Universidade de Fortaleza - UNIFOR

Inteligência Artificial Computacional

T296

Msc. Prof. Paulo Cirillo Souza Barbosa
Centro de Ciências Tecnológicas - CCT
Fortaleza, Ceará, Brasil

Sumário

Introdução.

- O que é o aprendizado de máquina?

Introdução.

- O que é o aprendizado de máquina? Pode-se definir como um conjunto de **métodos** (muitas vezes chamados de **modelos**), capazes de detectar padrões em dados, e utilizar esse padrão descoberto para predizer dados futuros, ou tomar outros tipos de decisão sob incertezas.
- Existem paradigmas diferentes em que estes modelos são regidos, porém, três visões possuem mais destaques na área: **aprendizado supervisionado**; **aprendizado não supervisionado**; **aprendizado por reforço**.
- Os modelos em si poderão ter um caráter **discriminativo** ou **gerativo**.
- No paradigma supervisionado, o objetivo é aprender as relações entre entradas e saídas de determinado problema.
- No paradigma não supervisionado, o objetivo está em identificar padrões em dados que não se tem **rótulos ou observações** associadas.
- Em aprendizado por reforço, o objetivo do modelo é desempenhar ações com base em entradas, no entanto, o mesmo deverá ser avaliado a cada ação escolhida.

A problemática da aprendizagem.

- Aprendizado de Máquina (*Machine Learning*) e Reconhecimento de Padrões.
- Exemplo de utilização Aprendizado de máquina.
- Quando utilizar Aprendizado de Máquina?

A problemática da aprendizagem.

- Aprendizado de Máquina (*Machine Learning*) e Reconhecimento de Padrões.
- Exemplo de utilização Aprendizado de máquina.
- Quando utilizar Aprendizado de Máquina?
  1. Um padrão existe.

A problemática da aprendizagem.

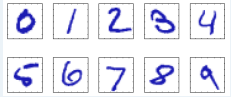
- Aprendizado de Máquina (*Machine Learning*) e Reconhecimento de Padrões.
- Exemplo de utilização Aprendizado de máquina.
- Quando utilizar Aprendizado de Máquina?
  1. Um padrão existe.
  2. Não é possível descrever matematicamente.

A problemática da aprendizagem.

- Aprendizado de Máquina (*Machine Learning*) e Reconhecimento de Padrões.
- Exemplo de utilização Aprendizado de máquina.
- Quando utilizar Aprendizado de Máquina?
  1. Um padrão existe.
  2. Não é possível descrever matematicamente.
  3. Precisa-se de dados.

Componentes principais do aprendizado

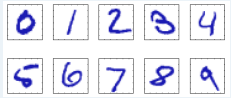
- Considere o seguinte problema (16x16):



- Entrada: **x**

Componentes principais do aprendizado

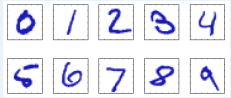
- Considere o seguinte problema (16x16):



- Entrada: **x** (vetor com dados referentes a um dígito)
- Saída: y

Componentes principais do aprendizado

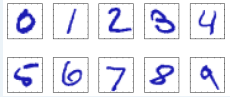
- Considere o seguinte problema (16x16):



- Entrada: **x** (vetor com dados referentes a um dígito)
- Saída: y (dígito referente: $0/1/2 \cdots 9$)
- Deseja-se uma função $f : \mathcal{X} \longrightarrow \mathcal{Y}$. Comumente chamada de função **objetivo** (*target function*)

Componentes principais do aprendizado

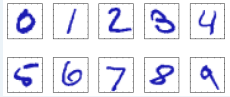
- Considere o seguinte problema (16x16):



- Entrada: **x** (vetor com dados referentes a um dígito)
- Saída: y (dígito referente: $0/1/2 \cdots 9$)
- Deseja-se uma função $f : \mathcal{X} \longrightarrow \mathcal{Y}$. Comumente chamada de função **objetivo** (*target function*) Função esta que é **desconhecida**.

Componentes principais do aprendizado

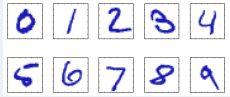
- Considere o seguinte problema (16x16):



- Entrada: **x** (vetor com dados referentes a um dígito)
- Saída: y (dígito referente: $0/1/2 \cdots 9$)
- Deseja-se uma função $f : \mathcal{X} \longrightarrow \mathcal{Y}$. Comumente chamada de função **objetivo** (*target function*) Função esta que é **desconhecida**.
- O que é preciso para aprender esse padrão?

Componentes principais do aprendizado

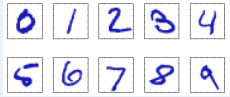
- Considere o seguinte problema (16x16):



- Entrada: **x** (vetor com dados referentes a um dígito)
- Saída: y (dígito referente: $0/1/2 \cdots 9$)
- Deseja-se uma função $f : \mathcal{X} \longrightarrow \mathcal{Y}$. Comumente chamada de função **objetivo** (*target function*) Função esta que é **desconhecida**.
- O que é preciso para aprender esse padrão? **Ora, Dados!!**

Componentes principais do aprendizado

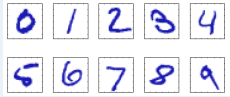
- Considere o seguinte problema (16x16):



- Entrada: **x** (vetor com dados referentes a um dígito)
- Saída: y (dígito referente: $0/1/2 \cdots 9$)
- Deseja-se uma função $f : \mathcal{X} \longrightarrow \mathcal{Y}$. Comumente chamada de função **objetivo** (*target function*) Função esta que é **desconhecida**.
- O que é preciso para aprender esse padrão? **Ora, Dados!!** $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)$

Componentes principais do aprendizado

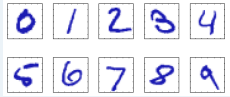
- Considere o seguinte problema (16x16):



- Entrada: **x** (vetor com dados referentes a um dígito)
- Saída: y (dígito referente: $0/1/2 \cdots 9$)
- Deseja-se uma função $f : \mathcal{X} \longrightarrow \mathcal{Y}$. Comumente chamada de função **objetivo** (*target function*) Função esta que é **desconhecida**.
- O que é preciso para aprender esse padrão? **Ora, Dados!!** $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)$
- Utiliza-se os dados para estimar a função hipótese: $g : \mathcal{X} \longrightarrow \mathcal{Y}$
- **Qual diferença entre f e g?**

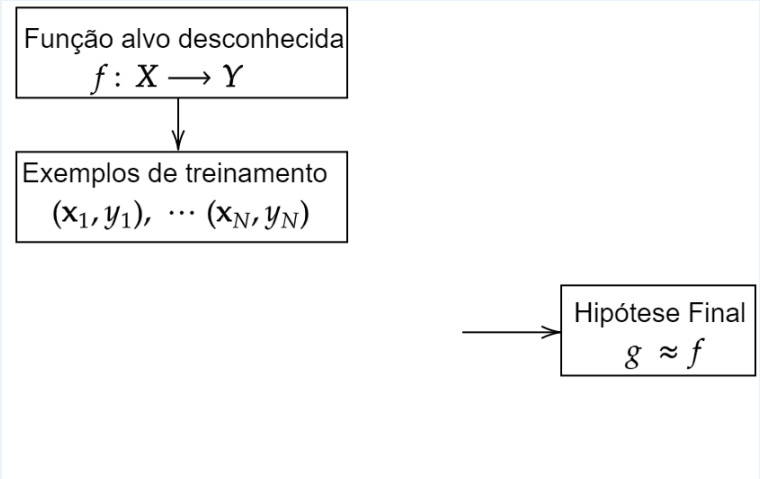
Componentes principais do aprendizado

- Considere o seguinte problema (16x16):

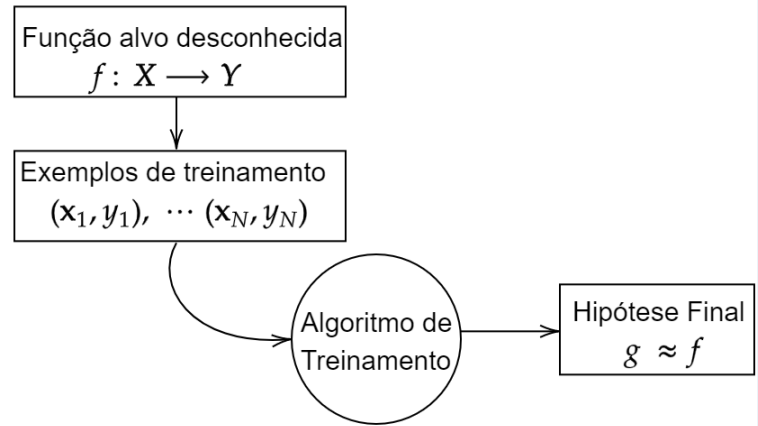


- Entrada: **x** (vetor com dados referentes a um dígito)
- Saída: y (dígito referente: $0/1/2 \cdots 9$)
- Deseja-se uma função $f : \mathcal{X} \longrightarrow \mathcal{Y}$. Comumente chamada de função **objetivo** (*target function*) Função esta que é **desconhecida**.
- O que é preciso para aprender esse padrão? **Ora, Dados!!** $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)$
- Utiliza-se os dados para estimar a função hipótese: $g : \mathcal{X} \longrightarrow \mathcal{Y}$
- **Qual diferença entre** f **e** g**?** g é uma aproximação de f.

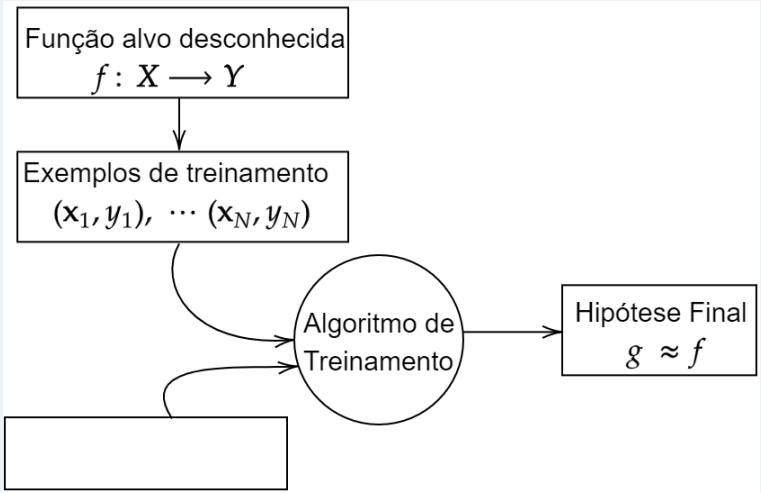
Componentes principais do aprendizado



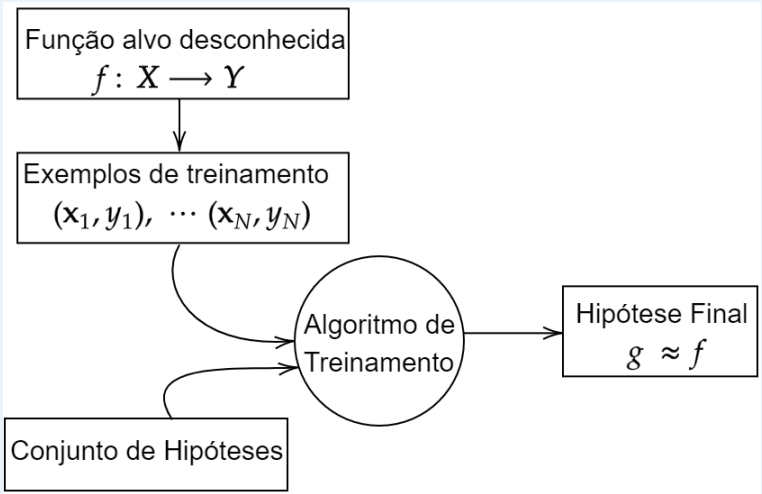
Componentes principais do aprendizado



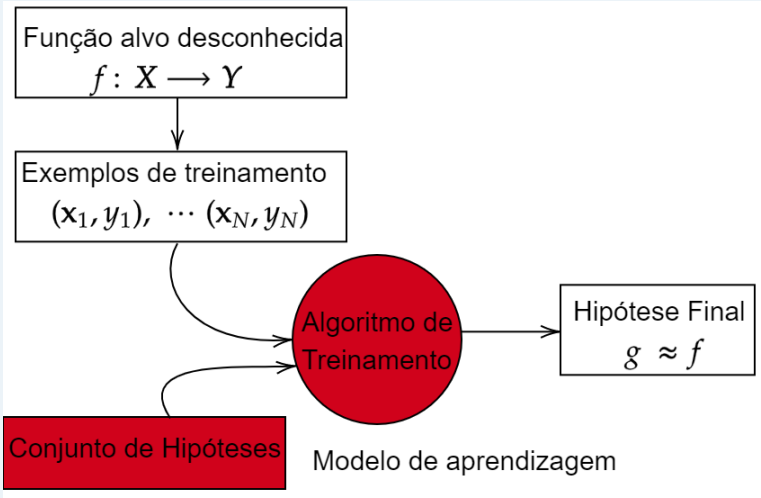
Componentes principais do aprendizado



Componentes principais do aprendizado



Componentes principais do aprendizado



Tipos de aprendizado

- Há diversos tipos, porém destacam-se **três** amplamente utilizados:
  1. Aprendizado Supervisionado (*Supervised Learning*).

Tipos de aprendizado

- Há diversos tipos, porém destacam-se **três** amplamente utilizados:
  1. Aprendizado Supervisionado (*Supervised Learning*).
  2. Aprendizado Não Supervisionado (*Unsupervised Learning*).

Tipos de aprendizado

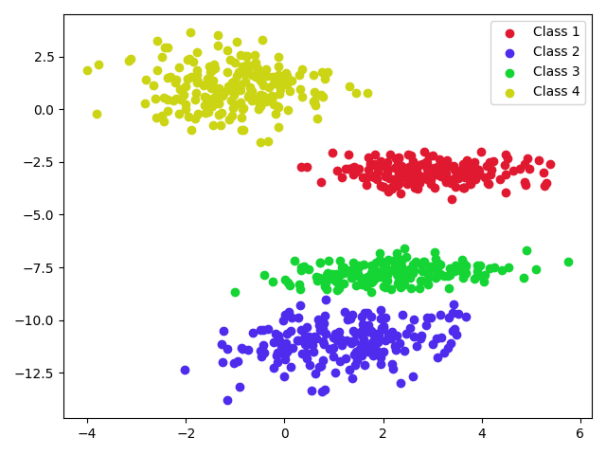
- Há diversos tipos, porém destacam-se **três** amplamente utilizados:
  1. Aprendizado Supervisionado (*Supervised Learning*).
  2. Aprendizado Não Supervisionado (*Unsupervised Learning*).
  3. Aprendizado Por Reforço (*Reinforcement Learning*).

Aprendizado Supervisionado

- (**entrada, saída correta**)

Aprendizado Supervisionado

- (**entrada, saída correta**)

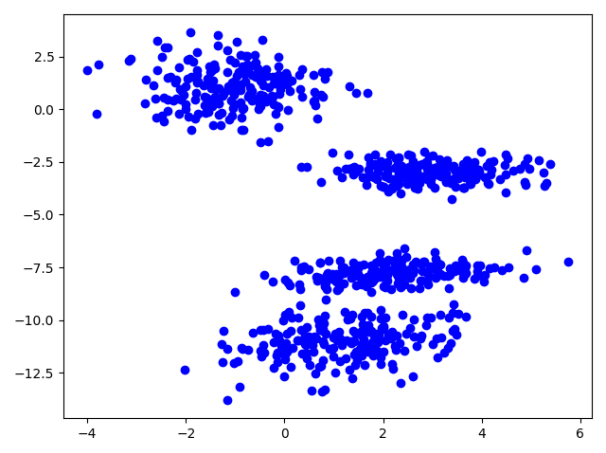


Aprendizado Não Supervisionado

- (**entrada**, **?**)

Aprendizado Não Supervisionado

- (**entrada**, **?**)



Aprendizado por Reforço

- (**entrada**, alguma **saída**, avaliar essa saída)
- Exemplos de aprendizado por reforço? (cotidiano).

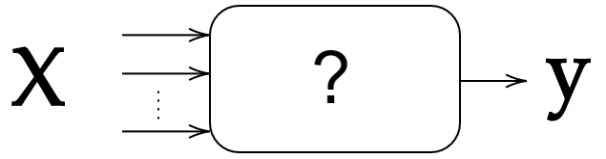
Aprendizado por Reforço

- (**entrada**, alguma **saída**, avaliar essa saída)
- Exemplos de aprendizado por reforço? (cotidiano).
- Exemplo clássico da utilização de aprendizado por reforço

Notações, Palavras-chave utilizadas.

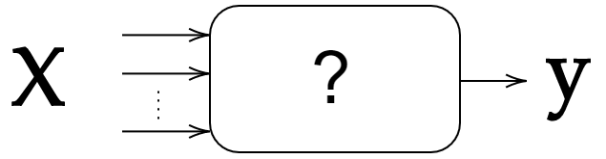
- Letras em minúsculo e não negritas (p) representam um número pertencente ao conjunto dos reais, ou seja, $p \in \mathbb{R}$.
- Letras minúsculas em negrito ($\mathbf{x}$), representam **vetores** que por natureza são colunas. Ou seja, aqueles vetores em que há uma quantidade específica de linhas e uma **única** coluna. Assim, $\mathbf{x} \in \mathbb{R}^{p \times 1}$.
- Letras maiúsculas e em negrito ($\mathbf{X}$) representa uma matriz de dimensão $m \times n$. Neste caso, com m linhas e n colunas. Ou seja, $\mathbf{X} \in \mathbb{R}^{m \times p}$.
- Funções poderão ser descritas com letras minúsculas ou maiúsculas seguidas de parentes e argumentos, por exemplo, $f(\cdot)$, $g(\cdot)$ ou $\mathbf{H}(\cdot)$.
- Subscritos i em x_i representa a $i-$ésima componente do vetor $\mathbf{x}$
- Subscritos i e j em x_{ij} representam, respectivamente, o elemento na matriz $\mathbf{X}$ alocado na $i-$ésima linha e $j-$ésima coluna.
- Subscrito i em $\mathbf{x}_i$ representa o $i-$ésimo vetor em uma matriz $\mathbf{X}$.

Regressão × Classificação



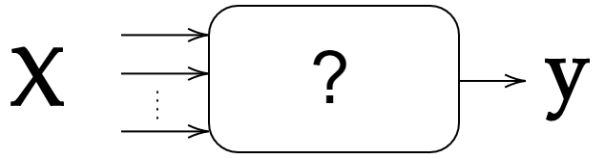
- Tem-se uma saída **y** que pode ser:
  1. Quantitativa
  2. Qualitativa (classes, categorias, fatores, variáveis discretas)

Regressão × Classificação



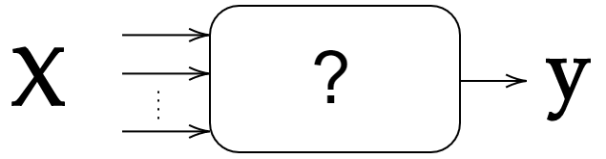
- Tem-se uma saída **y** que pode ser:
  1. Quantitativa
  2. Qualitativa (classes, categorias, fatores, variáveis discretas)
- Para ambos casos, o que é necessário para predizer a saída do sistema?

Regressão × Classificação



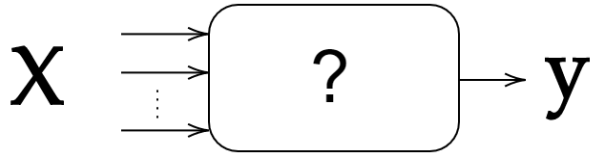
- Tem-se uma saída **y** que pode ser:
  1. Quantitativa
  2. Qualitativa (classes, categorias, fatores, variáveis discretas)
- Para ambos casos, o que é necessário para predizer a saída do sistema?
  1. **Ex1:** Dado alguma observação atmosférica do presente e passado, deseja-se predizer o **nível**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

Regressão × Classificação



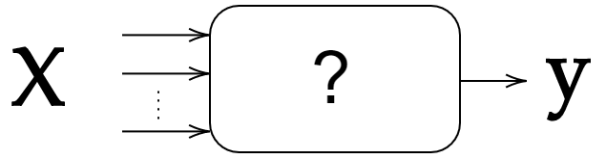
- Tem-se uma saída **y** que pode ser:
  1. Quantitativa
  2. Qualitativa (classes, categorias, fatores, variáveis discretas)
- Para ambos casos, o que é necessário para predizer a saída do sistema?
  1. **Ex1:** Dado alguma observação atmosférica do presente e passado, deseja-se predizer o **nível**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

Regressão × Classificação



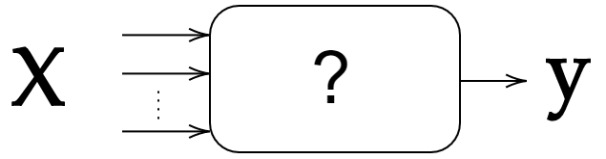
- Tem-se uma saída **y** que pode ser:
  1. Quantitativa
  2. Qualitativa (classes, categorias, fatores, variáveis discretas)
- Para ambos casos, o que é necessário para predizer a saída do sistema?
  1. **Ex1:** Dado alguma observação atmosférica do presente e passado, deseja-se predizer o **nível**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

Regressão × Classificação



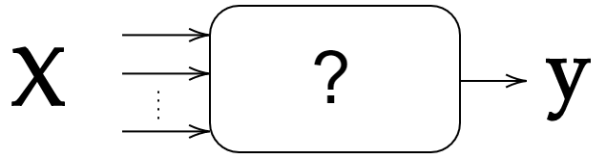
- Tratando de regressões (inicialmente pelo menos):
  1. Necessita-se dos dados de treinamento.

Regressão × Classificação



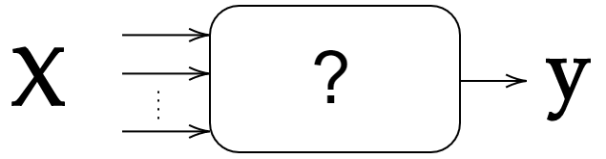
- Tratando de regressões (inicialmente pelo menos):
  1. Necessita-se dos dados de treinamento.
  2. Observam-se as saídas para um conjunto de amostras.

Regressão × Classificação



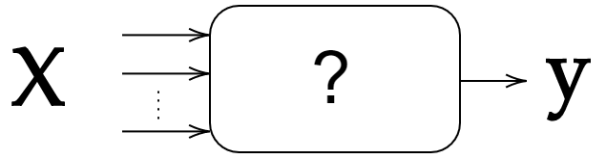
- Tratando de regressões (inicialmente pelo menos):
  1. Necessita-se dos dados de treinamento.
  2. Observam-se as saídas para um conjunto de amostras.
  3. Utiliza-se estes dados para construir o modelo preditivo.

Regressão × Classificação



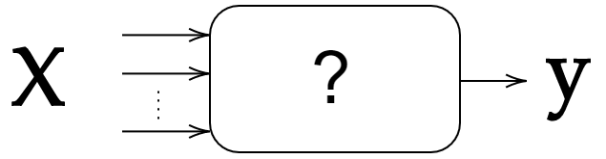
- Tratando de regressões (inicialmente pelo menos):
  1. Necessita-se dos dados de treinamento.
  2. Observam-se as saídas para um conjunto de amostras.
  3. Utiliza-se estes dados para construir o modelo preditivo.
  4. Este no que lhe concerne, pode realizar predições para amostras **desconhecidas**.

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 predições 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 predições 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 predições 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 predizer 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 predizer 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 $x_1, x_2, \cdots x_p \in \mathbb{R}$.

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 predizer 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 $x_1, x_2, \cdots x_p \in \mathbb{R}$.
  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

$$y = f(x_1, x_2, \cdots x_p | \beta_1, \cdots, \beta_p) + \varepsilon = f(\mathbf{x}|\boldsymbol{\beta}) + \varepsilon$$

Regressão Linear Simples/ Múltipla

$$y = f(x_1, x_2, \cdots x_p | \beta_1, \cdots, \beta_p) + \varepsilon = f(\mathbf{x}|\boldsymbol{\beta}) + \varepsilon$$

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

Regressão Linear Simples/ Múltipla

$$y = f(x_1, x_2, \cdots x_p | \beta_1, \cdots, \beta_p) + \varepsilon = f(\mathbf{x}|\boldsymbol{\beta}) + \varepsilon$$

- $f(\cdot|\cdot)$ é denominada a função de regressão.
- $\mathbf{x} \in \mathbb{R}^p$ é o vetor de variáveis regressoras.
- $\boldsymbol{\beta} \in \mathbb{R}^p$ é o vetor de parâmetros da função regressora.
- ε 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 $\{x_1, x_2, \cdots x_p\}$:

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

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

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

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

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*.
- 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

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

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

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

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

- Nesta, o β_0 é o intercepto (*intercept*), e β_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

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

Regressão Linear Simples/ Múltipla

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

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

- 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

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

- 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

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

- 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 $(\mathbf{x}_i, y_i), \quad i = 1, \cdots N$.
  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

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

- 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 $(\mathbf{x}_i, y_i), \quad i = 1, \cdots N$.
  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 predições é:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_n x_n = \hat{\beta}_0 + \sum_{i=0}^{p} \hat{\beta}_i x_i = \hat{\boldsymbol{\beta}}^T \mathbf{x}$$

Regressão Linear Simples/ Múltipla

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

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_n x_n = \hat{\beta}_0 + \sum_{i=0}^{p} \hat{\beta}_i x_i = \hat{\boldsymbol{\beta}}^T \mathbf{x}$$

- Qual diferença em termos de interpretação geométrica com relação ao modelo simples?
- Quais vantagens em se utilizar este modelo linear?
  1. Simplicidade do modelo.
  2. Interpretação direta ao parâmetro β_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 $\hat{\boldsymbol{\beta}}$, 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 $\hat{\boldsymbol{\beta}}$), precisa-se:
  1. N observações (amostras) do par entrada-saída $(\mathbf{x}_i, y_i), i = 1, \cdots N$.

Estimação através do método dos Mínimos Quadrados Ordinário.

- A estimação do modelo, consequentemente dos coeficientes de $\hat{\boldsymbol{\beta}}$, 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 $\hat{\boldsymbol{\beta}}$), precisa-se:
  1. N observações (amostras) do par entrada-saída $(\mathbf{x}_i, y_i), i = 1, \cdots N$.
  2. Em outras palavras, a i-ésima observação tem como resposta um escalar y_i a partir do vetor de variáveis correspondentes $\mathbf{x}_i = \begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{bmatrix}^T$, em que x_{ji} representa a $j-$ésima variável regressora da $i-$ésima amostra, em que $i = \{1, \cdots N\}$ e $j = \{1, \cdots, p\}$

Estimação através do método dos Mínimos Quadrados Ordinário.

- A estimação do modelo, consequentemente dos coeficientes de $\hat{\boldsymbol{\beta}}$, 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 $\hat{\boldsymbol{\beta}}$), precisa-se:
  1. N observações (amostras) do par entrada-saída $(\mathbf{x}_i, y_i), i = 1, \cdots N$.
  2. Em outras palavras, a i-ésima observação tem como resposta um escalar y_i a partir do vetor de variáveis correspondentes $\mathbf{x}_i = \begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{bmatrix}^T$, em que x_{ji} representa a $j-$ésima variável regressora da $i-$ésima amostra, em que $i = \{1, \cdots N\}$ e $j = \{1, \cdots, p\}$
  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 $\hat{\boldsymbol{\beta}}$, 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 $\hat{\boldsymbol{\beta}}$), precisa-se:
  1. N observações (amostras) do par entrada-saída $(\mathbf{x}_i, y_i), i = 1, \cdots N$.
  2. Em outras palavras, a i-ésima observação tem como resposta um escalar y_i a partir do vetor de variáveis correspondentes $\mathbf{x}_i = \begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{bmatrix}^T$, em que x_{ji} representa a $j-$ésima variável regressora da $i-$ésima amostra, em que $i = \{1, \cdots N\}$ e $j = \{1, \cdots, p\}$
  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 σ^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 $(\mathbf{x}_i, y_i), i = 1, \cdots N$.
  2. Em outras palavras, a i-ésima observação tem como resposta um escalar y_i a partir do vetor de variáveis correspondentes $\mathbf{x}_i = \begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{bmatrix}^T$, em que x_{ji} representa a $j-$ésima variável regressora da $i-$ésima amostra, em que $i = \{1, \cdots N\}$ e $j = \{1, \cdots, p\}$
  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 σ^2
  5. Assume-se que as observações em ε 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 $(\mathbf{x}_i, y_i), i = 1, \cdots N$.
  2. Em outras palavras, a i-ésima observação tem como resposta um escalar y_i a partir do vetor de variáveis correspondentes $\mathbf{x}_i = \begin{bmatrix} x_{i1} & x_{i2} & \cdots & x_{ip} \end{bmatrix}^T$, em que x_{ji} representa a $j-$ésima variável regressora da $i-$ésima amostra, em que $i = \{1, \cdots N\}$ e $j = \{1, \cdots, p\}$
  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 σ^2
  5. Assume-se que as observações em ε 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:

$$\begin{aligned}
\beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \varepsilon_1 &= y_1 \\
\beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \varepsilon_2 &= y_2 \\
&\vdots \\
\beta_0 + \beta_1 x_{N1} + \beta_2 x_{N2} + \cdots + \beta_p x_{Np} + \varepsilon_N &= y_N
\end{aligned}$$

- 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:

$$\begin{gathered}\beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \varepsilon_1 = y_1 \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \varepsilon_2 = y_2 \\ \vdots \\ \beta_0 + \beta_1 x_{N1} + \beta_2 x_{N2} + \cdots + \beta_p x_{Np} + \varepsilon_N = y_N\end{gathered}$$

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

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

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

Estimação através do método dos Mínimos Quadrados Ordinário.

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

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

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}_{N\times 1}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x12 & \cdots & x_{1p} \\ 1 & x_{21} & x22 & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & xN2 & \cdots & x_{Np} \end{bmatrix}_{N\times(p+1)}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}_{(p+1)\times 1}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix}_{N\times 1}$$

Estimação através do método dos Mínimos Quadrados Ordinário.

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

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

Estimação através do método dos Mínimos Quadrados Ordinário.

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

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

- 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 $J(\cdot)$:

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

Estimação através do método dos Mínimos Quadrados Ordinário.

$$J(\beta_0, \beta_1, \cdots, \beta_p) = \sum_{i=1}^{n} \varepsilon^2$$

$$= \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j \right)^2$$

- Ou em sua forma vetorial:

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

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

Estimação através do método dos Mínimos Quadrados Ordinário.

$$J(\beta_0, \beta_1, \cdots, \beta_p) = \sum_{i=1}^{n} \varepsilon^2$$

$$= \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j \right)^2$$

- Ou em sua forma vetorial:

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

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

Estimação através do método dos Mínimos Quadrados Ordinário.

$$J(\beta_0, \beta_1, \cdots, \beta_p) = \sum_{i=1}^{n} \varepsilon^2$$

$$= \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_j \right)^2$$

- Ou em sua forma vetorial:

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

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

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

Equações normais dos mínimos quadrados (Forma Escalar)

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

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

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

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:

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

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

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

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

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.

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

Estimação através do método dos Mínimos Quadrados Ordinário.

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

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

Estimação através do método dos Mínimos Quadrados Ordinário.

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

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

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

Estimação através do método dos Mínimos Quadrados Ordinário.

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

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

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

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

Estimação através do método dos Mínimos Quadrados Ordinário.

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

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

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

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

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

Estimação através do método dos Mínimos Quadrados Ordinário.

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

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

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

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

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

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

Estimação através do método dos Mínimos Quadrados Ordinário.

- O modelo preditivo pode realizar novas predições com: $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$

Estimação através do método dos Mínimos Quadrados Ordinário.

- O modelo preditivo pode realizar novas predições com: $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$

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, $i = 1, \cdots 10$? 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, $i = 1, \cdots 10$? 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

```python
import numpy as np
import matplotlib.pyplot as plt
#Algoritmo desenvolvido por: Paulo Cirillo Souza Barbosa.
#Exemplo de ajuste de reta utilizando modelo estimado pelo
# método dos mínimos quadrados ordinário.
x = np.array([480 ,500 ,380 ,1100,1100,230 ,490 ,250 ,300 ,510 ])
x.shape = (len(x),1)
y = np.array([180,150,170,350,460,60,240,90,110,250])
y.shape = (len(y),1)
plt.scatter(x,y,color='orange')
X = np.concatenate((np.ones((len(x),1)),x),axis=1)

#Estimação do modelo:
B = np.linalg.pinv(X.T@X)@X.T@y

x_axis = np.linspace(0,1200,1200)
x_axis.shape = (len(x_axis),1)
ones = np.ones((len(x_axis),1))
X_new = np.concatenate((ones,x_axis),axis=1)
Y_pred = X_new@B
plt.plot(x_axis,Y_pred,color='blue')
plt.show()
```

Estimação através do método dos Mínimos Quadrados Ordinário.

- O modelo preditivo pode realizar novas predições com: $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$

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. Análise Exploratória 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 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 Convolucional de uma CNN.

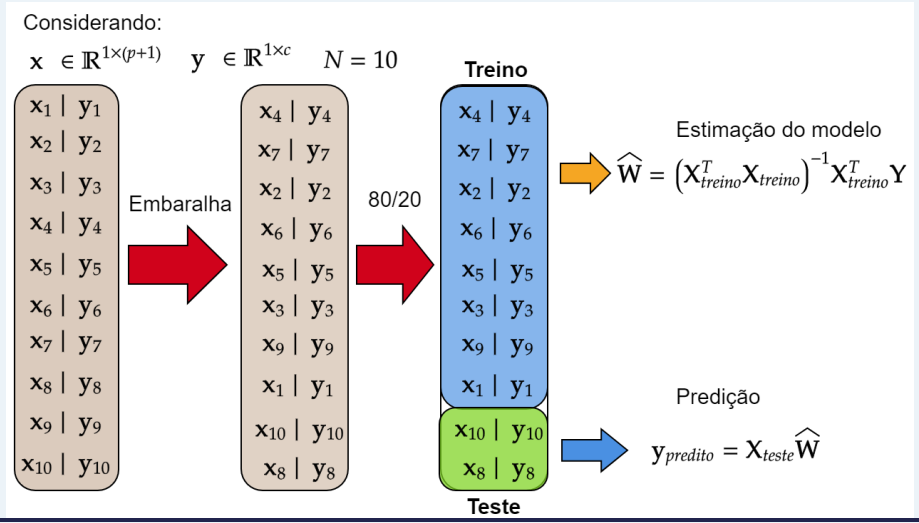
Formalizações de problemas reais.

- 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 Etapa da disciplina, serão expostos os métodos estatísticos como OLS e OLS regularizado.
     - Na AV3, serão estudados os modelos conexionistas, por exemplo, Perceptron Simples, ADALINE, Perceptron de Múltiplas Camadas e Rede Função de Base Radial.

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 *Leave-One-Out*.
  2. Validação cruzada com $k-$dobras.
  3. Validação por simulações de Monte Carlo.
  4. Validação de Amostragem Aleatória.

Validação de Amostragem Aleatória



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 $\mathbf{X}_{embaralhado}$ e $\mathbf{Y}_{embaralhado}$.
9: Segmente as amostras de **X** e **Y** **embaralhados** em $(\mathbf{X}_{treinamento}, \mathbf{Y}_{treinamento})$ e $(\mathbf{X}_{teste}, \mathbf{Y}_{teste})$ 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 ($\mathbf{Y}_{predito}$) e $\mathbf{Y}_{teste}$ (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 $(x_{numeroCarbono}, x_{pesoMolecular}, y_{solubilidade})$?
  5. Quais são as ações que você deve desempenhar?

Regressão Linear Múltipla - Implementação exemplo 3

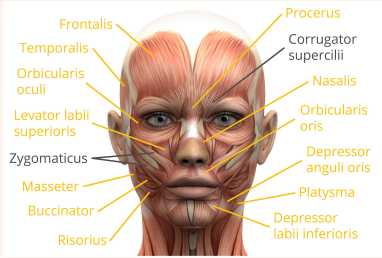
```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
DataX = pd.read_csv('solTrainX.txt',delimiter='\t')
DataY = pd.read_csv('solTrainY.txt',delimiter='\t')
y = DataY.values
x1 = DataX['MolWeight'].values
x1.shape = (len(x1),1)
x2 = DataX['NumCarbon'].values
x2.shape = (len(x2),1)
X = np.concatenate((x1,x2),axis=1)
X = np.concatenate((np.ones((X.shape[0],1)),X),axis=1)
B = np.linalg.pinv(X.T@X)@X.T@y
x_lim = np.linspace(0,600,200)
y_lim= np.linspace(0,30,200)
xx,yy = np.meshgrid(x_lim,y_lim)
zz = B[0] + B[1]*xx + B[2]*yy
fig = plt.figure()
ax = fig.add_subplot(projection='3d')
ax.scatter(x1,x2,y,color='#DD4040')
ax.set_xlabel("MolWeight")
ax.set_ylabel("NumCarbon")
ax.plot_surface(xx,yy,zz,cmap='viridis',rstride=10,cstride=10)
plt.show()
```

Regressão linear.

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

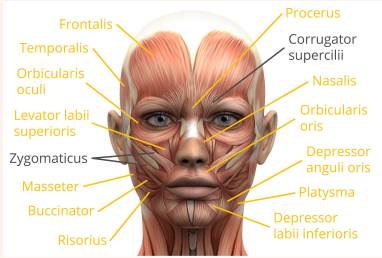
Definição do problema

- Considere que exista um sistema que consegue predizer 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 1**KHz**.
- 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** $\mathbf{y} = \mathbf{xW}$

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:

$$\mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} = \begin{bmatrix} \text{Dado lido pelo sensor 1} & \text{Dado lido pelo sensor 2} \end{bmatrix}$$

- Para o presente problema $x_j \in \{0, 1, 2, \cdots 4095\}$, $j = 1, 2$

Conjunto de dados.

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

$$\mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} = \begin{bmatrix} \text{Dado lido pelo sensor 1} & \text{Dado lido pelo sensor 2} \end{bmatrix}$$

- Para o presente problema $x_j \in \{0, 1, 2, \cdots 4095\}$, $j = 1, 2$ 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.

$$\text{Neutro: } \mathbf{y} = \begin{bmatrix} 1 & -1 & -1 & -1 & -1 \end{bmatrix}$$

$$\text{Sorrindo: } \mathbf{y} = \begin{bmatrix} -1 & 1 & -1 & -1 & -1 \end{bmatrix} \quad \text{Aberto: } \mathbf{y} = \begin{bmatrix} -1 & -1 & 1 & -1 & -1 \end{bmatrix}$$

$$\text{Surpreso: } \mathbf{y} = \begin{bmatrix} -1 & -1 & -1 & 1 & -1 \end{bmatrix} \quad \text{Rabugento: } \mathbf{y} = \begin{bmatrix} -1 & -1 & -1 & -1 & 1 \end{bmatrix}$$

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 $\mathbf{y}_i \in \mathbb{R}^{1\times 5}, i = 1, \cdots, 5000$.
- 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:

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

- 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 $\mathbf{y}_i \in \mathbb{R}^{1\times 5}, i = 1, \cdots, 5000$.
- 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:

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

- 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 $\mathbf{y}_i \in \mathbb{R}^{1\times 5}, i = 1, \cdots, 5000$.
- 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:

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

- Qual a ordem da matriz $\mathbf{W}$? **Exatamente!!!** Sua ordem é 2×5.
- 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 $\mathbf{y}_i \in \mathbb{R}^{1\times 5}, i = 1, \cdots, 5000$.
- 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:

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

- Qual a ordem da matriz $\mathbf{W}$? **Exatamente!!!** Sua ordem é 2×5.
- 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, $\mathbf{x}_i \in \mathbb{R}^{1\times(p+1)}$, $i = 1, \cdots, N$ 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, $\mathbf{x}_i \in \mathbb{R}^{1\times(p+1)}$, $i = 1, \cdots, N$ 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, $\mathbf{x}_i \in \mathbb{R}^{1\times(p+1)}$, $i = 1, \cdots, N$ 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 $\mathbf{Y}$?
- $\mathbf{X} \in \mathbb{R}^{5000\times3}$ e $\mathbf{Y} \in \mathbb{R}^{5000\times5}$.

Conjunto de dados.

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

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

Conjunto de dados.

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

$$\mathbf{Y}_{[5000\times5]} = \mathbf{X}_{[5000\times3]}\mathbf{W}_{[3\times5]}$$

- 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:

$$\mathbf{Y}_{[5000\times5]} = \mathbf{X}_{[5000\times3]}\mathbf{W}_{[3\times5]}$$

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

Conjunto de dados.

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

$$\mathbf{Y}_{[5000\times5]} = \mathbf{X}_{[5000\times3]}\mathbf{W}_{[3\times5]}$$

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

Conjunto de dados.

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

$$\mathbf{Y}_{[5000\times5]} = \mathbf{X}_{[5000\times3]}\mathbf{W}_{[3\times5]}$$

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

Conjunto de dados.

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

$$\mathbf{Y}_{[5000\times5]} = \mathbf{X}_{[5000\times3]}\mathbf{W}_{[3\times5]}$$

- 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:

$$\mathbf{Y}_{[5000\times5]} = \mathbf{X}_{[5000\times3]}\mathbf{W}_{[3\times5]}$$

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

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

$$\mathbf{X}^T_{[3\times5000]}\mathbf{Y}_{[5000\times5]} = \mathbf{X}^T_{[3\times5000]}\mathbf{X}_{[5000\times3]}\mathbf{W}_{[3\times5]}$$

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 λ 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.

$$\mathbf{x}_{novo} = \begin{bmatrix} 1 & sens1 & sens2 \end{bmatrix}$$

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

$$\begin{bmatrix} \hat{y}_1 & \hat{y}_2 & \hat{y}_3 & \hat{y}_4 & \hat{y}_5 \end{bmatrix}$$

- 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.

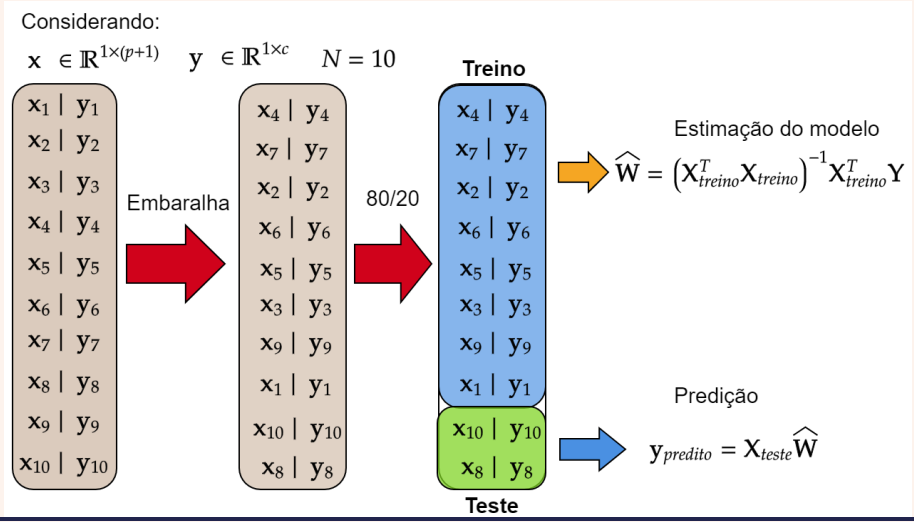
O que fazer?

- 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:

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

- Quão confiável é essa taxa de acerto?
- O que pode ser realizado para melhorar esta confiabilidade?

O que fazer?


Considerando:
$\mathbf{x} \in \mathbb{R}^{1\times(p+1)}$ $\mathbf{y} \in \mathbb{R}^{1\times c}$ $N = 10$
Embaralha
80/20
Treino
Teste
Estimação do modelo
$\widehat{\mathbf{W}} = \left(\mathbf{X}_{treino}^T \mathbf{X}_{treino}\right)^{-1} \mathbf{X}_{treino}^T \mathbf{Y}$
Predição
$\mathbf{y}_{predito} = \mathbf{X}_{teste} \widehat{\mathbf{W}}$

Dilema viés-variância.

- Já foi discutido que uma das maneiras de avaliar o desempenho de modelos de regressão, através da minimização da medida de desvios quadráticos, dada pela equação:

Dilema viés-variância.

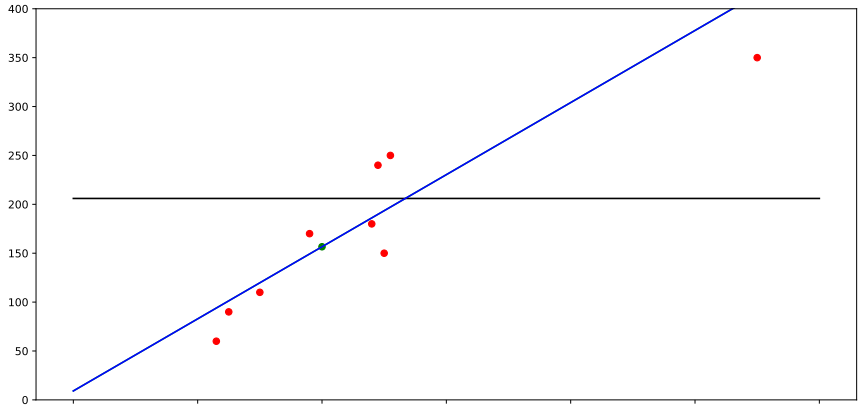
- Já foi discutido que uma das maneiras de avaliar o desempenho de modelos de regressão, através da minimização da medida de desvios quadráticos, dada pela equação:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(x_i))^2$$

- Nesta, $\hat{f}$ representa o modelo estimado, e x_i é a $i-$ésima amostra.
- Neste caso, MSE pode ser calculado tanto na etapa de treinamento quando na etapa de teste.
- Essa abordagem abre discussões importantes sobre a estimação de modelos em geral.
- No caso da estimação do MMQO, uma minimização de MSE de treino é interessante? E para métodos não-lineares como essa análise é realizada?

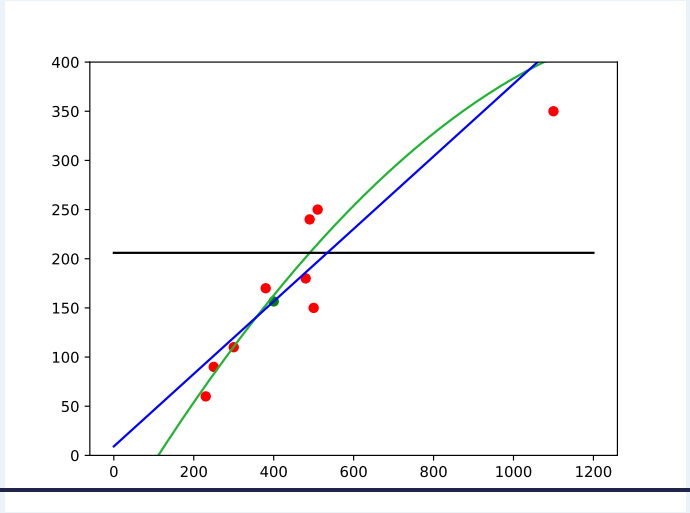
Dilema viés-variância.

- O que dizer sobre a imagem?



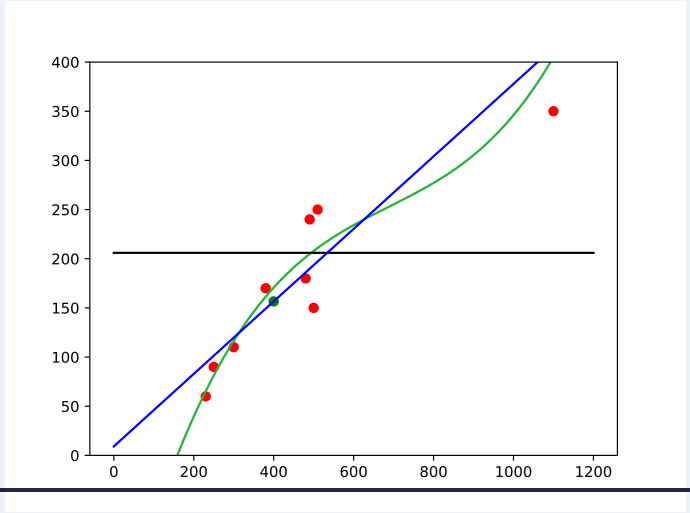
Dilema viés-variância.

- O que dizer sobre a imagem?



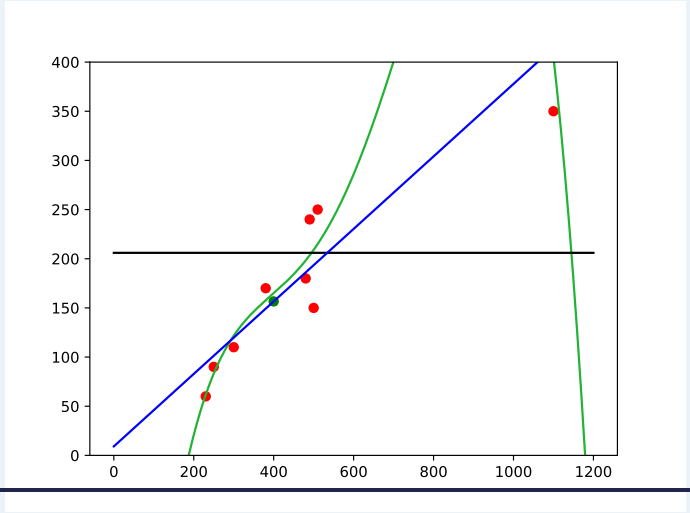
Dilema viés-variância.

- O que dizer sobre a imagem?



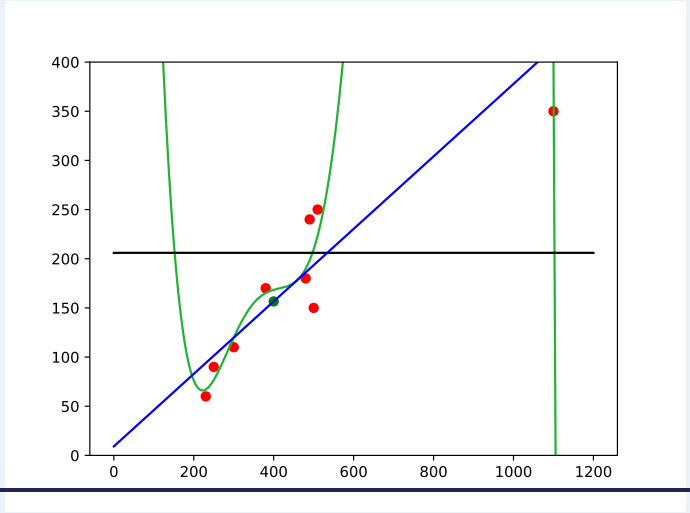
Dilema viés-variância.

- O que dizer sobre a imagem?



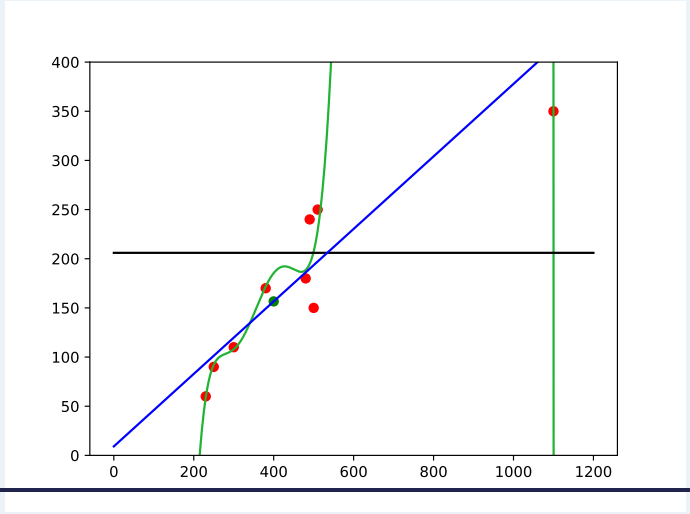
Dilema viés-variância.

- O que dizer sobre a imagem?



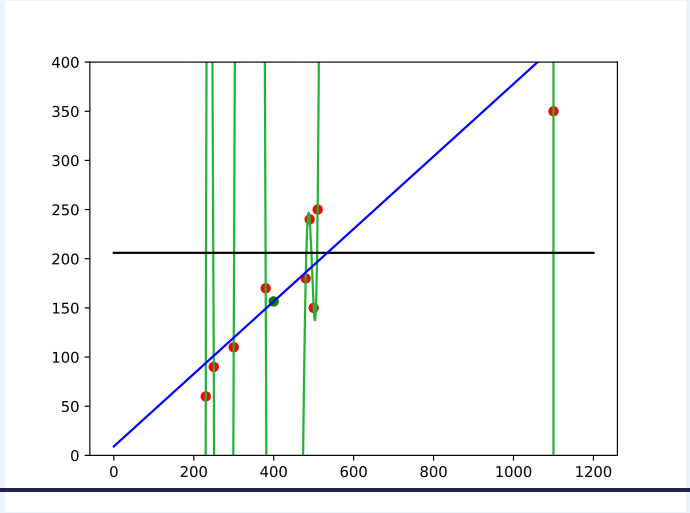
Dilema viés-variância.

- O que dizer sobre a imagem?

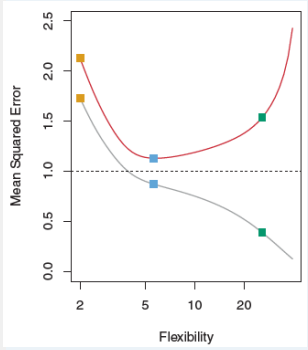


Dilema viés-variância.

- O que dizer sobre a imagem?

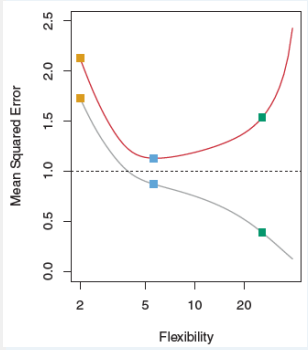


Dilema viés-variância.



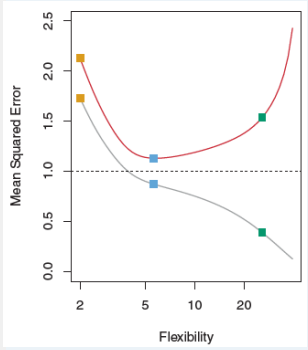
- Essa característica de declínio monotônico no MSE de treino e forma de U no MSE do teste, é presente em diversos métodos estatísticos, independente dos dados.
- A medida que a flexibilidade do modelo é aumentada, MSE de treinamento diminui, contudo, MSE do teste pode não desempenhar da mesma maneira.
- Quando um modelo proporciona um baixo MSE de treino e um alto MSE de teste, é dito que o modelo **sobreajusta** os dados (*overfitting*).
- Deve-se levar em consideração que ocorrendo ou não um *overfitting*, é esperado que o MSE de treinamento seja inferior ao MSE de teste.
- Alguém arrisca dizer o motivo?

Dilema viés-variância.



- Essa característica de declínio monotônico no MSE de treino e forma de U no MSE do teste, é presente em diversos métodos estatísticos, independente dos dados.
- A medida que a flexibilidade do modelo é aumentada, MSE de treinamento diminui, contudo, MSE do teste pode não desempenhar da mesma maneira.
- Quando um modelo proporciona um baixo MSE de treino e um alto MSE de teste, é dito que o modelo **sobreajusta** os dados (*overfitting*).
- Deve-se levar em consideração que ocorrendo ou não um *overfitting*, é esperado que o MSE de treinamento seja inferior ao MSE de teste.
- Alguém arrisca dizer o motivo? Exatamente!

Dilema viés-variância.



- Essa característica de declínio monotônico no MSE de treino e forma de U no MSE do teste, é presente em diversos métodos estatísticos, independente dos dados.
- A medida que a flexibilidade do modelo é aumentada, MSE de treinamento diminui, contudo, MSE do teste pode não desempenhar da mesma maneira.
- Quando um modelo proporciona um baixo MSE de treino e um alto MSE de teste, é dito que o modelo **sobreajusta** os dados (*overfitting*).
- Deve-se levar em consideração que ocorrendo ou não um *overfitting*, é esperado que o MSE de treinamento seja inferior ao MSE de teste.
- Alguém arrisca dizer o motivo? Exatamente! A maioria dos modelos busca a estimação de seus parâmetros através da minimização do MSE de treinamento.

Dilema viés-variância.

- A forma em *U* exibida anteriormente, é resultado de duas propriedades importantes em modelos de machine learning.
- Para minimizar o MSE de teste, é necessário encontrar um modelo que simultaneamente possua uma baixa variância e um baixo viés (*bias*).
- Nesse contexto, a **variância** do modelo é a capacidade que ele tem de modificar sua estimação quando utiliza-se diferentes conjuntos de dados. Pelas discussões realizadas, é interessante que essa estimação não cause uma diferença tão grande entre conjuntos de treinamento.
- Nesse contexto, *bias* trata-se da incapacidade do modelo de capturar as verdadeiras relações entre entrada e saída.
- Como exemplo, pode-se dizer que uma regressão linear, possui uma alta quantidade relativa de viés.
- Uma aproximação polinomial de alta ordem, consegue ser muito flexível e assim, possui baixo viés.

Classificador Bayesiano Gaussiano.

- A discussão realizada na última aula, sobre avaliação de MSE para treino/teste e dilema variância, foi realizada apenas para a configuração de regressão.
- Contudo, esses conceitos podem ser estendidos para a classificação com algumas modificações.
- Estas precisam ser realizadas, pois, as predições do modelo não são quantitativas.
- Uma maneira interessante para avaliar o desempenho do modelo em sua etapa de treinamento é pela taxa de erro (ou sua versão em função do acerto).

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

- 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 $\hat{y}$. Tal equação computa a fração de classificações incorretas.

Classificador Bayesiano Gaussiano.

- Considerando elementos no processo de classificação:
- Existem N pares $\{\mathbf{x}_i, y_i\}$
- $\mathbf{x}_i \in \mathbb{R}^p$ é o $i-$ésimo padrão de entrada.
- y_i é o rótulo da classe à qual pertence $\mathbf{x}_i$.
- Existem C classes que devem ser $C << N$. Ou seja, $\mathbf{y}_i \in \{y_1, y_2, \cdots, y_C\}$.

Classificador Bayesiano Gaussiano.

- Apenas como uma revisão rápida, a probabilidade *a priori* é definida como:

$$P(y_i) = \frac{\text{Possíveis Resultados Favorecendo o Evento } y_i}{\text{Número total de resultados possíveis}}$$

- A probabilidade de um evento é o número de vezes em que o resultado desejado pode ocorrer pelo total de resultados possíveis.
- Considerando um exemplo em que há um conjunto de dados com $N = 5000$ e $p = 2$ e $C = 5$, em que exista um balanceamento total de classes, qual é a probabilidade *a priori* para cada classe?

Classificador Bayesiano Gaussiano.

- Apenas como uma revisão rápida, a probabilidade:

$$P(\mathbf{x}_i|y_i) = ?$$

- representa uma função de densidade que explica como os dados estão organizados.
- Encontrar tal função não é uma tarefa simples, pois, necessita-se saber qual é a distribuição em que esses dados foram gerados.
- Contudo, uma suposição inicial e interessante é de que a distribuição é uma normal (gaussiana).
- Na área de reconhecimento de padrões, essa densidade de probabilidade é comumente chamada de função de **verossimilhança**.
- Desta maneira, pode-se utilizar a FDP da gaussiana para encontrar $P(\mathbf{x}_i|y_i)$.

Classificador Bayesiano Gaussiano.

- Apenas como uma revisão rápida, a probabilidade:

$$P(\mathbf{x}_i|y_i) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(\mathbf{x}_i-\mu_x)^2}{2\sigma^2}}$$

- representa uma função de densidade que explica como os dados estão organizados.
- Encontrar tal função não é uma tarefa simples, pois, necessita-se saber qual é a distribuição em que esses dados foram gerados.
- Contudo, uma suposição inicial e interessante é de que a distribuição é uma normal (gaussiana).
- Na área de reconhecimento de padrões, essa densidade de probabilidade é comumente chamada de função de **verossimilhança**.
- Desta maneira, pode-se utilizar a FDP da gaussiana para encontrar $P(\mathbf{x}_i|y_i)$.
- Em que σ é o desvio padrão e μ_x.

Classificador Bayesiano Gaussiano.

- Apenas como uma revisão rápida, a probabilidade (*a posteriori*):

$$P(y_i|\mathbf{x}_i) = \frac{P(y_i)P(\mathbf{x}_i|y_i)}{\mathbf{x}_i}$$

- É interpretada da seguinte forma: a partir de um novo dado $\mathbf{x}_i$, qual é a probabilidade de y_i ocorrer?
- Neste caso, pode-se utilizar o teorema de Bayes.

Classificador Bayesiano Gaussiano.

- Apenas como uma revisão rápida, a probabilidade (*a posteriori*):

$$P(y_i|\mathbf{x}_i) = \frac{P(y_i)P(\mathbf{x}_i|y_i)}{P(\mathbf{x}_i)} = \frac{priori \times verossimilhança}{\text{evidência}}$$

- É interpretada da seguinte forma: a partir de um novo dado $\mathbf{x}_i$, qual é a probabilidade de y_i ocorrer?
- Neste caso, pode-se utilizar o teorema de Bayes.
- Na prática o interesse está no numerador dessa equação, pois, o denominador não é dependente de y e o padrão de entrada é dado.

Classificador Bayesiano Gaussiano.

- Um critério para tomada de decisão comumente utilizado é o critério de *máxima a posteriori*.
- Assim, para um dado $\mathbf{x}$, realiza-se a atribuição à $c-$ésima classe para aquela em que a densidade a posteriori é a maior,

$$\hat{y} = \arg\max \{P(y_c|\mathbf{x})\} \qquad i = 1, \cdots, C$$

- Novamente, o operador *arg max* retorna o maior argumento (índice).

Classificador Bayesiano Gaussiano.

- Um critério para tomada de decisão comumente utilizado é o critério de *máxima a posteriori*.
- Assim, para um dado $\mathbf{x_i}$, realiza-se a atribuição à $c-$ésima classe para aquela em que a densidade a posteriori é a maior,

$$\hat{y} = \arg\max \{P(y_c) \prod_{i=1}^{p} P(y_c|\mathbf{x}_i)\} \quad c = 1, \cdots, C$$

- Novamente, o operador *arg max* retorna o maior argumento (índice).
- Considerando um problema com $p > 1$ preditores, pode-se reescrever

Classificador Bayesiano Gaussiano.

- Um critério para tomada de decisão comumente utilizado é o critério de *máxima a posteriori*.
- Assim, para um dado $\mathbf{x_i}$, realiza-se a atribuição à $c-$ésima classe para aquela em que a densidade a posteriori é a maior,

$$\hat{y} = \arg\max \{P(y_c) \prod_{i=1}^{p} P(y_c|\mathbf{x}_i)\} \quad c = 1, \cdots, C$$

- Novamente, o operador *arg max* retorna o maior argumento (índice).
- Considerando um problema com $p > 1$ preditores, pode-se reescrever.
- É comum que ao aplicar o produtório, os valores tenham um limite para zero. Desta maneira aplica-se a função log natural:

$$\hat{y} = \arg\max \ln \left[\{P(y_c) \prod_{i=1}^{p} P(y_c|\mathbf{x}_i) \right] \} \quad c = 1, \cdots, C$$

Classificador Bayesiano Gaussiano.

Algorithm 2: Pseudocódigo da do modelo classificador gaussiano.

1: Dividir os dados em treinamento e teste.
2: Para o conjunto de dados para treinamento, calcular as probabilidades a priori para cada classe.
3: Faça o cálculo de médias e desvio-padrão para cada massa de dados associado às classes.
4: **for** cada amostra $\mathbf{x}_{teste}$ no conjunto $\mathbf{X}_{teste}$ **do**
5: $\quad \hat{y} \longleftarrow \arg\max \ln \left[\{P(y_c) \prod_{i=1}^{p} P(y_c|\mathbf{X}_{teste} \right] \} \qquad c = 1, \cdots, C$
6: **end for**
7: Compute $\frac{1}{N} \sum_{i=1}^{N} I(y_i \neq \hat{y}_i)$ para os conjuntos de treino e teste.