Pós-Graduação

Visão Global da Tecnologia da Informação

Aula 3

Prof. Jones Egydio

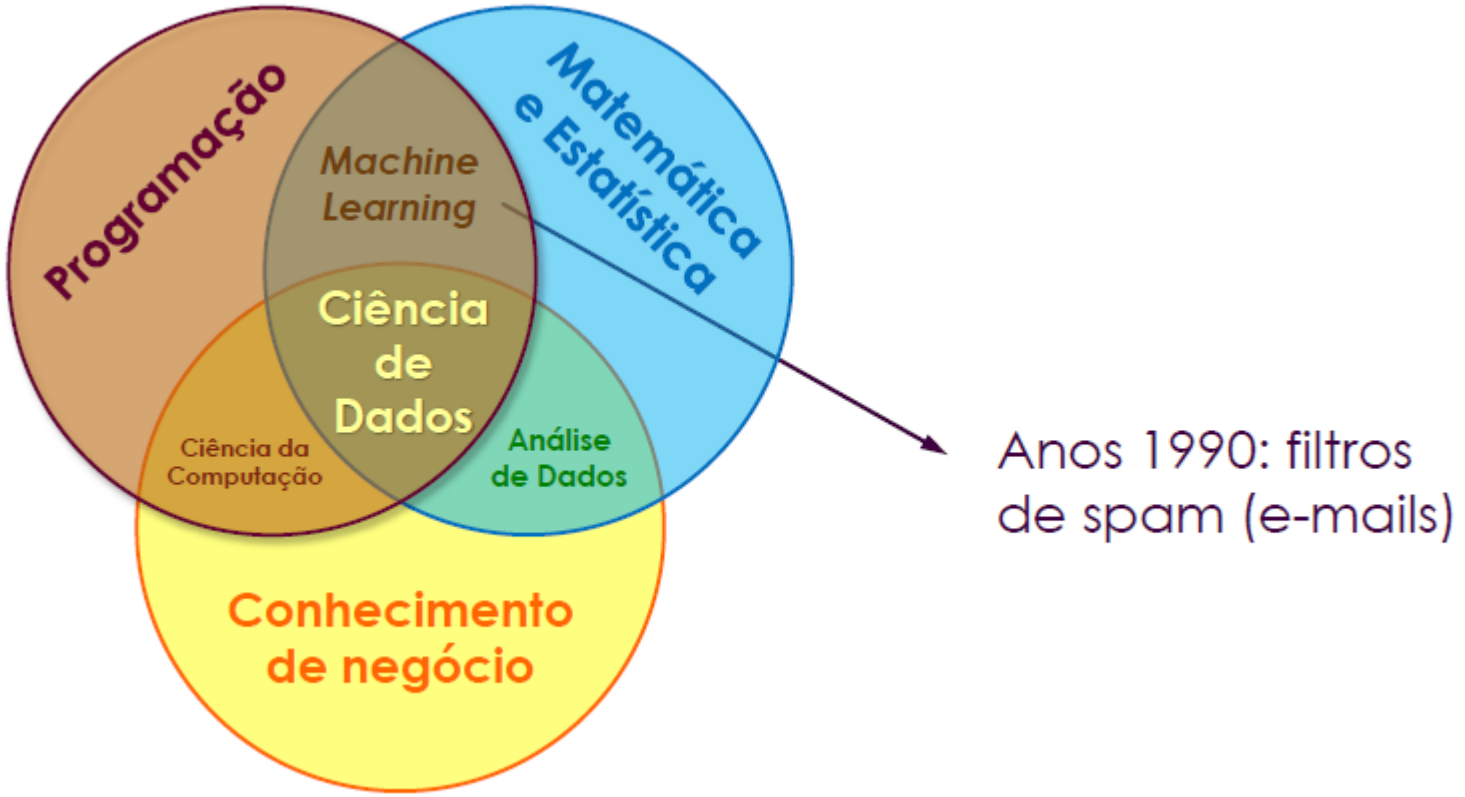
jones.egydio@maua.br

Maio/2023

Objetivos

- Introdução aos modelos de regressão;
- Aplicações com dados de exemplo;
- Trabalhar com modelos de *Machine Learning* no Python;
- Aplicações utilizando dados externos;
- Análise dos resultados;
- Atividade para entrega.

Machine Learning



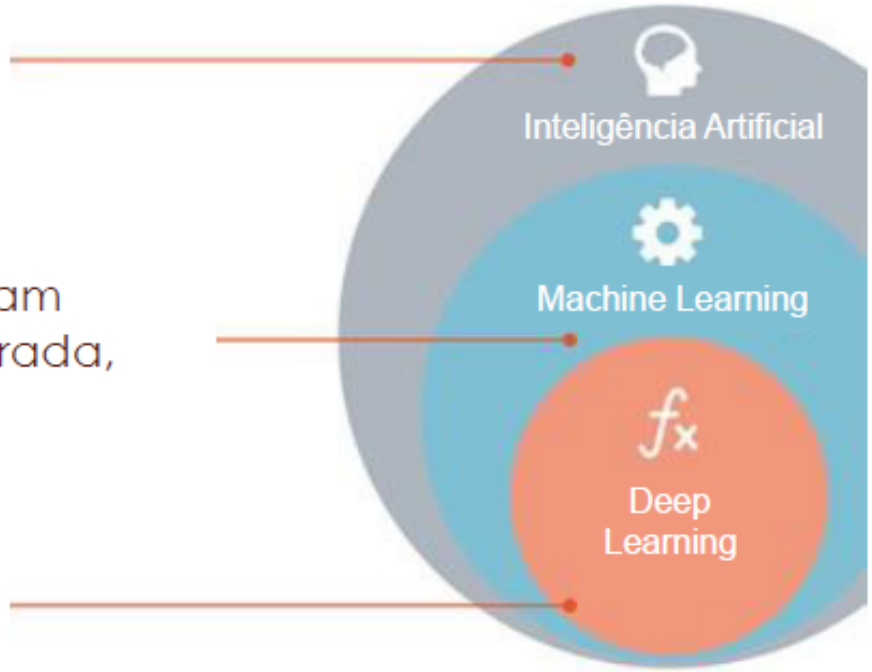
Fonte: PEREIRA, W., Notas de Aula – Machine Learning.

Machine Learning

Inteligência artificial
Técnicas que permitem aos computadores imitar comportamentos humanos.

Machine Learning
Conjunto de técnicas da IA que capacitam máquinas a aprender com dados de entrada, sem uma programação explícita.

Deep Learning
Conjunto de técnicas de ML que usam redes neurais multinível para gerar soluções altamente especializadas.



Fonte: PEREIRA, W., Notas de Aula – Machine Learning.

Machine Learning

- Algumas aplicações:

  Saúde: diagnósticos médicos;

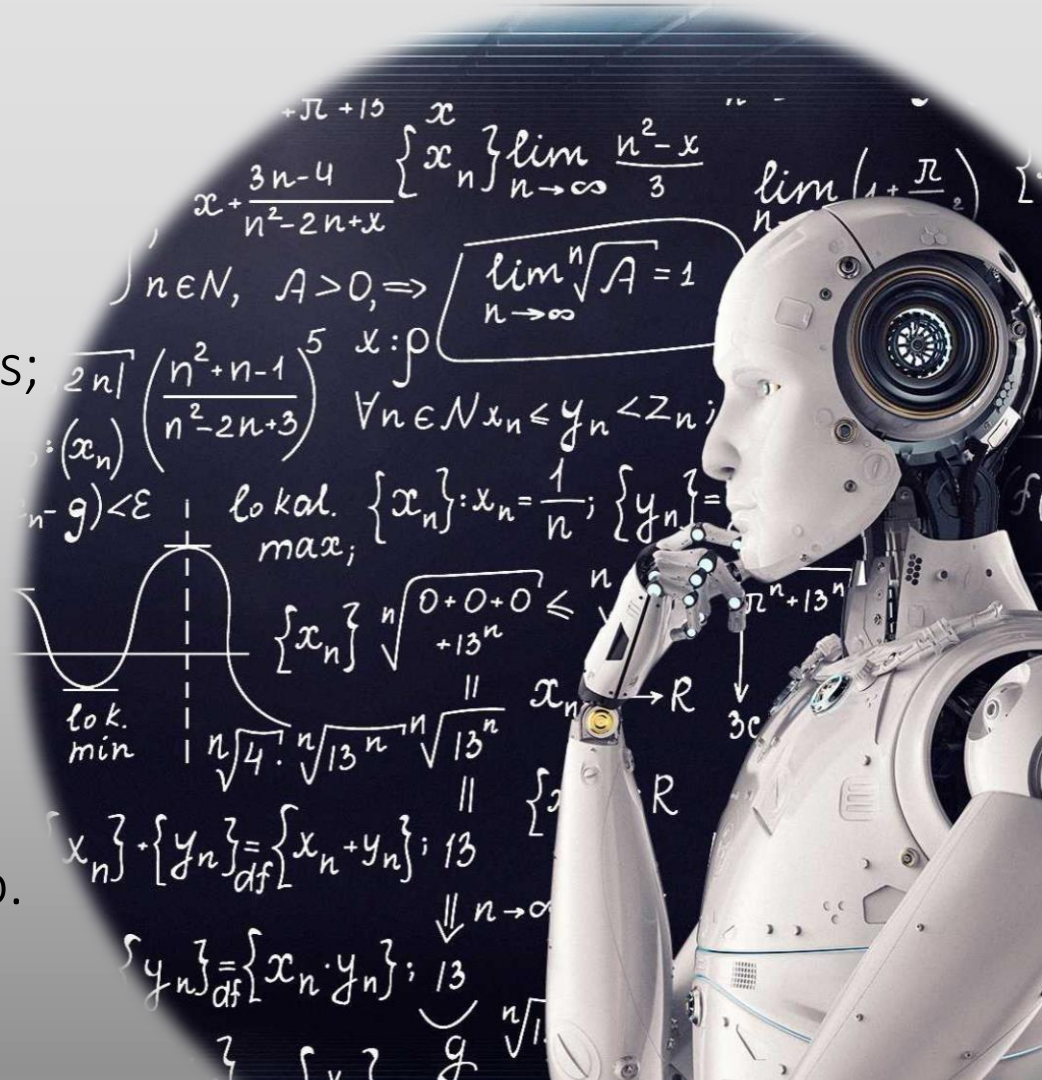
  Finanças: análise de crédito, detecção de fraudes;

  Varejo: previsão de vendas e gestão de estoque;

  Governo: cidades inteligentes;

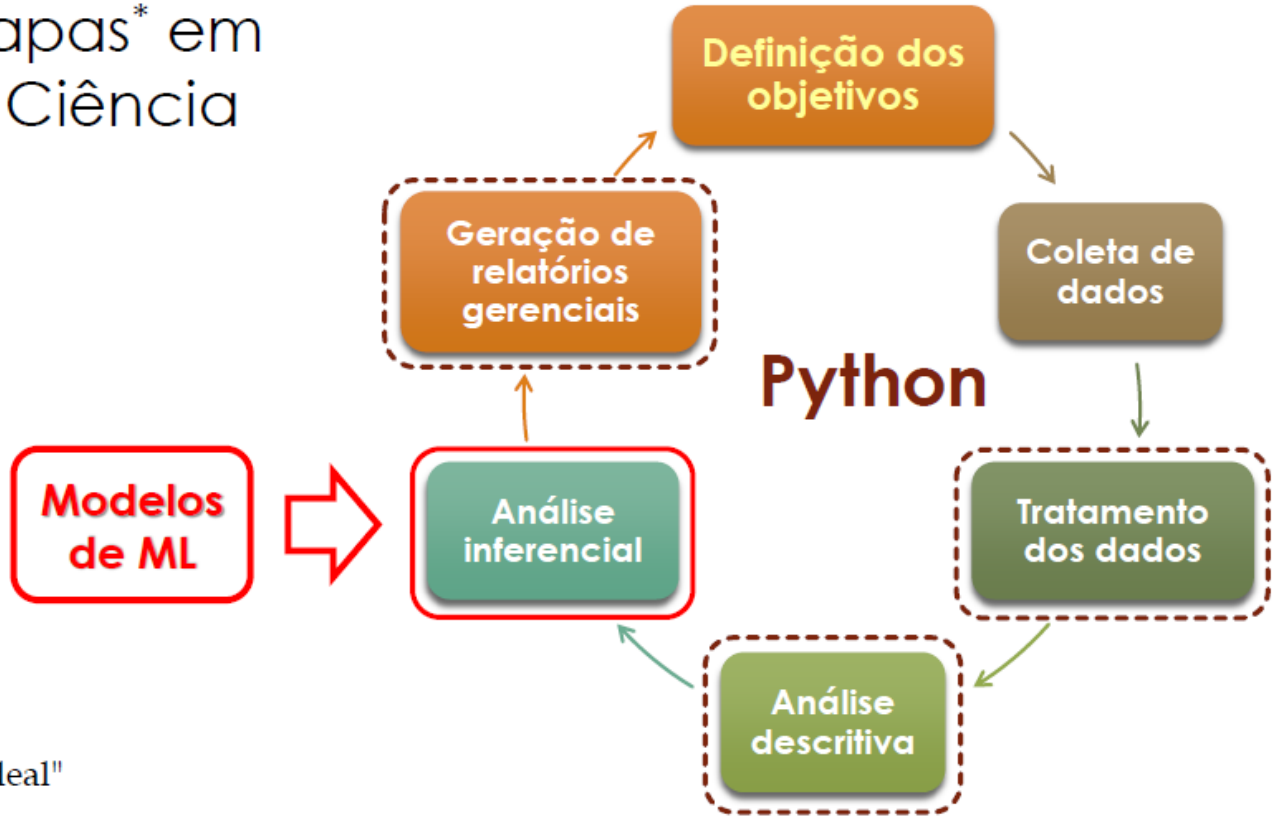
  Redes de comunicação: detecção de invasões;

  Comércio eletrônico: sistemas de recomendação.



Machine Learning

Ciclo de etapas* em projetos de Ciência de Dados:



* Ciclo simplificado e "ideal"

Fonte: PEREIRA, W., Notas de Aula – Machine Learning.

Machine Learning

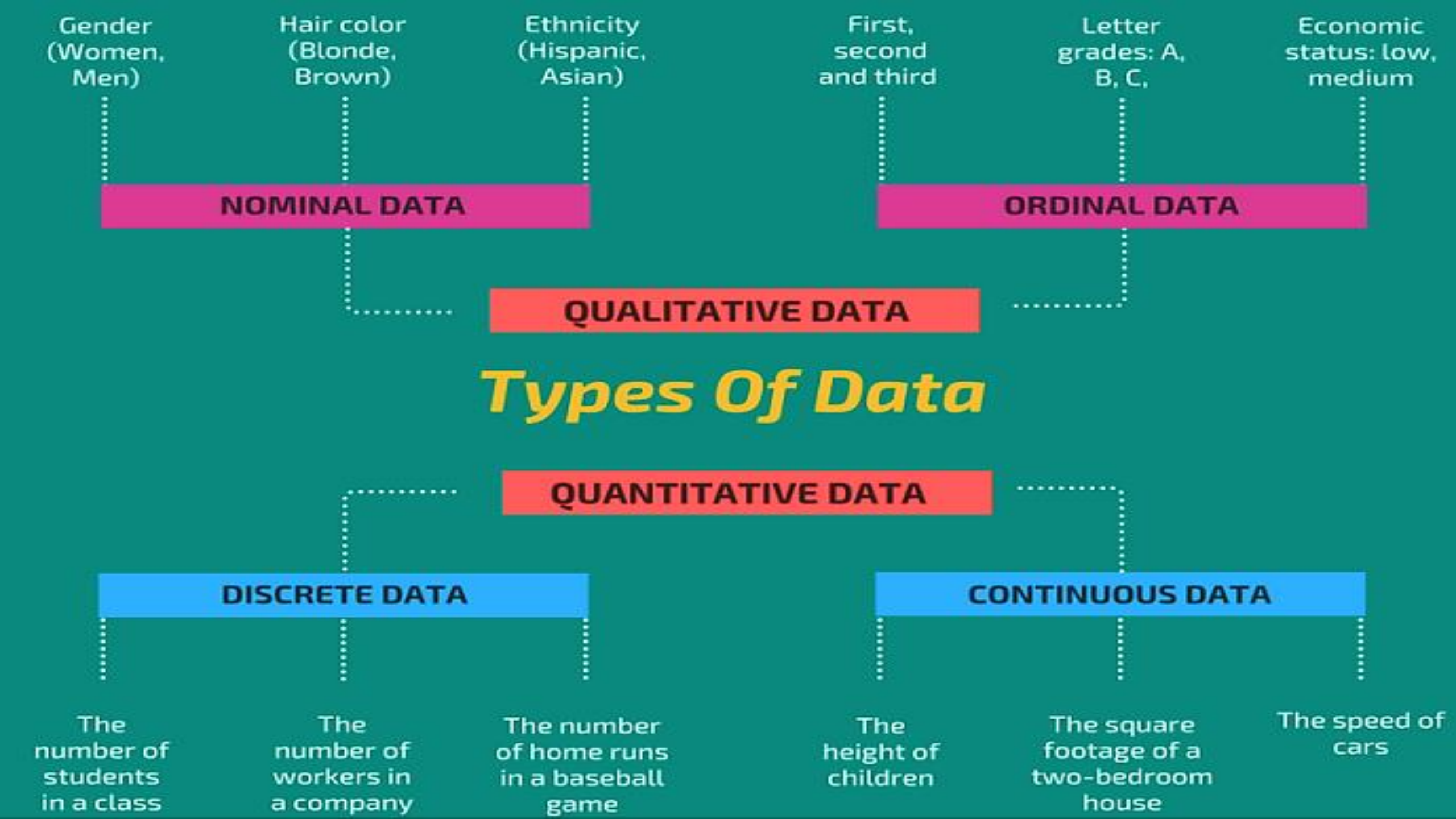
A **análise inferencial** pertence ao ramo da Estatística responsável por gerar modelos de comportamentos populacionais a partir de resultados amostrais:

Fonte: PEREIRA, W., Notas de Aula – Machine Learning.

Machine Learning

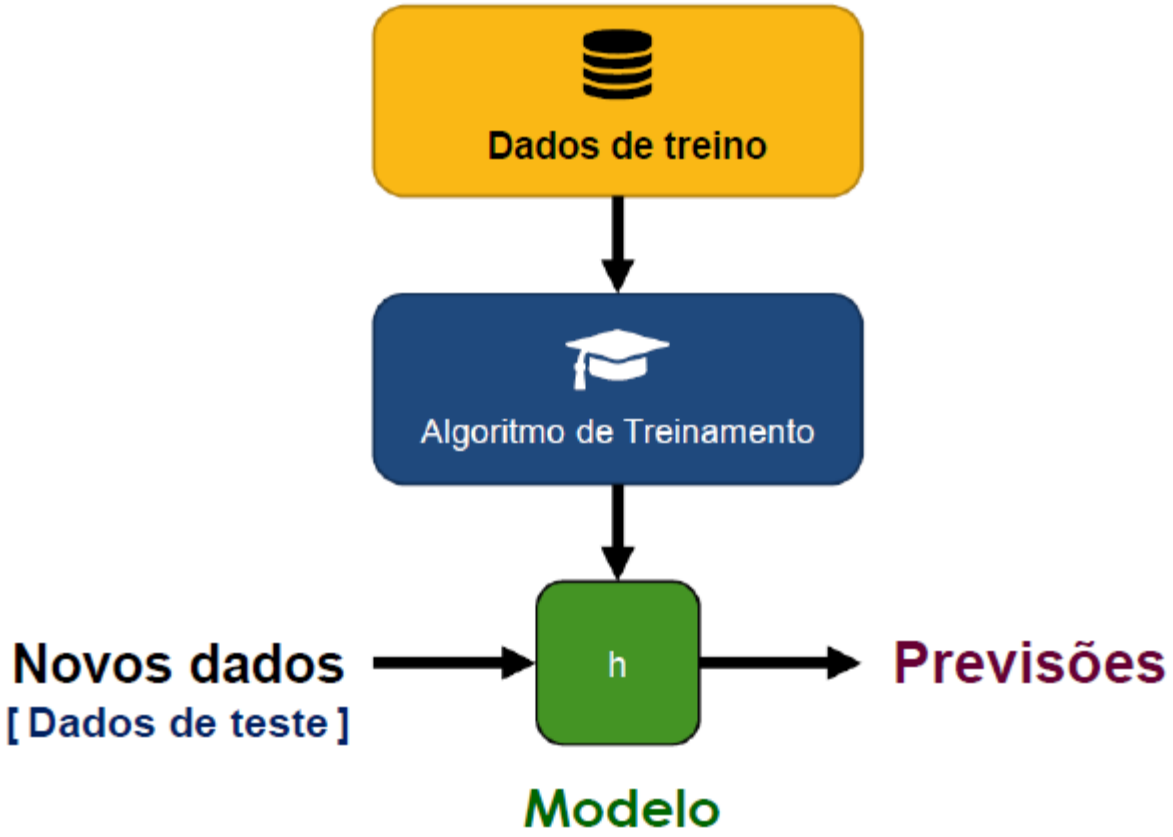
- **Em geral, não se estudam populações. Motivos:**
  - Falta de tempo e recursos (humanos/financeiros) para coletar os dados de interesse;
  - Evitar o eventual desperdício de produtos (exemplo: testes de durabilidade);
  - Questões éticas (ex.: ensaios clínicos de medicamentos);
  - Possibilidade de se obter resultados com graus controlados de confiabilidade a partir de amostras.

Em alguns casos, no entanto, amostras não bastam (ex.: censo populacional).


Gender (Women, Men)
Hair color (Blonde, Brown)
Ethnicity (Hispanic, Asian)
First, second and third
Letter grades: A, B, C,
Economic status: low, medium
NOMINAL DATA
ORDINAL DATA
QUALITATIVE DATA
Types Of Data
QUANTITATIVE DATA
DISCRETE DATA
CONTINUOUS DATA
The number of students in a class
The number of workers in a company
The number of home runs in a baseball game
The height of children
The square footage of a two-bedroom house
The speed of cars

Machine Learning

- Modelos: representações simplificadas do comportamento de sistemas.
- Permitem entender as relações entre as variáveis envolvidas e prever as respostas para novas entradas.



Machine Learning

- Tipos de variáveis:

Variáveis determinísticas	Variáveis aleatórias
$y = 25 + 0,5x$	X = idade de uma criança
Custo de R$ 25,00 acrescido de 0,5 para cada unidade de x.	y = tamanho do vocabulário da criança

Machine Learning

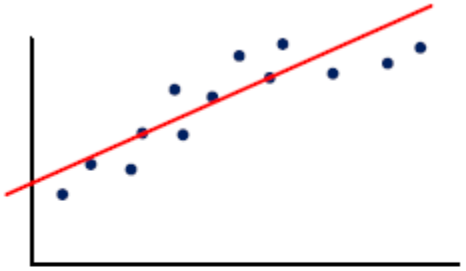
- Serão abordados modelos de regressão, que são expressões matemáticas que tentam explicar como os valores de uma variável resposta (ou dependente) são afetados pelos valores de uma ou mais variáveis explicativas (ou independentes).
- Só faz sentido criar um modelo de regressão entre variáveis se existir alguma associação entre elas. A investigação inicial é feita, geralmente, de forma gráfica.
- Testes estatísticos podem ser usados para comprovar a significância das relações.

Machine Learning

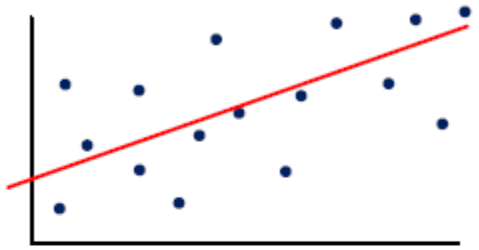
- Os primeiros modelos de regressão a serem discutidos envolverão apenas variáveis quantitativas. A investigação gráfica das possíveis associações serão feitas por meio de diagramas de dispersão.
- A associação mais simples entre variáveis é a associação linear, em que o aumento de uma das variáveis é acompanhado por um aumento ou diminuição da outra e vice-versa, com taxa fixa de crescimento ou decrescimento. Note que esta análise não depende da relação de causa e efeito entre as variáveis.

Machine Learning

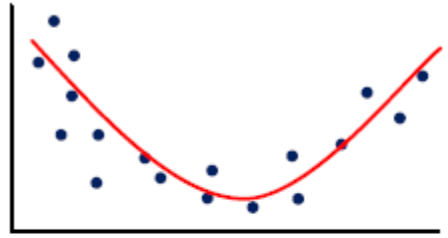
- **Exemplos de tipos de associação entre variáveis:**



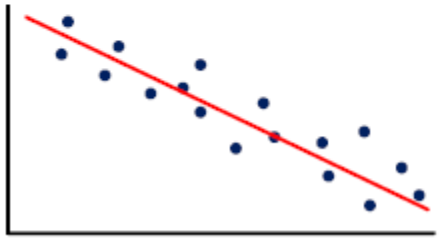
Linear, positiva e forte



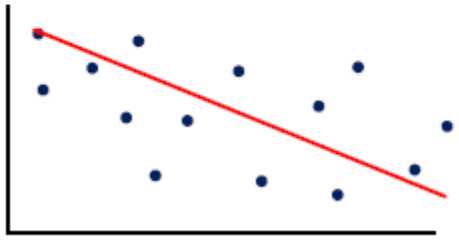
Linear, positiva e fraca



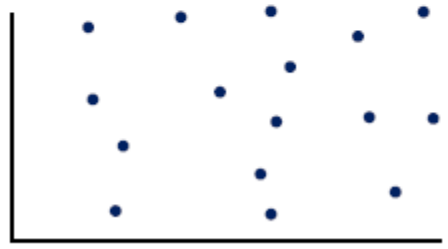
Não linear (quadrática)



Linear, negativa e forte



Linear, negativa e fraca



Muito fraca/inexistente

Machine Learning

- O grau de relacionamento linear entre duas variáveis é tipicamente medido pelo coeficiente de correlação de Pearson, denotado por *r*. O valor deste coeficiente varia entre -1 e 1, sendo que, quanto mais próximo de 1 ou -1, maior a força da correlação positiva ou negativa, respectivamente. Critério proposto:

$\|r\| \leq 0,40$	**Correlação fraca**
$0,40 \leq \|r\| \leq 0,70$	**Correlação moderada**
$\|r\| \geq 0,70$	**Correlação forte**

Machine Learning

- Assim, um modelo de regressão linear simples de uma variável quantitativa *y* (dependente) em função de uma variável (independente) *x*, da forma

$$\hat{y} = \theta_0 + \theta_1 x + \varepsilon$$

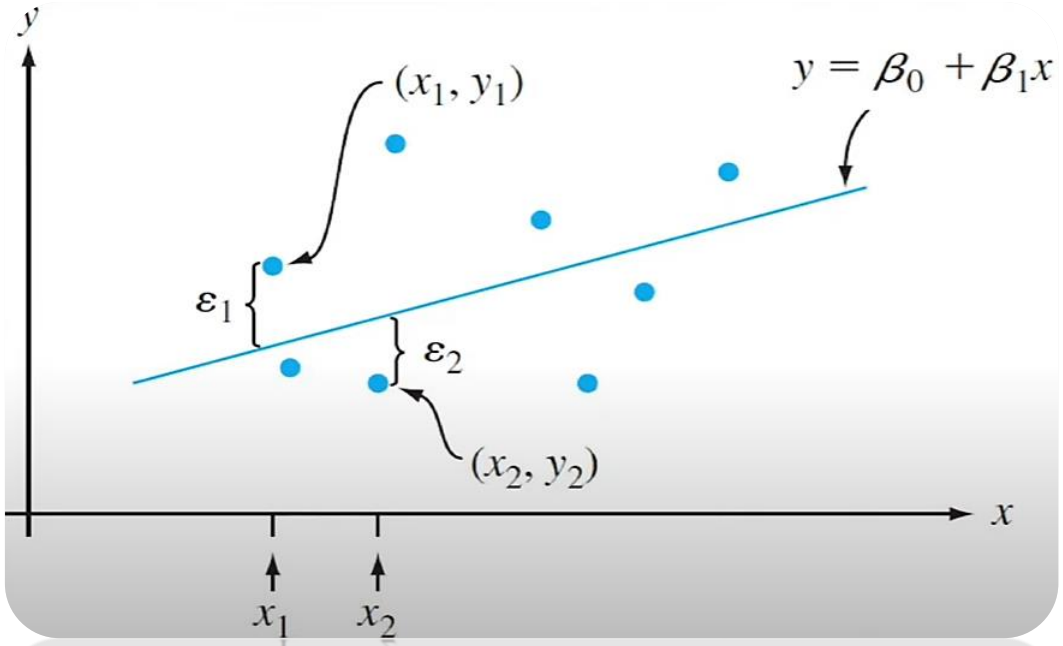
pode ser obtido computacionalmente para estimar a verdadeira relação

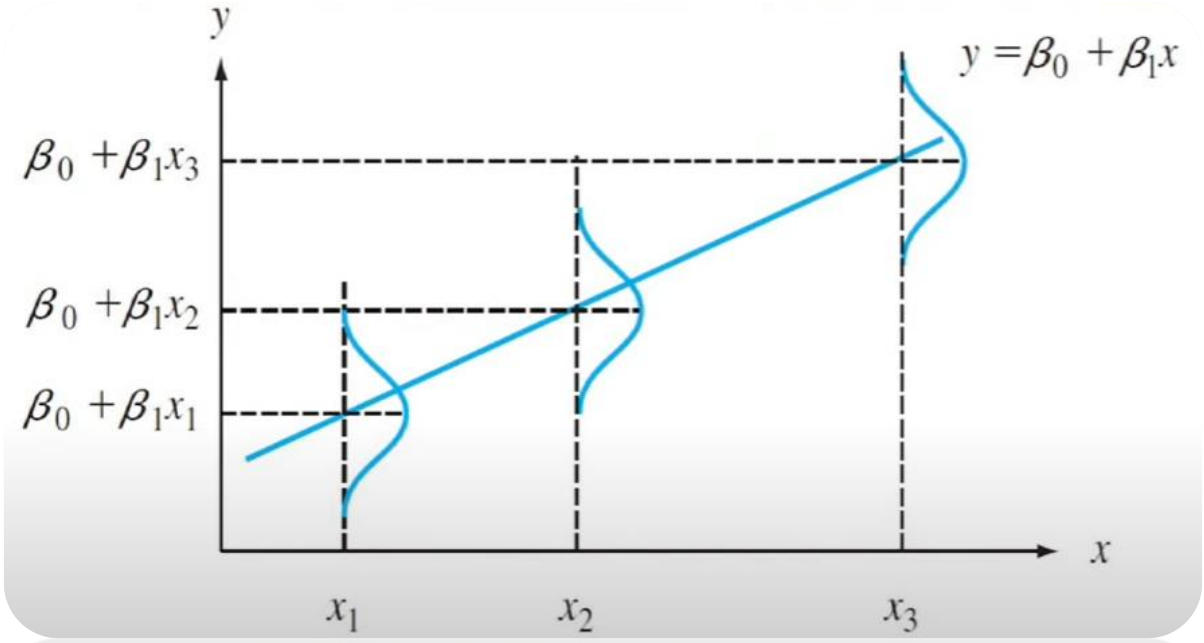
$$\hat{y} = b_0 + b_1 x + \varepsilon$$

sendo ε uma componente aleatória, ou desvio aleatório, ou termo do erro aleatório. Possui uma distribuição normal com $E(\varepsilon) = 0$ e $V(\varepsilon) = \sigma^2$.

Machine Learning

$$\hat{y} = b_0 + b_1 x + \varepsilon$$



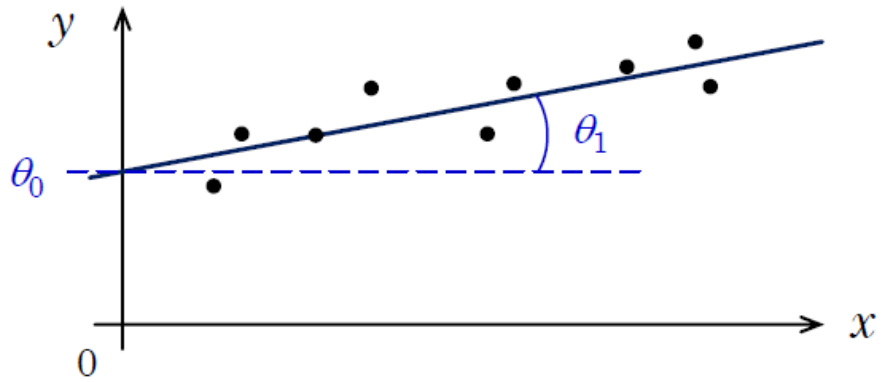


Machine Learning

- Os valores θ_0 e θ_1 são estimativas da relação (supostamente) linear entre as variáveis *x* e *y*. De forma geral:

- θ_1 = coeficiente angular: representa a variação estimada no valor de y para uma variação unitária no valor de x.

- θ_0 = coeficiente linear: valor estimado de *y* para *x = 0*.



Machine Learning

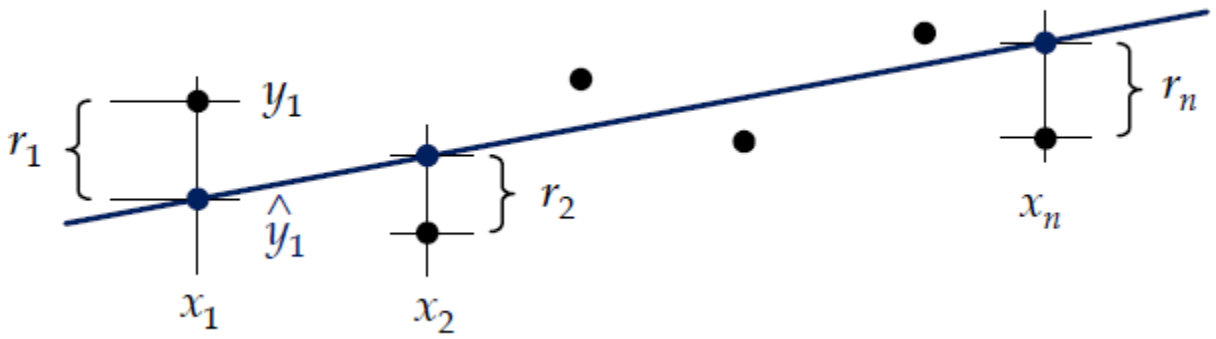
- O processo de cálculo dos coeficientes θ_0 e θ_1 (e outras informações que serão discutidas futuramente) é chamado de **treinamento do modelo** e os dados usados no processo são chamados de **dados de treino**.
- O resultado do treinamento de um modelo depende diretamente do algoritmo de treinamento e dos dados de treino utilizados.

n valores

x	y
x_1	y_1
x_2	y_2
...	...
x_n	y_n

Machine Learning

- Os coeficientes dos modelos de regressão linear geralmente são calculados por um algoritmo que minimiza a soma dos quadrados dos resíduos de ajuste (r_i), dados pelas diferenças entre os valores observados (y_i) e os valores previstos pelo modelo ($\hat{y}_i$).



$$\hat{y} = \theta_0 + \theta_1 x$$

n = tamanho da amostra de dados

Machine Learning

- **No Python, é rotineiramente utilizado para esses modelos:**

Biblioteca contendo definições e métodos de cálculo para *arrays* e matrizes multidimensionais.

Biblioteca para criação de gráficos e visualizações de dados, feita para a linguagem Python e sua extensão de matemática, **NumPy**.

Biblioteca criada para a linguagem Python para manipulação e análise de dados em tabelas numéricas e séries temporais.

Biblioteca de algoritmos de aprendizado de máquina para a linguagem Python mais adotada atualmente.

Exemplo 1

- **Objetivo**
  - Trabalhar com um modelo de regressão linear simples no Python.

Modelo esperado:

$$\hat{y} = -1696{,}192 + 9349{,}40x$$

Para $x = 0{,}5$:

$$\hat{y} = -1696{,}192 + 9349{,}40(0{,}5) = 2978{,}51$$

Colaboratory

x	y
15,5	2.158,7
23,75	1.678,15
8	2.316
17	2.061,3
5	2.207,5
...	...
21,5	1.753,7

Imagem ilustrativa

Exemplo 2

- **Objetivo**
  - Trabalhar com um modelo de regressão linear múltipla no Python.

Modelo esperado:

$$\hat{y} = -8{,}188 + 0{,}8347x_1 + 0{,}041x_2$$

Para $x_1 = 30$ e $x_2 = 4{,}5$:

$$\hat{y} = -8{,}188 + 0{,}8347(30) + 0{,}041(4{,}5) = 17{,}040$$

x_1	x_2	y
2	50	9,95
8	110	24,45
11	120	31,75
10	550	35
8	295	25,02
...	...	...
5	400	21,15

Imagem ilustrativa

Colaboratory

Machine Learning

- Reforçando o que foi discutido anteriormente: só faz sentido criar um modelo de regressão linear entre variáveis que apresentem um relacionamento linear. A avaliação pode ser feita graficamente ou por testes estatísticos sobre o coeficiente de correlação linear.

- No entanto, mesmo quando a relação entre as variáveis não for linear, em muitos casos é possível torná-la linear por meio de uma transformação conveniente. Os modelos que admitem este tipo de transformação são chamados de intrinsecamente lineares.

Machine Learning

- Após a transformação, os dados alterados são ajustados por um modelo de regressão linear. Se o modelo for significativo e tiver boa qualidade de ajuste, calculam-se os parâmetros do modelo (não linear) original.
- A função não linear a ser estimada (e, consequentemente, a transformação a ser aplicada) pode ser escolhida por observação gráfica ou por conhecimento do modelo teórico do fenômeno em questão.
- A seguir será mostrado um exemplo de transformação e, na sequência, outras formas usadas na prática.

Machine Learning

Exemplo 3. Considere que as variáveis x e y estão relacionadas pela **função potência**: $y = \alpha x^{\beta}$. Neste caso, é possível utilizar a seguinte transformação:

$$ln(y) = ln(\alpha x^{\beta}) \Rightarrow ln(y) = ln(\alpha) + \beta ln(x)$$

Em seguida, efetua-se uma mudança de variáveis:

$$z = ln(y);\ a = ln(\alpha);\ b = \beta;\ w = ln(x)$$

Com isso, obtém-se o modelo linear: $z = a + bw$.

Machine Learning

Transformação e resultados:

x	y
1,0	3,5
2,0	5,4
3,0	9,0
4,0	14,8
5,0	24,0
6,0	29,3

$y = \alpha x^{\beta}$

w = ln(x)	z = ln(y)
0	1,25276
0,69315	1,6864
1,09861	2,19722
1,38629	2,69463
1,60944	3,17805
1,79176	3,37759

$z = a + bw$

Numpy: função ***log***

Regressão linear:

$z = 1{,}0445 + 1{,}2341w$

$R^2 = 0{,}9515$

Desfazendo a transformação:

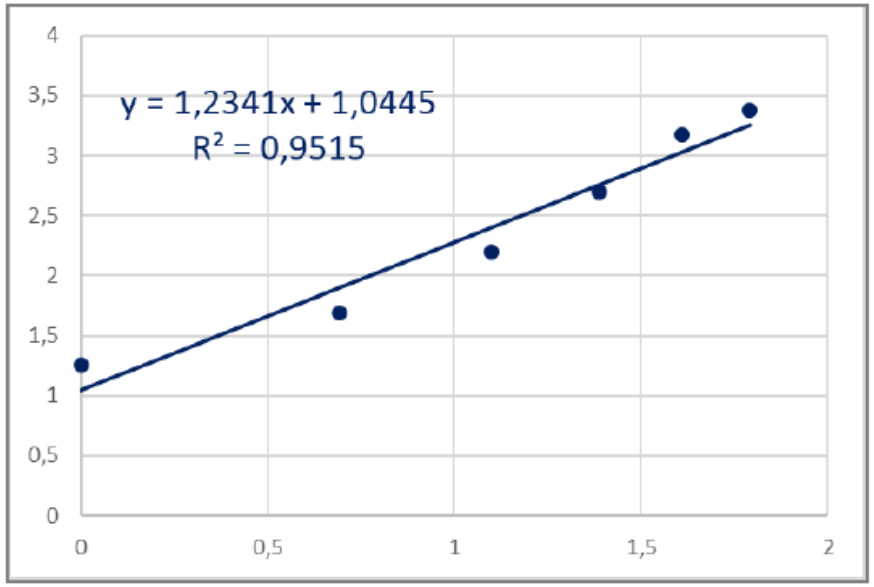
$\beta = b = \mathbf{1,2341}$

$\alpha = e^{a} = e^{1{,}0445} = \mathbf{2,842}$

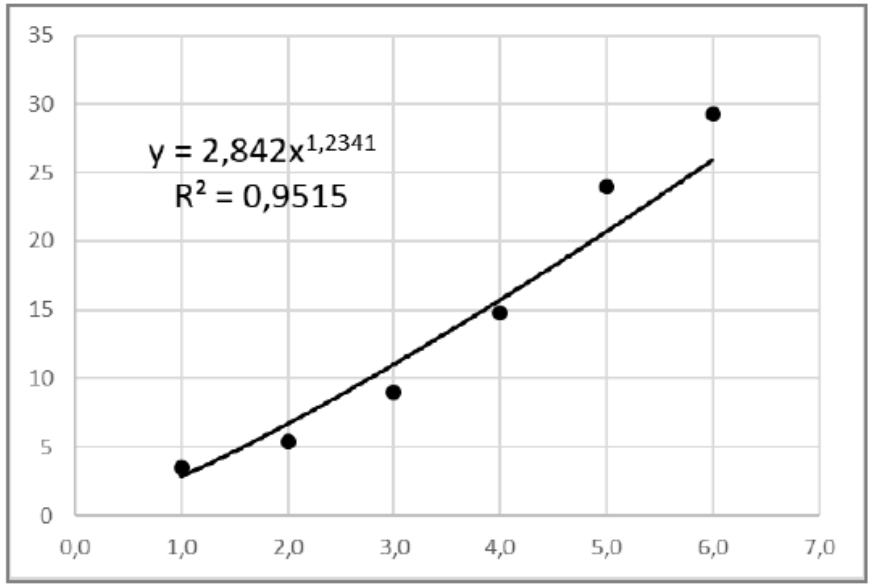
$$y = 2{,}842x^{1{,}2341}$$

Machine Learning

Dados transformados



Dados originais



Machine Learning

Caso polinomial. Suponha que as variáveis x e y se relacionem pelo polinômio: $\hat{y} = b_0 + b_1 x + b_2 x^2$. Neste caso, uma regressão linear múltipla pode ser usada, transformando x e x^2 em duas "novas" variáveis: $x_1 = x;\ x_2 = x^2$.

- Os três coeficientes da regressão linear de y em função de x_1 e x_2 serão, respectivamente, os coeficientes do polinômio de segundo grau.
- Este tipo de transformação tende a gerar modelos de regressão não significativos devido à correlação natural entre as variáveis (afinal, uma é função da outra).

Machine Learning

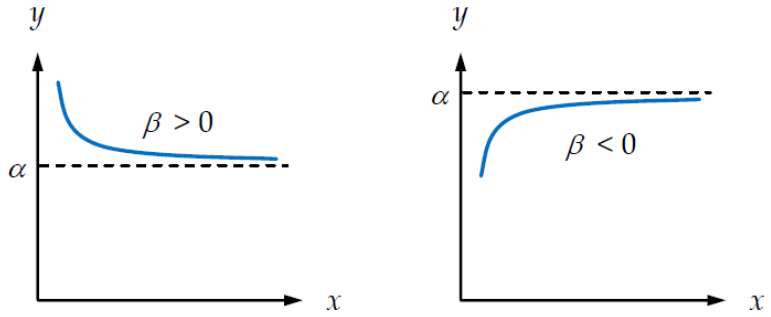
Função hiperbólica: $y = \alpha + \beta/x$

Transformação: $w = 1/x$

Variáveis: $z = y$; $a = \alpha$; $b = \beta$; $w = 1/x$

Modelo linear: $z = a + bw$

Parâmetros do modelo original = modelo linear



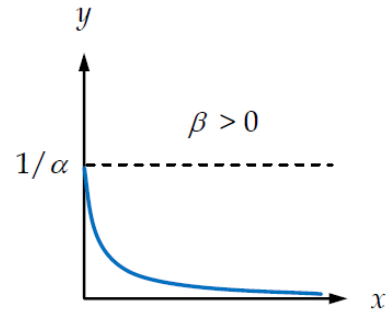
Funções do tipo: $y = 1/(\alpha + \beta x)$

Transformação: $z = 1/y$

Variáveis: $z = 1/y$; $a = \alpha$; $b = \beta$; $w = x$

Modelo linear: $z = a + bw$

Parâmetros do modelo original = modelo linear



Machine Learning

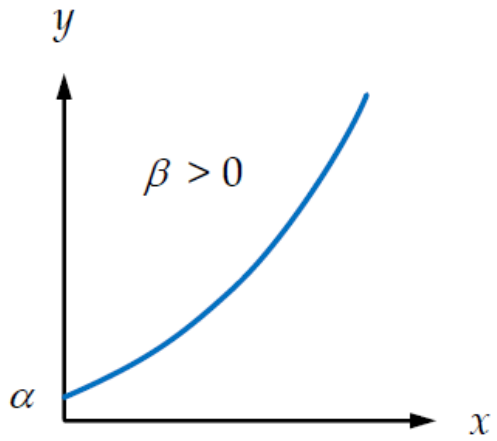
Função exponencial: $y = \alpha e^{\beta x}$

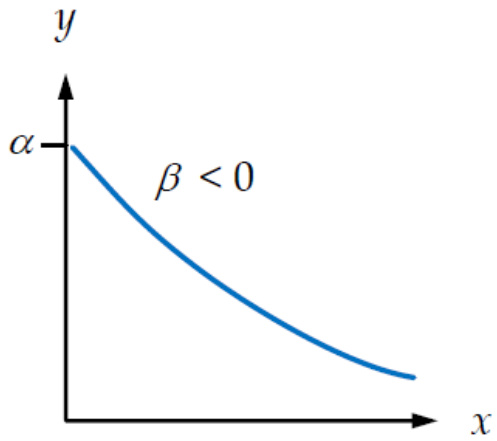
Transformação: $ln(y) = ln(\alpha) + \beta x$

Variáveis: $z = ln(y)$; $a = ln(\alpha)$; $b = \beta$; $w = x$

Modelo linear: $z = a + bw$

Parâmetros do modelo original: $\alpha = e^{a}$; $\beta = b$





Machine Learning

- Atividade:
  - Resolver os dois exercícios descritos no notebook [2023_PG_VGTI]_AT_Lec_03.ipynb, disponível para download juntamente com o material da aula na página da disciplina no Open LMS.
  - Prazo de entrega: 11/06/2023.

Referências bibliográficas

- BRUCE, P., BRUNCE, A., Estatística Prática para Cientistas de Dados, Alta Books, São Paulo, 2019;
- PEREIRA, W., Notas de Aula - Machine Learning Métodos não-probabilísticos, 2022;
- MONTGOMERY, D. C., RUNGER, G. C. Estatística aplicada e probabilidade para engenheiros. 5. ed. Rio de Janeiro, RJ: LTC, 2013. 521 p.;
- DEVORE, J. L. Probabilidade e estatística: para engenharia e ciências. São Paulo: Thomson, 2006. 692 p.;
- GÉRON, A. Hands-on machine learning with Scikit-Learn & Tensor-Flow: concepts, tolls, and techniques to build intelligent systems. Sebastopol, CA: O'Reilly, c2017. 548 p.
- MARCHESE, R. M., Notas de Aula - ADA, São Paulo, 2023. Disponível em: https://github.com/renatapink/DS_Hypera_960/tree/main

Obrigado!