$$
$$\vdots$$

$$N\hat{\beta}_0 \sum_{i=1}^N x_{ip} + \hat{\beta}_1 \sum_{i=1}^N x_{ip} + \hat{\beta}_2 \sum_{i=1}^N x_{ip}x_{i2} + \cdots + \hat{\beta}_p \sum_{i=1}^N x_{ip}^2 = \sum_{i=1}^N x_{ip}y_i,$$



## Estimação através do método dos Mínimos Quadrados Ordinário.



- *méthode des moindres carrés* - **Legendre** (1805).



- Formalização matemática completa do método - **Gauss** (1809).



Estimação através do método dos Mínimos Quadrados Ordinário.

$$J(\beta) = \|\epsilon\|_2^2 = \epsilon^T \epsilon = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$



Estimação através do método dos Mínimos Quadrados Ordinário.

$$\begin{aligned} J(\beta) &= \|\epsilon\|_2^2 = \epsilon^T \epsilon = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \\ &= \mathbf{y}^T \mathbf{y} - 2\beta^T \mathbf{X}^T \mathbf{y} + \beta^T \mathbf{X}^T \mathbf{X} \beta \end{aligned}$$



Estimação através do método dos Mínimos Quadrados Ordinário.

$$J(\beta) = \|\epsilon\|_2^2 = \epsilon^T \epsilon = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

$$= \mathbf{y}^T \mathbf{y} - 2\beta^T \mathbf{X}^T \mathbf{y} + \beta^T \mathbf{X}^T \mathbf{X} \beta$$

$$\frac{\partial J}{\partial \beta} = \mathbf{0} - 2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X} \beta = 0$$





Estimação através do método dos Mínimos Quadrados Ordinário.

$$J(\beta) = \|\epsilon\|_2^2 = \epsilon^T \epsilon = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

$$= \mathbf{y}^T \mathbf{y} - 2\beta^T \mathbf{X}^T \mathbf{y} + \beta^T \mathbf{X}^T \mathbf{X} \beta$$

$$\frac{\partial J}{\partial \beta} = \mathbf{0} - 2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X} \beta = 0$$

$$\mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y}$$



## Estimação através do método dos Mínimos Quadrados Ordinário.

$$J(\beta) = \|\epsilon\|_2^2 = \epsilon^T \epsilon = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

$$= \mathbf{y}^T \mathbf{y} - 2\beta^T \mathbf{X}^T \mathbf{y} + \beta^T \mathbf{X}^T \mathbf{X} \beta$$

$$\frac{\partial J}{\partial \beta} = \mathbf{0} - 2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X} \beta = 0$$

$$\mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y}$$

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$



## Estimação através do método dos Mínimos Quadrados Ordinário.

$$J(\beta) = \|\epsilon\|_2^2 = \epsilon^T \epsilon = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

$$= \mathbf{y}^T \mathbf{y} - 2\beta^T \mathbf{X}^T \mathbf{y} + \beta^T \mathbf{X}^T \mathbf{X} \beta$$

$$\frac{\partial J}{\partial \beta} = \mathbf{0} - 2\mathbf{X}^T \mathbf{y} + 2\mathbf{X}^T \mathbf{X} \beta = 0$$

$$\mathbf{X}^T \mathbf{X} \beta = \mathbf{X}^T \mathbf{y}$$

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

$$\hat{\beta} = \mathbf{X}^\dagger \mathbf{y}$$



Estimação através do método dos Mínimos Quadrados Ordinário.

- O modelo preditivo pode realizar novas previsões com:  $\hat{y} = \mathbf{x}\hat{\beta}$



Estimação através do método dos Mínimos Quadrados Ordinário.

- O modelo preditivo pode realizar novas previsões com:  $\hat{y} = \mathbf{x}\hat{\beta}$

Exemplo prático para regressão linear simples.

Índice	País	Cigarro per capita	Mortes por milhão de pessoas
1	Austrália	480	180
2	Canadá	500	150
3	Dinamarca	380	170
4	Finlândia	1100	350
5	Grã Bretanha	1100	460
6	Islândia	230	60
7	Holanda	490	240
8	Noruega	250	90
9	Suécia	300	110
10	Suíça	510	250

Tabela 1: Consumo per capita de cigarros em vários países em 1930 e as taxas de morte por câncer de pulmão em 1950 (Freedman et al., 2007).



## Exemplo prático para regressão linear simples.

Índice	País	Cigarro per capita	Mortes por milhão de pessoas
1	Austrália	480	180
2	Canadá	500	150
3	Dinamarca	380	170
4	Finlândia	1100	350
5	Grã Bretanha	1100	460
6	Islândia	230	60
7	Holanda	490	240
8	Noruega	250	90
9	Suécia	300	110
10	Suíça	510	250

Tabela 2: Consumo per capita de cigarros em vários países em 1930 e as taxas de morte por câncer de pulmão em 1950 (Freedman et al., 2007).



### Exemplo prático para regressão linear simples.

- Para o presente conjunto de dados, informe os valores de  $p$  e  $N$ . Além disso, existem quantos parâmetros da função regressora?



### Exemplo prático para regressão linear simples.

- Para o presente conjunto de dados, informe os valores de  $p$  e  $N$ . Além disso, existem quantos parâmetros da função regressora?
- Pode-se plotar um gráfico de dispersão dada a relação entre  $x_i$  e  $y_i$ ,  $i = 1, \dots, 10$ ? Se sim, faça.





### Exemplo prático para regressão linear simples.

- Para o presente conjunto de dados, informe os valores de  $p$  e  $N$ . Além disso, existem quantos parâmetros da função regressora?
- Pode-se plotar um gráfico de dispersão dada a relação entre  $x_i$  e  $y_i$ ,  $i = 1, \dots, 10$ ? Se sim, faça.
- Através do método dos mínimos quadrados ordinário, faça a estimação dos parâmetros da função regressora.
- Trace a reta que melhor ajusta estes dados.
- Dado uma nova amostra  $x_{11} = 400$ , qual valor de  $y_{11}$ ? Plote o ponto no gráfico de dispersão.
- **Se você estiver de posse de dados que possuam dois preditores e saídas quantitativas, é possível gerar um gráfico do modelo?**



## Regressão Linear Simples - Implementação exemplo 2

```
1 import numpy as np
2 import matplotlib.pyplot as plt
3 #Algoritmo desenvolvido por: Paulo Cirillo Souza Barbosa.
4 #Exemplo de ajuste de reta utilizando modelo estimado pelo
5 # método dos mínimos quadrados ordinário.
6 x = np.array([480 ,500 ,380 ,1100,1100,230 ,490 ,250 ,300 ,510 ])
7 x.shape = (len(x),1)
8 y = np.array([180,150,170,350,460,60,240,90,110,250])
9 y.shape = (len(y),1)
10 plt.scatter(x,y,color='orange')
11 X = np.concatenate((np.ones((len(x),1)),x),axis=1)
12
13 #Estimação do modelo:
14 B = np.linalg.pinv(X.T@X)@X.T@y
15
16 x_axis = np.linspace(0,1200,1200)
17 x_axis.shape = (len(x_axis),1)
18 ones = np.ones((len(x_axis),1))
19 X_new = np.concatenate((ones,x_axis),axis=1)
20 Y_pred = X_new@B
21 plt.plot(x_axis,Y_pred,color='blue')
22 plt.show()
```



### Estimação através do método dos Mínimos Quadrados Ordinário.

- O modelo preditivo pode realizar novas previsões com:  $\hat{y} = \mathbf{x}\hat{\beta}$

### Formalização de problemas em geral.

- Após essa descrição sólida que se utiliza de conceitos da álgebra linear, cálculo diferencial e estatística, faz sentido sair um pouco da abstração para entender as motivações para se utilizar modelos (no sentido genérico) de IA.
- O ato de desenvolver um modelo **generalista** é vantajoso no sentido de que um mesmo modelo consegue aprender relações de diferentes problemas e consegue também se adaptar a mudanças daquele problema.
- Os slides seguintes, possuem um teor prático e que está de "mãos-dadas" com os fundamentos exibidos nos slides anteriores.
- Assim, a pretensão é que você consiga realizar conexões sobre **como, quando e porquê** utilizar tais modelos generalistas.



### Passos comuns ao desenvolver algoritmos que aprendem a partir de dados.

- Foi comentado no passado que os passos comuns para desenvolver um modelo de IA generalista, tem a seguinte sequência:
  - 1 Coleta de dados.
    - Fornecidos por demanda.
    - Participação do processo de aquisição dos dados.
    - Dados sintetizados (**Orange**).



## Passos comuns ao desenvolver algoritmos que aprendem a partir de dados.

- Foi comentado no passado que os passos comuns para desenvolver um modelo de IA generalista, tem a seguinte sequência:
  - 1 Coleta de dados.
    - Fornecidos por demanda.
    - Participação do processo de aquisição dos dados.
    - Dados sintetizados (**Orange**).
  - 2 Visualização inicial dos dados.
    - Fazer afirmações e/ou tirar conclusões (**inferência**) a partir da população de dados.
    - Identificar padrões, correlações, e tendências.
    - Identificar as possíveis anomalias e *outliers*.
    - Para isso, faz-se o uso de gráficos como, por exemplo, espalhamento(*scatter*), *box-plot*, *violin-plot*, **histograma**, **matriz de coeficientes de correlação**
    - Caso haja uma alta dimensionalidade dos dados, pode-se utilizar abordagens como gráficos tridimensionais paralelos, ou métodos como *t-SNE*.



## Formalizações de problemas reais.

- Foi comentado no passado que os passos comuns para desenvolver um modelo de IA generalista, tem a seguinte sequência:
  - 3 Pré-processamento.
    - Remoção/Preenchimento de amostras com informações faltantes.
    - Remoção de possíveis anomalias e *outliers*.
    - Escalonamento/Padronização dos dados.
    - Balanceamento de classes.
    - Codificação de características (*feature encoding*).



## Formalizações de problemas reais.

- Foi comentado no passado que os passos comuns para desenvolver um modelo de IA generalista, tem a seguinte sequência:
  - 3 Pré-processamento.
    - Remoção/Preenchimento de amostras com informações faltantes.
    - Remoção de possíveis anomalias e *outliers*.
    - Escalonamento/Padronização dos dados.
    - Balanceamento de classes.
    - Codificação de características (*feature encoding*).
  - 4 Processamento.
    - Filtros, por exemplo, Passa-faixas, Médias, Gaussiano, Kalman, ou outros.
    - Redução de dimensionalidade, por exemplo, PCA, LDA, ...



## Formalizações de problemas reais.

- Foi comentado no passado que os passos comuns para desenvolver um modelo de IA generalista, tem a seguinte sequência:
  - 4 Extração e Seleção de Características.
    - Projeto de novas *features* que conseguem capturar as informações relevantes nos dados.
    - Para atingir isso, podem ser aplicados métodos de aprendizado de máquina também, mas com a tarefa de extrair e selecionar as melhores características que serão enviadas ao modelo classificador/regressor.
    - É de interesse que nessa etapa utilize-se métodos baseados no aprendizado não supervisionado e com uma quantidade pequena ou nula de hiperparâmetros.
    - Exemplos de tais métodos são: matriz de covariância, PSD, Funções de Kernel, camada Convolutacional de uma CNN.





## Formalizações de problemas reais.

- Foi comentado no passado que os passos comuns para desenvolver um modelo de IA generalista, tem a seguinte sequência:
  - 5 Modelo de classificação/regressão
    - Etapa em que o modelo treina a partir de exemplos, ou seja, seu aprendizado é condicionado as características das informações fornecidas ao modelo.
    - Nessa primeira etapa da disciplina, serão expostos os métodos estatísticos como OLS, Redes Bayesianas, Classificadores de Distância( $k$ -NN, DMC).
    - Na AV2, serão estudados os modelos conexionistas, por exemplo, Perceptron Simples, ADALINE, Perceptron de Múltiplas Camadas e Rede Função de Base Radial.
- Como dito anteriormente, a área de IA consegue abranger **muitas** subáreas, assim, além dos modelos de aprendizado de máquina (seja estatístico ou conexionista), ao final da disciplina serão estudados modelos conhecidos como Inteligência Artificial Baseada em Busca ou IA Heurística. Nessa etapa serão estudados métodos como, busca aleatória local, busca aleatória global, têmpera simulada, algoritmos genéticos, algoritmos de otimização através de enxame de partículas.



### Formalizações de problemas reais.

- Quando se deseja resolver um problema associado a tarefa de regressão/classificação, aplicam-se os passos descritos nos slides anteriores.
- É de interesse prático tentar identificar um modelo que consegue resolver o problema, essa definição é de total controle do projetista do sistema. Contudo, muitas vezes ao realizar uma análise visual dos dados, é possível levantar hipóteses de qual(is) modelo(s) conseguem resolver o problema.
- Dessa maneira, é importante medir como tais modelos desempenham para decidir qual é o ideal para solução do problema.
- Uma maneira de identificar qual modelo tem o melhor desempenho num grupo de modelos, é utilizar técnicas de validação.
  - 1 Validação incremental.
  - 2 Validação cruzada com  $k$ -dobras.
  - 3 Validação de Monte Carlo.
  - 4 Validação de Amostragem Aleatória.



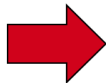
## Validação de Amostragem Aleatória

Considerando:

$$\mathbf{x} \in \mathbb{R}^{1 \times (p+1)} \quad \mathbf{y} \in \mathbb{R}^{1 \times c} \quad N = 10$$

$\mathbf{x}_1 | \mathbf{y}_1$   
 $\mathbf{x}_2 | \mathbf{y}_2$   
 $\mathbf{x}_3 | \mathbf{y}_3$   
 $\mathbf{x}_4 | \mathbf{y}_4$   
 $\mathbf{x}_5 | \mathbf{y}_5$   
 $\mathbf{x}_6 | \mathbf{y}_6$   
 $\mathbf{x}_7 | \mathbf{y}_7$   
 $\mathbf{x}_8 | \mathbf{y}_8$   
 $\mathbf{x}_9 | \mathbf{y}_9$   
 $\mathbf{x}_{10} | \mathbf{y}_{10}$

Embaralha



$\mathbf{x}_4 | \mathbf{y}_4$   
 $\mathbf{x}_7 | \mathbf{y}_7$   
 $\mathbf{x}_2 | \mathbf{y}_2$   
 $\mathbf{x}_6 | \mathbf{y}_6$   
 $\mathbf{x}_5 | \mathbf{y}_5$   
 $\mathbf{x}_3 | \mathbf{y}_3$   
 $\mathbf{x}_9 | \mathbf{y}_9$   
 $\mathbf{x}_1 | \mathbf{y}_1$   
 $\mathbf{x}_{10} | \mathbf{y}_{10}$   
 $\mathbf{x}_8 | \mathbf{y}_8$

80/20



Treino

$\mathbf{x}_4 | \mathbf{y}_4$   
 $\mathbf{x}_7 | \mathbf{y}_7$   
 $\mathbf{x}_2 | \mathbf{y}_2$   
 $\mathbf{x}_6 | \mathbf{y}_6$   
 $\mathbf{x}_5 | \mathbf{y}_5$   
 $\mathbf{x}_3 | \mathbf{y}_3$   
 $\mathbf{x}_9 | \mathbf{y}_9$   
 $\mathbf{x}_1 | \mathbf{y}_1$

Teste

$\mathbf{x}_{10} | \mathbf{y}_{10}$   
 $\mathbf{x}_8 | \mathbf{y}_8$

Estimação do modelo

$$\hat{\mathbf{W}} = (\mathbf{X}_{treino}^T \mathbf{X}_{treino})^{-1} \mathbf{X}_{treino}^T \mathbf{Y}$$

Predição

$$\mathbf{y}_{predito} = \mathbf{X}_{teste} \hat{\mathbf{W}}$$



## Pseudocódigo para validação de modelos.

### Algorithm 1: Pseudocódigo Treinamento, Teste e Validação.

- 1: Colete dados.
- 2: Organize os dados em Variáveis Regressoras ( $X$ ) e Variável Objetivo( $Y$ ).
- 3: Faça uma análise por inspeção visual dos dados.
- 4: Aplique o pré-processamento se necessário.
- 5: Defina a quantidade  $R$  de rodadas.
- 6: Crie uma lista vazia para cada modelo a ser testado representando uma medida de erro/acerto.
- 7: **for**  $r$  começando de 0 até  $R$  **do**
- 8:     Embaralhe amostras de  $X$  e  $Y$  em novas variáveis  $X_{embaralhado}$  e  $Y_{embaralhado}$ .
- 9:     Segmente as amostras de  $X$  e  $Y$  **embaralhados** em  $(X_{treinamento}, Y_{treinamento})$  e  $(X_{teste}, Y_{teste})$  utilizando uma proporção definida (90/10, 80/20, 70/30).
- 10:     Treine os modelos escolhidos utilizando apenas os dados de treinamento.
- 11:     Aplique cada dado de teste nos modelos treinados.
- 12:     Produza uma medida de erro/acerto baseado na resposta do modelo ( $Y_{predito}$ ) e  $Y_{teste}$  (para regressão, essa medida pode ser a **média de desvios quadráticos**).
- 13:     Armazene essa medida na lista criada e considerando cada modelo.
- 14: **end for**
- 15: Compute a média, desvio padrão, maior valor e menor valor das  $R$  medidas de erro/acerto existentes em cada lista preenchida. É interessante também plotar um gráfico dessas informações.



## Exemplo Contextualizado.

- Considere o seguinte exemplo hipotético:
  - 1 Você é um projetista de modelos de IA, e seu cliente te pediu para desenvolver um sistema que lhe auxilia em suas pesquisas.
  - 2 Esse cliente é um químico que faz diversos experimentos de solubilidade a partir de relações estruturais de componentes químicos.
  - 3 Em um momento inicial, o cliente resolveu apenas lhe fornecer parte dos experimentos e lhe enviou dados referentes a  $N = 951$  amostras com os preditores **peso molecular** e **quantidade de carbono** (logo,  $p = 2$ ). Em conjunto dessas informações, ele também lhe enviou as  $N = 951$  medições de solubilidade realizadas.
  - 4 Com essas informações, é possível desenvolver um sistema inteligente que **aprende** as relações entre  $(x_{\text{numeroCarbono}}, x_{\text{pesoMolecular}}, y_{\text{solubilidade}})$ ?
  - 5 Quais são as ações que você deve desempenhar?



## Regressão Linear Múltipla - Implementação exemplo 3

```
1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4 DataX = pd.read_csv('solTrainX.txt',delimiter='\t')
5 DataY = pd.read_csv('solTrainY.txt',delimiter='\t')
6 y = DataY.values
7 x1 = DataX['MolWeight'].values
8 x1.shape = (len(x1),1)
9 x2 = DataX['NumCarbon'].values
10 x2.shape = (len(x2),1)
11 X = np.concatenate((x1,x2),axis=1)
12 X = np.concatenate((np.ones((X.shape[0],1)),X),axis=1)
13 B = np.linalg.pinv(X.T@X)@X.T@y
14 x_lim = np.linspace(0,600,200)
15 y_lim = np.linspace(0,30,200)
16 xx,yy = np.meshgrid(x_lim,y_lim)
17 zz = B[0] + B[1]*xx + B[2]*yy
18 fig = plt.figure()
19 ax = fig.add_subplot(projection='3d')
20 ax.scatter(x1,x2,y,color='#DD4040')
21 ax.set_xlabel("MolWeight")
22 ax.set_ylabel("NumCarbon")
23 ax.plot_surface(xx,yy,zz,cmap='viridis',rstride=10,cstride=10)
24 plt.show()
```

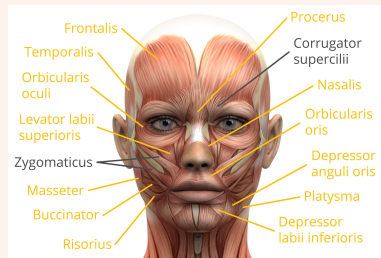


## Regressão linear.

- É possível utilizar o método dos mínimos quadrados para estimar um "sistema" que resolve um problema de **classificação**?

## Definição do problema

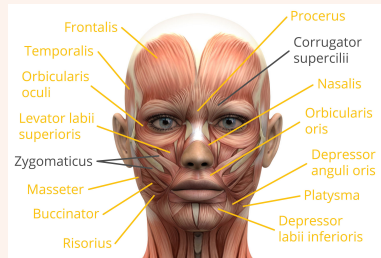
- Considere que exista um sistema que consegue prever expressões faciais forçadas, dado sinais de eletromiografia.
- Imagine que esse sistema faça a aquisição dos sinais através de dois sensores posicionados no corrugador do supercílio e no zigomático maior.





## Definição do problema

- Considere que o dispositivo que faz as aquisições do sinal com uma taxa de amostragem de 1KHz.
- Em uma determinada análise, um pesquisador se submeteu ao processo de aquisição dos sinais para cinco diferentes gestos.
- Considere que para cada gesto a coleta dos dados foi realizada durante 1 segundo.







### Conjunto de dados.

- Cada valor de biopotencial medido, pode ser entendido como uma variável que o sistema usa para decidir qual gesto é posto.
- Pode-se organizar as informações em uma matriz, referente as



### Conjunto de dados.

- Cada valor de biopotencial medido, pode ser entendido como uma variável que o sistema usa para decidir qual gesto é posto.
- Pode-se organizar as informações em uma matriz, referente as 5000 amostras com as variáveis de biopotencial medidas.
- De posse do presente conjunto de dados, usando a Álgebra Linear é possível construir esse sistema que consegue classificar os cinco diferentes gestos?



### Conjunto de dados.

- Cada valor de biopotencial medido, pode ser entendido como uma variável que o sistema usa para decidir qual gesto é posto.
- Pode-se organizar as informações em uma matriz, referente as 5000 amostras com as variáveis de biopotencial medidas.
- De posse do presente conjunto de dados, usando a Álgebra Linear é possível construir esse sistema que consegue classificar os cinco diferentes gestos?
- É necessário formular o problema de classificação com uma **transformação linear**  
 $y = xW$



## Conjunto de dados.

- Para **uma** amostra, o vetor de características pode ser representado por:

$$\mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} =$$



## Conjunto de dados.

- Para **uma** amostra, o vetor de características pode ser representado por:  
$$\mathbf{x} = [x_1 \ x_2] = [\text{Dado lido pelo sensor 1} \ \text{Dado lido pelo sensor 2}]$$
- Para o presente problema  $x_j \in \{0, 1, 2, \dots, 4095\}, j = 1, 2$



## Conjunto de dados.

- Para **uma** amostra, o vetor de características pode ser representado por:  
$$\mathbf{x} = [x_1 \ x_2] = [\text{Dado lido pelo sensor 1} \ \text{Dado lido pelo sensor 2}]$$
- Para o presente problema  $x_j \in \{0, 1, 2, \dots, 4095\}$ ,  $j = 1, 2$  que é referente a uma leitura correspondente a 12 bits do conversor A/D.



## Conjunto de dados.

- Considerando que cada amostra do conjunto de dados, é rotulada com um **texto** referente ao gesto daquela amostra. No presente conjunto existem amostras referentes aos gestos: **Neutro, Sorrindo, Aberto, Surpreso e Rabugento.**



## Conjunto de dados.

- Considerando que cada amostra do conjunto de dados, é rotulada com um **texto** referente ao gesto daquela amostra. No presente conjunto existem amostras referentes aos gestos: **Neutro**, **Sorrindo**, **Aberto**, **Surpreso** e **Rabugento**.
- Para **uma** amostra, também associa-se o vetor-código (o rótulo), que possui dimensão ( $c = 5$ ), ou seja, um identificador para o gesto posto.

$$\text{Neutro: } \mathbf{y} = [1 \quad -1 \quad -1 \quad -1 \quad -1]$$

$$\text{Sorrindo: } \mathbf{y} = [-1 \quad 1 \quad -1 \quad -1 \quad -1] \quad \text{Aberto: } \mathbf{y} = [-1 \quad -1 \quad 1 \quad -1 \quad -1]$$

$$\text{Surpreso: } \mathbf{y} = [-1 \quad -1 \quad -1 \quad 1 \quad -1] \quad \text{Rabugento: } \mathbf{y} = [-1 \quad -1 \quad -1 \quad -1 \quad 1]$$





## Conjunto de dados.

- Note que o conjunto de dados tem  $N = 5000$  vetores  $\mathbf{x}_i \in \mathbb{R}^{1 \times 2}$  e 5000 vetores  $\mathbf{y}_i \in \mathbb{R}^{1 \times 5}, i = 1, \dots, 5000$ .
- O índice  $i$  denota a  $i$ -ésima amostra presente no conjunto de dados.
- É de interesse determinar uma matriz  $\mathbf{W}$  que para um determinado dado vetor de entrada (amostra)  $\mathbf{x}_i$  forneça uma predição do vetor-código associado ao gesto correspondente:

$$\mathbf{y}_i = \mathbf{x}_i \mathbf{W}, \quad \forall k = 1, \dots, N = 5000$$

- Qual a ordem da matriz  $\mathbf{W}$ ?



## Conjunto de dados.

- Note que o conjunto de dados tem  $N = 5000$  vetores  $\mathbf{x}_i \in \mathbb{R}^{1 \times 2}$  e 5000 vetores  $\mathbf{y}_i \in \mathbb{R}^{1 \times 5}, i = 1, \dots, 5000$ .
- O índice  $i$  denota a  $i$ -ésima amostra presente no conjunto de dados.
- É de interesse determinar uma matriz  $\mathbf{W}$  que para um determinado dado vetor de entrada (amostra)  $\mathbf{x}_i$  forneça uma predição do vetor-código associado ao gesto correspondente:

$$\mathbf{y}_i = \mathbf{x}_i \mathbf{W}, \quad \forall k = 1, \dots, N = 5000$$

- Qual a ordem da matriz  $\mathbf{W}$ ? **Exatamente!!!**



## Conjunto de dados.

- Note que o conjunto de dados tem  $N = 5000$  vetores  $\mathbf{x}_i \in \mathbb{R}^{1 \times 2}$  e 5000 vetores  $\mathbf{y}_i \in \mathbb{R}^{1 \times 5}, i = 1, \dots, 5000$ .
- O índice  $i$  denota a  $i$ -ésima amostra presente no conjunto de dados.
- É de interesse determinar uma matriz  $\mathbf{W}$  que para um determinado dado vetor de entrada (amostra)  $\mathbf{x}_i$  forneça uma predição do vetor-código associado ao gesto correspondente:

$$\mathbf{y}_i = \mathbf{x}_i \mathbf{W}, \quad \forall k = 1, \dots, N = 5000$$

- Qual a ordem da matriz  $\mathbf{W}$ ? **Exatamente!!!** Sua ordem é  $2 \times 5$ .
- Esta matriz representa a versão matemática que rotula gestos faciais.
- Qual a problemática desse modelo em específico?



## Conjunto de dados.

- Note que o conjunto de dados tem  $N = 5000$  vetores  $\mathbf{x}_i \in \mathbb{R}^{1 \times 2}$  e 5000 vetores  $\mathbf{y}_i \in \mathbb{R}^{1 \times 5}, i = 1, \dots, 5000$ .
- O índice  $i$  denota a  $i$ -ésima amostra presente no conjunto de dados.
- É de interesse determinar uma matriz  $\mathbf{W}$  que para um determinado dado vetor de entrada (amostra)  $\mathbf{x}_i$  forneça uma predição do vetor-código associado ao gesto correspondente:

$$\mathbf{y}_i = \mathbf{x}_i \mathbf{W}, \quad \forall k = 1, \dots, N = 5000$$

- Qual a ordem da matriz  $\mathbf{W}$ ? **Exatamente!!!** Sua ordem é  $2 \times 5$ .
- Esta matriz representa a versão matemática que rotula gestos faciais.
- Qual a problemática desse modelo em específico? Exatamente também, o hiperplano que tentará dividir as classes está limitado à origem.



## Conjunto de dados.

- Desta maneira, inicialmente deve-se para cada amostra considerar a existência de  $w_0$ , ou seja, fazer com que  $x_0 = 1$ .
- Logo,  $\mathbf{x}_i \in \mathbb{R}^{1 \times (p+1)}$ ,  $i = 1, \dots, N$  e  $p =$



## Conjunto de dados.

- Desta maneira, inicialmente deve-se para cada amostra considerar a existência de  $w_0$ , ou seja, fazer com que  $x_0 = 1$ .
- Logo,  $\mathbf{x}_i \in \mathbb{R}^{1 \times (p+1)}$ ,  $i = 1, \dots, N$  e  $p = 2$
- Pode-se organizar todas as observações e seus rótulos nas linhas das matrizes  $\mathbf{X}$  e  $\mathbf{Y}$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_{5000} \end{bmatrix} \quad \text{e} \quad \mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_{5000} \end{bmatrix}$$

- Portanto, quais são as ordens de  $\mathbf{X}$  e  $\mathbf{Y}$ ?



## Conjunto de dados.

- Desta maneira, inicialmente deve-se para cada amostra considerar a existência de  $w_0$ , ou seja, fazer com que  $x_0 = 1$ .
- Logo,  $\mathbf{x}_i \in \mathbb{R}^{1 \times (p+1)}$ ,  $i = 1, \dots, N$  e  $p = 2$
- Pode-se organizar todas as observações e seus rótulos nas linhas das matrizes  $\mathbf{X}$  e  $\mathbf{Y}$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_{5000} \end{bmatrix} \quad \text{e} \quad \mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_{5000} \end{bmatrix}$$

- Portanto, quais são as ordens de  $\mathbf{X}$  e  $\mathbf{Y}$ ?
- $\mathbf{X} \in \mathbb{R}^{5000 \times 3}$  e  $\mathbf{Y} \in \mathbb{R}^{5000 \times 5}$ .



## Conjunto de dados.

- A versão matricial da transformação  $\mathbf{y}_k = \mathbf{x}_k \mathbf{W}$  é dada por:

$$\mathbf{Y}_{[5000 \times 5]} = \mathbf{X}_{[5000 \times 3]} \mathbf{W}_{[3 \times 5]}$$





## Conjunto de dados.

- A versão matricial da transformação  $\mathbf{y}_k = \mathbf{x}_k \mathbf{W}$  é dada por:

$$\mathbf{Y}_{[5000 \times 5]} = \mathbf{X}_{[5000 \times 3]} \mathbf{W}_{[3 \times 5]}$$

- Note que as matrizes  $\mathbf{X}$  e  $\mathbf{Y}$  são montadas a partir do conhecimento dos dados, e a matriz  $\mathbf{W}$  é



## Conjunto de dados.

- A versão matricial da transformação  $\mathbf{y}_k = \mathbf{x}_k \mathbf{W}$  é dada por:

$$\mathbf{Y}_{[5000 \times 5]} = \mathbf{X}_{[5000 \times 3]} \mathbf{W}_{[3 \times 5]}$$

- Note que as matrizes  $\mathbf{X}$  e  $\mathbf{Y}$  são montadas a partir do conhecimento dos dados, e a matriz  $\mathbf{W}$  é **DESCONHECIDA**.
- Perguntas: A matriz  $\mathbf{X}$  é quadrada?



## Conjunto de dados.

- A versão matricial da transformação  $\mathbf{y}_k = \mathbf{x}_k \mathbf{W}$  é dada por:

$$\mathbf{Y}_{[5000 \times 5]} = \mathbf{X}_{[5000 \times 3]} \mathbf{W}_{[3 \times 5]}$$

- Note que as matrizes  $\mathbf{X}$  e  $\mathbf{Y}$  são montadas a partir do conhecimento dos dados, e a matriz  $\mathbf{W}$  é **DESCONHECIDA**.
- Perguntas: A matriz  $\mathbf{X}$  é quadrada? Pode-se obter sua inversa de modo a isolar a matriz  $\mathbf{W}$ ?



## Conjunto de dados.

- A versão matricial da transformação  $\mathbf{y}_k = \mathbf{x}_k \mathbf{W}$  é dada por:

$$\mathbf{Y}_{[5000 \times 5]} = \mathbf{X}_{[5000 \times 3]} \mathbf{W}_{[3 \times 5]}$$

- Note que as matrizes  $\mathbf{X}$  e  $\mathbf{Y}$  são montadas a partir do conhecimento dos dados, e a matriz  $\mathbf{W}$  é **DESCONHECIDA**.
- Perguntas: A matriz  $\mathbf{X}$  é quadrada? Pode-se obter sua inversa de modo a isolar a matriz  $\mathbf{W}$ ?
- E se  $\mathbf{X}$  fosse uma matriz quadrada?



## Conjunto de dados.

- A versão matricial da transformação  $\mathbf{y}_k = \mathbf{x}_k \mathbf{W}$  é dada por:

$$\mathbf{Y}_{[5000 \times 5]} = \mathbf{X}_{[5000 \times 3]} \mathbf{W}_{[3 \times 5]}$$

- Note que as matrizes  $\mathbf{X}$  e  $\mathbf{Y}$  são montadas a partir do conhecimento dos dados, e a matriz  $\mathbf{W}$  é **DESCONHECIDA**.
- Perguntas: A matriz  $\mathbf{X}$  é quadrada? Pode-se obter sua inversa de modo a isolar a matriz  $\mathbf{W}$ ?
- E se  $\mathbf{X}$  fosse uma matriz quadrada? Para isolar a matriz  $\mathbf{W}$ , pode-se usar o artifício que se segue.

$$\mathbf{X}^T \mathbf{Y} = \mathbf{X}^T \mathbf{X} \mathbf{W}$$



## Conjunto de dados.

- A versão matricial da transformação  $\mathbf{y}_k = \mathbf{x}_k \mathbf{W}$  é dada por:

$$\mathbf{Y}_{[5000 \times 5]} = \mathbf{X}_{[5000 \times 3]} \mathbf{W}_{[3 \times 5]}$$

- Note que as matrizes  $\mathbf{X}$  e  $\mathbf{Y}$  são montadas a partir do conhecimento dos dados, e a matriz  $\mathbf{W}$  é **DESCONHECIDA**.
- Perguntas: A matriz  $\mathbf{X}$  é quadrada? Pode-se obter sua inversa de modo a isolar a matriz  $\mathbf{W}$ ?
- E se  $\mathbf{X}$  fosse uma matriz quadrada? Para isolar a matriz  $\mathbf{W}$ , pode-se usar o artifício que se segue.

$$\mathbf{X}^T \mathbf{Y} = \mathbf{X}^T \mathbf{X} \mathbf{W}$$

$$\mathbf{X}_{[3 \times 5000]}^T \mathbf{Y}_{[5000 \times 5]} = \mathbf{X}_{[3 \times 5000]}^T \mathbf{X}_{[5000 \times 3]} \mathbf{W}_{[3 \times 5]}$$



## Conjunto de dados.

- O que acontece quando se computa  $\mathbf{X}^T \mathbf{X}$ ??



## Conjunto de dados.

- O que acontece quando se computa  $\mathbf{X}^T \mathbf{X}$ ??
- Como se trata de uma matriz quadrada, pode-se computar a sua inversa.
- Portanto, ao multiplicar ambos os lados da equação pela inversa de  $\mathbf{X}^T \mathbf{X}$ , tem-se:

$$(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \mathbf{W}$$





## Conjunto de dados.

- O que acontece quando se computa  $\mathbf{X}^T \mathbf{X}$ ??
- Como se trata de uma matriz quadrada, pode-se computar a sua inversa.
- Portanto, ao multiplicar ambos os lados da equação pela inversa de  $\mathbf{X}^T \mathbf{X}$ , tem-se:

$$(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \mathbf{W}$$

$$\mathbf{W} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

- Como seria uma formalização para um problema genérico? Com  $N$  amostras,  $p$  preditores e  $C$  classes.



Abrindo um parêntesis com notificação importante.

- Vejam, é importante aqui destacar um detalhe já discutido em sala de aula.
- A depender da construção do conjunto de dados, as ordens das matrizes ou vetores, poderiam estar transpostas.
- Exemplo: imagine que os dados disponíveis sejam:  $\mathbf{X} \in \mathbb{R}^{p \times N}$ ,  $\mathbf{Y} \in \mathbb{R}^{C \times N}$ . O que fazer nesse caso?



Abrindo um parêntesis com notificação importante.

- Vejam, é importante aqui destacar um detalhe já discutido em sala de aula.
- A depender da construção do conjunto de dados, as ordens das matrizes ou vetores, poderiam estar transpostas.
- Exemplo: imagine que os dados disponíveis sejam:  $\mathbf{X} \in \mathbb{R}^{p \times N}$ ,  $\mathbf{Y} \in \mathbb{R}^{C \times N}$ . O que fazer nesse caso?
- Bom, a princípio pode-se pensar em transpor as matrizes de modo a manter a estrutura exibida nos slides anteriores.
- Contudo, pela álgebra linear, pode-se estimar os coeficientes de  $\mathbf{W}$  pelo mesmo princípio:

$$\mathbf{Y} = \mathbf{W}\mathbf{X}$$

- Logo, pode-se tentar isolar  $\mathbf{W}$  na tentativa de criar uma matriz quadrada com inversa.



Abrindo um parêntesis com notificação importante.

- Logo, pode-se tentar isolar  $\mathbf{W}$  na tentativa de criar uma matriz quadrada  $\mathbf{X}\mathbf{X}^T$  com inversa. Então, faz-se

$$\mathbf{Y} = \mathbf{W}\mathbf{X}$$



Abrindo um parêntesis com notificação importante.

- Logo, pode-se tentar isolar  $\mathbf{W}$  na tentativa de criar uma matriz quadrada  $\mathbf{X}\mathbf{X}^T$  com inversa. Então, faz-se

$$\mathbf{Y} = \mathbf{W}\mathbf{X}$$

$$\mathbf{Y}\mathbf{X}^T = \mathbf{W}\mathbf{X}\mathbf{X}^T$$

$$\mathbf{Y}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1} = \mathbf{W}\mathbf{X}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}$$

$$\mathbf{W} = \mathbf{Y}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}$$

- Qual diferença deste  $\mathbf{W}$  para o estimado nos slides anteriores? As informações dos coeficientes estarão compostas em ambos vetores (ou matrizes)?



## ...Continuando...Regularização por Tikhonov.

- Qual(is) problemática(s) que está(ão) relacionada(s) a matriz  $\mathbf{X}^T \mathbf{X}$  (para o caso  $\mathbf{X} \in \mathbb{R}^{N \times (p+1)}$ )



## ...Continuando...Regularização por Tikhonov.

- Qual(is) problemática(s) que está(ão) relacionada(s) a matriz  $\mathbf{X}^T \mathbf{X}$  (para o caso  $\mathbf{X} \in \mathbb{R}^{N \times (p+1)}$ )
- Para evitar o mau-condicionamento, costuma-se fazer o uso da regularização por Tikhonov.
- Nesta, a matriz  $\mathbf{W}$  passa a ser estimada por meio da seguinte expressão:

$$\mathbf{W} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{Y}$$

- Em que  $0 < \lambda \leq 1$  é chamada de constante de regularização e  $\mathbf{I} \in \mathbb{R}^{p \times p}$
- Perceba que  $\lambda$  é um **hiperparâmetro**.
- Um hiperparâmetro é um parâmetro do modelo que deve ser **pré-definido** para que os parâmetros da função discriminante propriamente dita possam ser estimados.
- Qual problemática ao se ter hiperparâmetros?



## Predição a partir do modelo gerado.

- De posse da matriz  $\mathbf{W}$ , podemos utilizá-la como componente de software de tomada de decisão para classificar as expressões faciais.
- Matematicamente, isto pode ser realizado por meio da seguinte equação:

$$\mathbf{y} = \mathbf{x}\mathbf{W}$$

- Cada uma das saídas individuais poderiam ainda ser escritas como:

$$y_i = \mathbf{x}\mathbf{w}_i$$

- Essa expressão chama-se função discriminante linear da  $i$ -ésima classe.





## Predição a partir do modelo gerado.

- Desta maneira, imagine que um sinal seja adquirido referente aos dois sensores.

$$\mathbf{x}_{novo} = [1 \quad sens1 \quad sens2]$$

- Ao multiplicarmos estes pela matriz  $\mathbf{W}$ , obtemos o vetor de saídas  $\mathbf{y}_{novo} = \mathbf{x}_{novo} \mathbf{W}$ .

$$[\hat{y}_1 \quad \hat{y}_2 \quad \hat{y}_3 \quad \hat{y}_4 \quad \hat{y}_5]$$

- Por tratar de um produto interno, utiliza-se como regra de decisão a seguinte expressão:

$$j^* = \text{índice da classe de } \mathbf{x}_{novo} = \arg \max \{\hat{y}_j\} \forall j$$

- A função *max* retorna o maior valor entre todas as saídas  $\hat{y}_j$ , e a função *arg* retorna seu índice.



## O que fazer?

- O que fazer, quando há a posse dos dados?
- Destaca-se que o conjunto de dados é disposto da seguinte maneira:
  - 1 Os dados estão estruturados em um arquivo *.json*.
  - 2 Será disponibilizado inicialmente, os dados referentes a duas classes.
  - 3 Para cada gesto, tem-se 1000 amostras do sensor 1 e 1000 do sensor 2.
  - 4 Portanto, inicialmente é necessário organizar os dados da seguinte maneira  $\mathbf{X} \in \mathbb{B}^{N \times P}$  e  $\mathbf{Y} \in \mathbb{B}^{N \times C}$ .
- As amostras então, devem ser embaralhadas.
- Em sequência, deve-se dividir as amostras  $N$  do conjunto de dados em treino/teste (comumente nas proporções 80/20 ou 70/30 ou 90/10) .
- Pode-se calcular a Taxa de acerto do classificador usando o seguinte cálculo:

$$Tx_A = \frac{\text{Qtd de predições corretas}}{\text{Total de amostras de teste}}$$



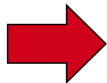
## O que fazer?

Considerando:

$$\mathbf{x} \in \mathbb{R}^{1 \times (p+1)} \quad \mathbf{y} \in \mathbb{R}^{1 \times c} \quad N = 10$$

$\mathbf{x}_1 | \mathbf{y}_1$   
 $\mathbf{x}_2 | \mathbf{y}_2$   
 $\mathbf{x}_3 | \mathbf{y}_3$   
 $\mathbf{x}_4 | \mathbf{y}_4$   
 $\mathbf{x}_5 | \mathbf{y}_5$   
 $\mathbf{x}_6 | \mathbf{y}_6$   
 $\mathbf{x}_7 | \mathbf{y}_7$   
 $\mathbf{x}_8 | \mathbf{y}_8$   
 $\mathbf{x}_9 | \mathbf{y}_9$   
 $\mathbf{x}_{10} | \mathbf{y}_{10}$

Embaralha



$\mathbf{x}_4 | \mathbf{y}_4$   
 $\mathbf{x}_7 | \mathbf{y}_7$   
 $\mathbf{x}_2 | \mathbf{y}_2$   
 $\mathbf{x}_6 | \mathbf{y}_6$   
 $\mathbf{x}_5 | \mathbf{y}_5$   
 $\mathbf{x}_3 | \mathbf{y}_3$   
 $\mathbf{x}_9 | \mathbf{y}_9$   
 $\mathbf{x}_1 | \mathbf{y}_1$   
 $\mathbf{x}_{10} | \mathbf{y}_{10}$   
 $\mathbf{x}_8 | \mathbf{y}_8$

80/20



Treino

$\mathbf{x}_4 | \mathbf{y}_4$   
 $\mathbf{x}_7 | \mathbf{y}_7$   
 $\mathbf{x}_2 | \mathbf{y}_2$   
 $\mathbf{x}_6 | \mathbf{y}_6$   
 $\mathbf{x}_5 | \mathbf{y}_5$   
 $\mathbf{x}_3 | \mathbf{y}_3$   
 $\mathbf{x}_9 | \mathbf{y}_9$   
 $\mathbf{x}_1 | \mathbf{y}_1$

Teste

$\mathbf{x}_{10} | \mathbf{y}_{10}$   
 $\mathbf{x}_8 | \mathbf{y}_8$



Estimação do modelo

$$\hat{\mathbf{W}} = (\mathbf{X}_{treino}^T \mathbf{X}_{treino})^{-1} \mathbf{X}_{treino}^T \mathbf{Y}$$



Predição

$$\mathbf{y}_{predito} = \mathbf{X}_{teste} \hat{\mathbf{W}}$$



## Modelo de classificação baseado em dissimilaridade.

- Como uma revisão de conceitos, considere que exista um conjunto de dados  $\mathbf{X} \in \mathbb{R}^{N \times p}$  que estão associados à  $N$  rótulos (podendo ser numéricos ou não).
- Ou seja, o conjunto de dados completo, é uma massa de dados  $\mathcal{X} = (\mathbf{X}_n, \mathbf{Y}_n)$  sendo  $n = 1, 2, \dots, N$ .
- Um algoritmo baseado em dissimilaridade tenta resolver o problema de classificação ao utilizar a visão de "*quanto menor for o valor, mais próximo ele é*".
- As distâncias, em uma maneira geral, são medidas de dissimilaridades.
- Um dos modelos mais elementares que utilizam a distância euclidiana como dissimilaridade, é o algoritmo baseado em vizinhanças.
- O modelo apresentado inicialmente, é chamado de  $K$ – vizinhos mais próximos, ou do inglês,  $K$ –*Nearest Neighbor* (KNN). Contudo, ao decorrer dos slides será apresentado uma versão livre de hiperparâmetro.



## Modelo de classificação baseado em dissimilaridade.

- Quando ainda estava-se falando de regressão, foi utilizado o MSE para computar uma medida de desempenho para o modelo de regressão.
- Uma maneira de computar uma medida de desempenho para classificação é utilizar a equação,

$$\frac{1}{N_t} \sum_{i=1}^{N_t} I(y_i \neq \hat{y}_i),$$

que representa a taxa de erro do modelo.

- Nesta Equação,  $I(y_i \neq \hat{y}_i)$  representa apenas uma função que retorna 1, caso o  $i$ -ésimo rótulo de  $y$  seja diferente do  $i$ -ésimo valor predito em  $\hat{y}$ . Ou seja, a presente equação computa a fração de classificações incorretas.
- Comumente essa equação é chamada de taxa de erro para treinamento/teste.



## Modelo de classificação baseado em dissimilaridade.

- Em operação, um modelo de classificação (que desempenha bem sua tarefa), **consegue discriminar uma amostra à classe mais provável, dado seus valores de preditores.**
- Em outras palavras, deve-se rotular uma amostra com  $p$  preditores para a classe  $c$  que

$$Pr(y_i = c | \mathbf{x}_i = x_0, x_1, \dots, x_p)$$

tenha o maior valor.

- Neste exemplo, a equação exibida trata-se da probabilidade condicional que  $y_i$  seja  $c$ , dada a amostra  $x_i$  com preditores  $x_0, x_1, \dots, x_p$ .
- O classificador KNN estima a probabilidade condicional de  $y_i$  dado  $\mathbf{x}_i$  e classifica a amostra para àquela classe com o maior valor de probabilidade estimada.



## Modelo de classificação baseado em dissimilaridade.

- Dado um valor positivo  $K$  e uma amostra de **teste**  $\mathbf{x}_{teste}$ , o modelo  $K$ -NN primeiro identifica os  $K$  pontos no conjunto de treinamento que estão mais próximos da amostra de  $\mathbf{x}_{teste}$ , ao computar a distância euclidiana.
- Com esses  $K$  elementos identificados, faz-se a contagem da quantidade de elementos para cada uma das classes e em seguida computa-se as probabilidades utilizando a equação,

$$Pr(y_i = c | \mathbf{x}_i = x_0, x_1, \dots, x_p) = \frac{1}{K} \sum I(y_i = c)$$

- Assim, a amostra  $\mathbf{x}_{teste}$  é rotulada com a classe com a maior valor de probabilidade.



## Pseudocódigo para validação de modelos.

**Algorithm 2:** Pseudocódigo do modelo K-NN.

- 1: Armazene em memória os dados para treinamento  $\mathbf{X}_{treino}$  e cada um de seus  $N$  rótulos  $\mathbf{y}_{treino}$ .
- 2: **for** cada amostra do conjunto de teste  $\mathbf{X}_{teste}$  **do**
- 3:    $\mathbf{x}_{teste_i}$  representa a  $i$ -ésima amostra de teste.
- 4:   **for** cada amostra do conjunto de treinamento **do**
- 5:      $\mathbf{x}_{treino_j}$  representa a  $j$ -ésima amostra de treinamento.
- 6:     Compute a distância euclidiana entre  $\mathbf{x}_{teste_i}$  e  $\mathbf{x}_{treino_j}$  ( $\|\mathbf{x}_{teste_i} - \mathbf{x}_{treino_j}\|$ ).
- 7:     Armazene esse valor computado na memória.
- 8:   **end for**
- 9:   Identifique os  $K$  elementos que possuem a menor distância calculada.
- 9:   Busque nos rótulos de treinamento ( $\mathbf{y}_{treinamento}$ ) a qual classe tais elementos com menor distância pertencem.
- 9:   Compute a probabilidade exibida no slide anterior.
- 9:   Faça a atribuição da classe de  $\mathbf{x}_{teste_i}$  para a classe com maior valor de probabilidade.
- 10: **end for**
- 11: Após todas amostras de teste serem discriminadas, compute a taxa de erro/acerto.





## Pseudocódigo para validação de modelos.

- Dúvida comum, o que se trata a etapa de treinamento deste modelo?



### Pseudocódigo para validação de modelos.

- Dúvida comum, o que se trata a etapa de treinamento deste modelo?
- Quais são as interpretações para diferentes valores de  $K$ ?
- A escolha do hiperparâmetro  $K$  possui um efeito drástico no comportamento do resultado do classificador.
- Quando  $K$  é baixo, por exemplo,  $K = 1$  o modelo é extremamente flexível e assim, o classificador pode realizar mais previsões incorretas para os dados de teste. A medida que  $K$  aumenta, o modelo se aproxima da rigidez.
- Quais problemáticas associadas ao classificador  $K$ -NN?



### Pseudocódigo para validação de modelos.

- Dúvida comum, o que se trata a etapa de treinamento deste modelo?
- Quais são as interpretações para diferentes valores de  $K$ ?
- A escolha do hiperparâmetro  $K$  possui um efeito drástico no comportamento do resultado do classificador.
- Quando  $K$  é baixo, por exemplo,  $K = 1$  o modelo é extremamente flexível e assim, o classificador pode realizar mais previsões incorretas para os dados de teste. A medida que  $K$  aumenta, o modelo se aproxima da rigidez.
- Quais problemáticas associadas ao classificador  $K$ -NN?
- Definição de hiperparâmetro ótimo.
- Custo computacional associado!
- Uma versão de classificador que utiliza dissimilaridade é o DMC (Distância Mínima ao Centróide).



## Pseudocódigo para validação de modelos.

**Algorithm 3:** Pseudocódigo do modelo K-NN.

- 1: Armazene em memória os dados para treinamento  $\mathbf{X}_{treino}$  e cada um de seus  $N$  rótulos  $\mathbf{y}_{treino}$ .
- 2: Para cada amostra, compute o ponto médio do conjunto de treinamento ( $\mathbf{m}_c$ ).
- 3: **for** cada amostra do conjunto de teste  $\mathbf{X}_{teste}$  **do**
- 4:    $\mathbf{x}_{teste_i}$  representa a  $i$ —ésima amostra de teste.
- 5:   Compute as distâncias entre  $\mathbf{x}_{teste_i}$  e cada  $\mathbf{m}_c$ .
- 6:   Identifique o elemento que possui a menor distância calculada.
- 6:   Busque nos rótulos de treinamento ( $\mathbf{y}_{treinamento}$ ) a qual classe esse elemento pertence.
- 6:   Faça a atribuição da classe de  $\mathbf{x}_{teste_i}$  para a classe identificada no passo anterior.
- 7: **end for**
- 8: Após todas amostras de teste serem discriminadas, compute a taxa de erro/acerto.