

**MBA
USP
ESALQ**

Técnicas de Machine Learning

Prof. Dr. Wilson Tarantin Junior

*A responsabilidade pela idoneidade, originalidade e licitude dos conteúdos didáticos apresentados é do professor.

Proibida a reprodução, total ou parcial, sem autorização. Lei nº 9610/98

Machine learning

- Definição

- Podem ser encontradas muitas definições para o conceito *machine learning*
- Porém, em termos de seus objetivos, pode-se entender como a utilização dos dados e de algoritmos para a produção de informações que serão relevantes para a tomada de decisão
 - Informações úteis para uma melhor tomada de decisão (*data-driven decision making*)
 - Por exemplo: criando modelos preditivos e/ou analisando a interdependência entre os dados

Banco de dados

- Definição e composição
 - O banco de dados é o objeto onde estão armazenadas as informações de interesse para a análise ou estudo em questão
 - Em muitos casos, o banco de dados contém uma **amostra**, ou seja, é um subconjunto extraído da população
 - O banco de dados é composto por variáveis e por observações
 - **Observações**: as unidades que têm suas características e atributos medidos
 - **Variáveis**: características/atributos observados, medidos ou categorizados

Banco de dados: exemplos

- Bancos de dados sobre:

- Pessoas
- Países
- Empresas
- Tarefas
- Ações da bolsa
- ...

Banco de dados

- Estrutura para uso
 - Normalmente, o banco de dados é estruturado com as variáveis em colunas e as observações em linhas em uma estrutura tabular

	Idade	Altura	Cidade	Profissão	Renda	...
Pessoa 1						
Pessoa 2						
Pessoa 3						
Pessoa 4						
Pessoa 5						
Pessoa 6						
Pessoa 7						
Pessoa 8						
Pessoa 9						
Pessoa 10						
...						

Tipos de variáveis

- As variáveis podem ser divididas em
 - **Métricas:** são as variáveis quantitativas, isto é, apresentam características que podem ser mensuradas ou contadas
 - **Não métricas:** são as variáveis qualitativas, sendo que indicam características que não podem ser medidas. Tais variáveis contêm categorias, por isto, muitas vezes, são chamadas de variáveis categóricas
 - **A identificação do tipo de variável é fundamental para a escolha da técnica que será utilizada na análise dos dados**

Tipos de variáveis: exemplos

- **Métricas (quantitativas)**

- Idade em anos
- Renda mensal em Reais
- Número de habitantes no município
- Distância em metros entre duas cidades
- Rentabilidade percentual diária de uma ação na bolsa

- **Não métricas (qualitativas)**

- Nacionalidade
- Cor do veículo
- Profissão
- Grau de escolaridade
- Respostas sim ou não a um questionário
- Escalas likert

Variáveis qualitativas

- Características principais
 - As variáveis qualitativas têm sua representação feita por meio de tabelas de distribuição de frequências ou gráficos
 - Não é possível calcular medidas de resumo como média ou desvio padrão para variáveis qualitativas
 - As tabelas de frequências apresentam as contagens observadas por categoria da variável

Variáveis quantitativas

- Características principais
 - As variáveis quantitativas podem ser representadas por diversas ferramentas, como gráficos, medidas de posição e dispersão
 - A seguir, alguns exemplos de estatísticas descritivas
 - Medidas de posição: média, mediana, quartis
 - Medidas de dispersão: variância e desvio padrão

Detalhando as variáveis

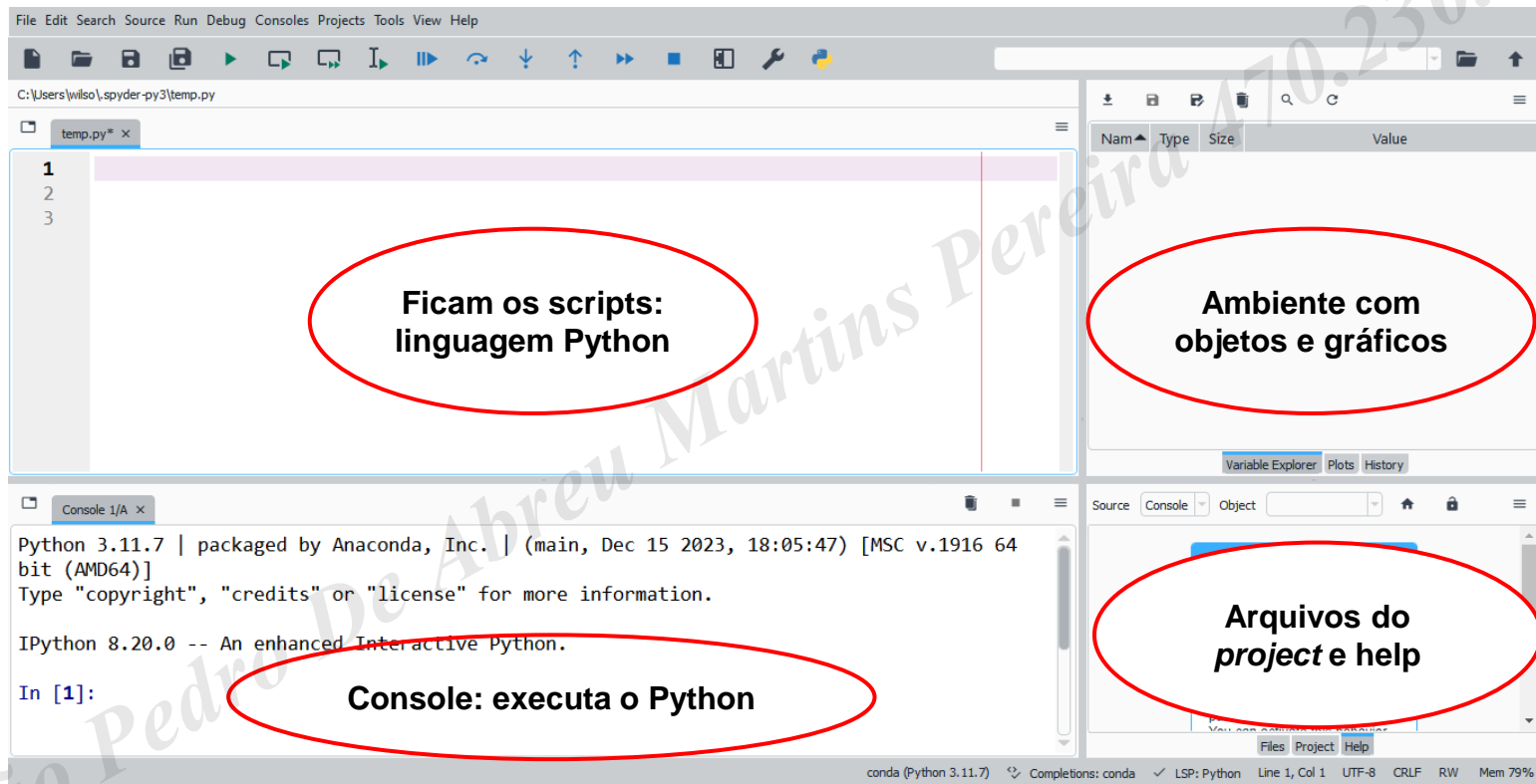
- Outras características relevantes
 - Variáveis qualitativas: dicotômica ou policotômica; nominal ou ordinal
 - Dicotômica: duas categorias (binária); Policotômica: mais de duas categorias
 - Nominal: não estabelece relação de grandeza/ordem; Ordinal: estabelece ordem
 - Variáveis quantitativas: discretas ou contínuas
 - Discretas: possuem conjunto finito e numerável de valores, em geral, são obtidas a partir de dados de contagem (0, 1, 2, 3, 4, 5...)
 - Contínuas: assumem valores pertencentes ao intervalo de números reais

Introdução ao Spyder IDE

Apresentação

- **Python:** é a linguagem de programação que vamos utilizar
- Neste curso, vamos implementá-lo por meio do **Spyder (IDE)**
 - Software que torna o uso do Python mais simples para o usuário

Ajuste: View > Window layouts > Rstudio



Acessando um *project*

- Sempre que acessar um projeto em execução, acesse o *project*:
 1. Retire a pasta do arquivo compactado (caso esteja);
 2. No Spyder acesse: Projects > Open Project > Selecione a pasta
 3. Os arquivos contidos no *project* aparecerão no ambiente

Pacotes

- Alguns pacotes que utilizaremos:
 - Pandas: manipulação e análise de dados
 - Numpy: funções matemáticas e dados
 - Matplotlib: visualização de dados em gráficos
 - Seaborn: também é um pacote gráfico
 - Plotly: gráficos interativos

Documentação

- Leituras e documentação para consulta
 - <https://pandas.pydata.org/docs/index.html>
 - <https://numpy.org/doc/stable/>
 - <https://matplotlib.org/>
 - <https://seaborn.pydata.org/>
 - <https://plotly.com/python/>

Modelos Lineares de Regressão Simples e Múltipla

Modelos lineares de regressão

- Modelos supervisionados de machine learning
 - Conhecidos como modelos confirmatórios ou técnicas de dependência
 - O objetivo é estimar modelos, equações, com o intuito de elaborar previsões
 - Portanto, há inferência dos resultados para outras observações fora da amostra
 - Define-se uma relação $Y = f(X)$
 - **Y**: chamada de variável dependente, é a variável a ser explicada (*target*)
 - **X**: chamadas de variáveis explicativas, são as preditoras (*features*)

Quando aplicar o modelo

- A regressão linear é aplicada quando a **variável dependente é quantitativa**
 - O objetivo é explicar o comportamento de Y em função de um conjunto de X
 - Estabelece-se uma relação linear entre as variáveis
- Regressão linear simples e múltipla
 - A regressão linear simples contém apenas uma variável explicativa
 - A regressão linear múltipla contém mais de uma variável explicativa

Modelo geral de regressão linear

$$Y_i = a + b_1 \cdot X_{1i} + b_2 \cdot X_{2i} + \dots + b_k \cdot X_{ki} + u_i$$

- Y é a variável dependente quantitativa
- a representa a constante (intercepto)
- b_k representam os coeficientes para cada variável explicativa
- X_k representam as variáveis explicativas do modelo
- u_i representa o termo de erro do modelo (resíduo)
 - k é o número de variáveis explicativas e i refere-se às observações em análise
- As variáveis explicativas (X) podem ser métricas ou categóricas

Mínimos quadrados ordinários (MQO)

- O algoritmo estimará os parâmetros α e β do modelo

$$\hat{Y}_i = \alpha + \beta \cdot X_i$$

- Pode-se definir o resíduo do modelo para dada observação i

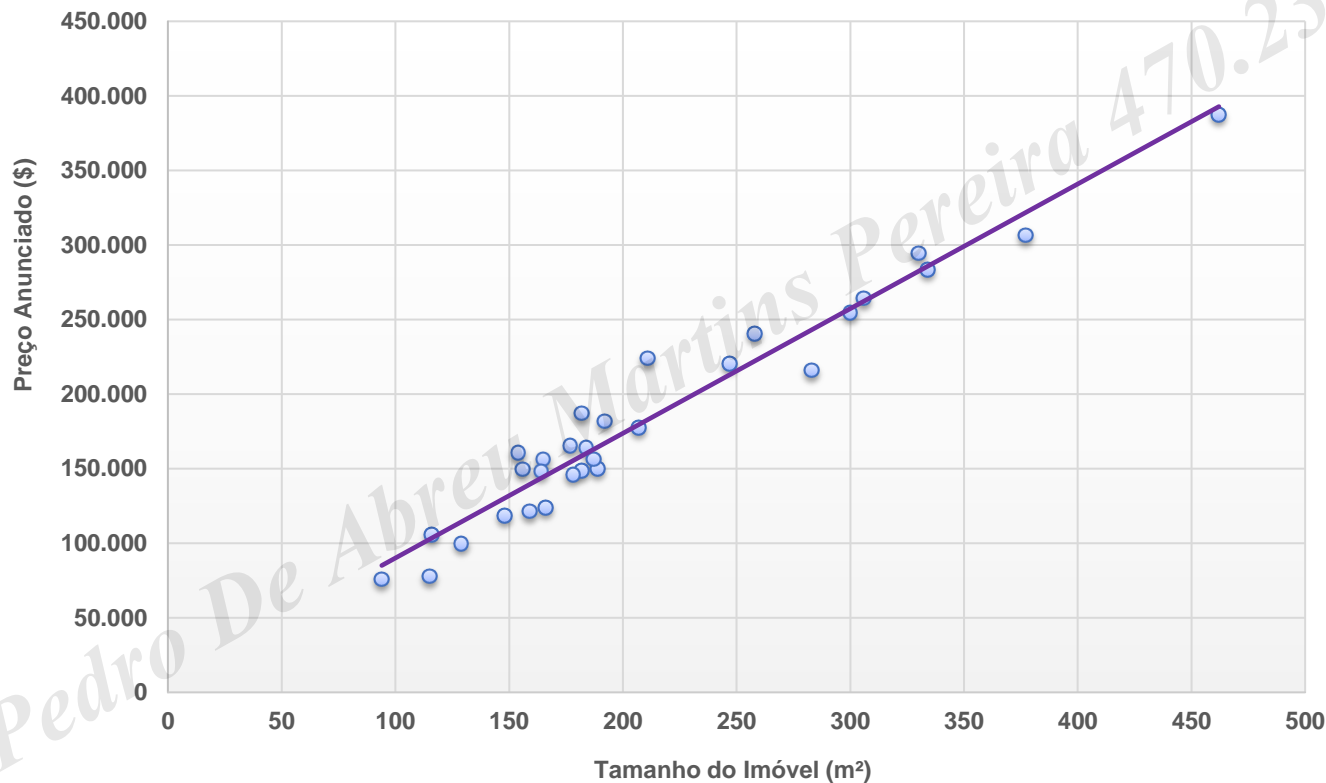
$$u_i = Y_i - \hat{Y}_i$$

- Condições para a estimação dos parâmetros do modelo (MQO)

1. A somatória dos resíduos deve ser igual a zero
2. A somatória dos resíduos ao quadrado é a mínima possível

Pode ser
visto como
P.O.

Visualizando graficamente



Elementos de um modelo

- Interpretaremos
 - Coeficientes estimados
 - Significância geral do modelo (teste F – ANOVA)
 - Significância dos parâmetros (testes t)
 - Intervalos de confiança
 - Poder explicativo do modelo (R^2)

Parâmetros do modelo

- Interpretação dos parâmetros α e β
 - α é o coeficiente linear, ou seja, o valor de Y caso todas as X=0
 - Muitas vezes, o α pode ser interpretado como a projeção da reta no eixo Y, uma vez que não encontram-se observações da amostra com todas as variáveis X=0
 - β são os coeficientes angulares, ou seja, a inclinação da reta
 - Na regressão múltipla, os β são interpretados na condição *ceteris paribus*, ou seja, o efeito daquela X sobre Y mantidas todas as demais variáveis constantes
 - Destaca-se que a interpretação dos parâmetros do modelo deve ocorrer sem a extrapolação dos dados, isto é, vale dentro do limite de variação das variáveis

Teste F (ANOVA)

- Avalia a significância geral do modelo de regressão, ou seja, se pelo menos um dos β estimados é estatisticamente diferente de zero

$$F = \frac{\frac{SQR}{(k-1)}}{\frac{SQU}{(n-k)}}$$

SQR: Soma dos Quadrados da Regressão
SQU: Soma dos Quadrados dos Resíduos
k: nº de parâmetros do modelo (inclui α)
n: tamanho da amostra

- H_0 : $\beta_1 = \beta_2 = \dots = \beta_k = 0$
- H_1 : existe pelo menos um $\beta_j \neq 0$
- Normalmente, adota-se o nível de significância de 5% para o teste
 - Se o p-valor do teste $F < 0.05$, rejeita-se H_0

Teste F (ANOVA)

$$SQT = SQR + SQU$$

- **Soma dos Quadrados Totais (SQT):** variação de Y em torno de sua média
- **Soma dos Quadrados da Regressão (SQR):** variação de Y considerando as variáveis X
- **Soma dos Quadrados dos Resíduos (SQU):** variação de Y que não é explicada pelo modelo

$$\underbrace{\sum_{i=1}^n (Y_i - \bar{Y})^2}_{SQT} = \underbrace{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}_{SQR} + \underbrace{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}_{SQU}$$

Teste t

- Avalia a significância individual dos parâmetros estimados

$$t_{\alpha} = \frac{\alpha}{s.e.(\alpha)} \quad t_{\beta_j} = \frac{\beta_j}{s.e.(\beta_j)}$$

- $H_0: \alpha = 0$
- $H_1: \alpha \neq 0$
- $H_0: \beta_j = 0$
- $H_1: \beta_j \neq 0$

- Adota-se o nível de significância de 5% para o teste
 - Se o p-valor do teste $t < 0.05$, rejeita-se H_0
 - **Mesmo que não tenha significância, o α não deve ser removido do modelo!**

Intervalos de confiança

- Para o nível de confiança escolhido, é o intervalo de valores que contém o verdadeiro parâmetro populacional

$$\alpha \pm t \times s.e.(\alpha)$$

$$\beta_j \pm t \times s.e.(\beta_j)$$

- t é o valor crítico bicaudal da distribuição t de Student para o nível de confiança escolhido na análise, com $n - k$ graus de liberdade
 - Normalmente, observa-se o nível de confiança de 95% (nível de significância de 5%)

Coeficiente de explicação (R^2)

- O R^2 apresenta o poder explicativo do modelo, ou seja, o percentual da variabilidade de Y que é explicado pela variação das variáveis X

$$R^2 = \frac{SQR}{SQR + SQU}$$

- O R^2 varia entre 0 e 1: valores mais próximos de 1 indicam maior capacidade preditiva
 - O R^2 não deve ser analisado no sentido de validar ou não o modelo, pois, em muitos campos do conhecimento, é comum não obter valores muito elevados
- R^2 ajustado para comparação entre modelos: $R^2_{ajust} = 1 - \frac{n-1}{n-k} (1 - R^2)$
 - Ajusta-se a quantidade k de parâmetros (incluindo o α) e o tamanho da amostra n

Variáveis explicativas categóricas

- Quando há variáveis X categóricas, é necessário transformá-las em *dummies*
- Dummy*: variável binária (1 ou 0) indicando a presença ou ausência do atributo

ID	Variável A		Variável B		
	Categ. 1	Categ. 2	Categ. 1	Categ. 2	Categ. 3
1	1	0	0	1	0
2	0	1	0	0	1
3	0	1	1	0	0
4	1	0	0	0	1
5	0	1	0	1	0
6	1	0	1	0	0

Na regressão, utiliza-se o procedimento de **n-1 dummies**, ou seja, uma das categorias de cada variável categórica fica como a referência de sua variável no intercepto

Referência

Fávero, Luiz Paulo; Belfiore, Patrícia. (2024). Manual de análise de dados: estatística e machine learning com Excel®, SPSS®, Stata®, R® e Python®. 2 ed. Rio de Janeiro: LTC.

OBRIGADO!