

(1) Introdução ao Curso de Ciência de Dados

João Pinheiro

2001 Engenharia

2001engenharia@gmail.com

13 de fevereiro de 2025

Conteúdo do Curso I

- 1 Objetivos
- 2 Público Alvo
- 3 Pré-requisitos
- 4 Introdução e Conceitos Básicos
 - Conceitos Básicos
 - Introdução do Aprendizado de Máquina
 - Introdução ao Ambiente de programação (Anaconda, Colab, Kaggle)
 - Programação em Python
 - Estatística Básica
 - Funções Matemáticas e Álgebra Linear
- 5 Big Data
 - Ambientes Big Data - Spark, Hive, Hadoop
 - Conceitos de Tratamentos de Dados, ETL e SQL
- 6 Pré-processamento e Visualização de Dados
 - Data Cleaning, Amostragem
 - Introdução à Redução de Dimensionalidade

Conteúdo do Curso II

- Scaling, Encoding
- Standardization, Normalization, Regex

7 Aprendizado Supervisionado

- Introdução e Conceitos Iniciais, Viés-Variância Tradeoff
- Treino, Validação, Teste
- Regressão e Regressão Linear
- Métricas de Regressão
- Classificação
- Métricas de Classificação
- Regressão Linear Múltipla
- Regularizações, Ridge, Lasso, Elastic Net
- Modelos Lineares Generalizados GLM
- Regressão Polinomial
- Regressão Logística
- LDA, QDA

Conteúdo do Curso III

- KNN - vizinhos mais próximos
- Naive Bayes
- Arvore de Decisão, CART
- Support Vector Machine (SVM) e Support Vector Regression (SVR)
- Ensemble
- Random Forest
- Gradient Boosting Machines - XGBoost, CatBoost, LightGBM

8 Avaliação, melhoria e interpretabilidade do modelo

- Métodos Resampling (Validação Cruzada, out-of-sample, Kfold)
- Tuning do Modelo, Otimização de Hiperparâmetro
- Interpretabilidade do modelo

9 Aprendizado Não Supervisionado

- Introdução de Conceitos Iniciais
- Clusterização com K-means
- Agrupamento baseado em densidade (DBSCAN)

Conteúdo do Curso IV

- Mistura gaussiana (GMM)
- Agrupamento Hierárquico
- Redução de dimensionalidade (PCA)

10 Redes Neurais e Aprendizado Profundo (Deep Learning)

- Redes Neurais, Perceptron
- Tensorflow, PyTorch
- Multilayer Perceptron
- Redes Neurais Convolucionais e Arquiteturas
- Regularização, Normalização e Transferência de Aprendizado
- Autoencoders
- Rede neural convolucional (CNN)
- Rede Neural Recorrente (RNN)
- Redes Adversárias Generativas (GANs)
- Sequence to Sequence e Mecanismo de Atenção

11 Visão Computacional

Conteúdo do Curso V

12 Processamento de Linguagem Natural (NLP)

13 Grande Modelo de Linguagem (LLMs)

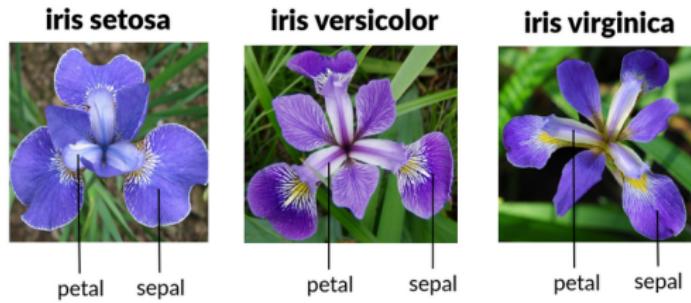
Objetivos

- Entender um problema a ser resolvido e transforma-lo em um problema de aprendizado de máquina.
- Cada aula teórica será acompanhada de uma prática em programação.
- O curso será transformado em um **livro no Github**, podendo ser consultado a qualquer momento e atualizado por vocês.
- Além disso teremos grandes **projetos** em cada um dos macro temas, possibilitando um maior aprendizado e esses projetos podem ser utilizados como portfólio de vocês.

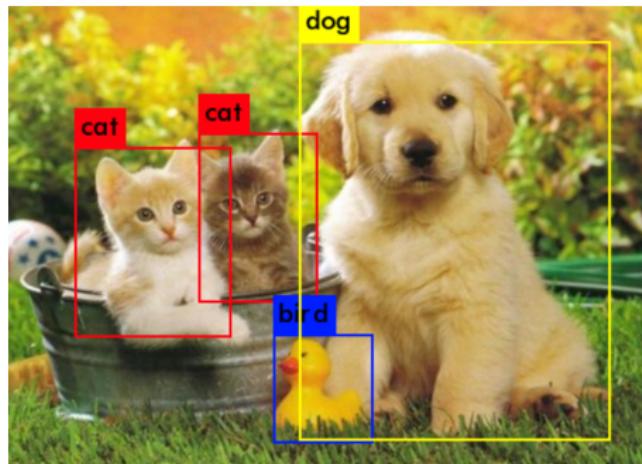
Predição de Preço de Imóveis



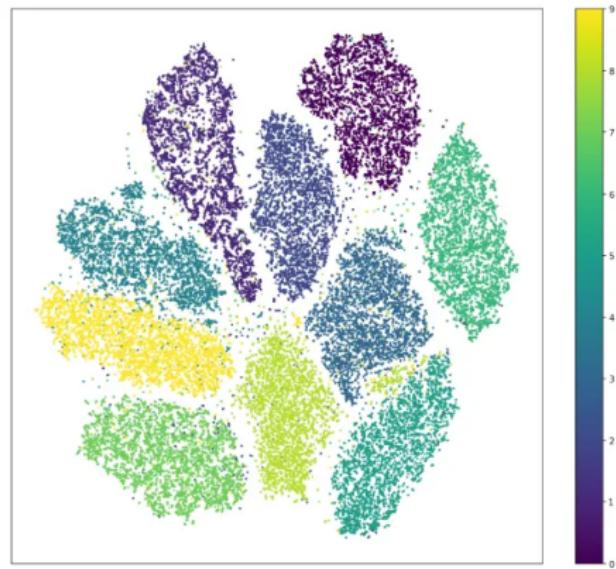
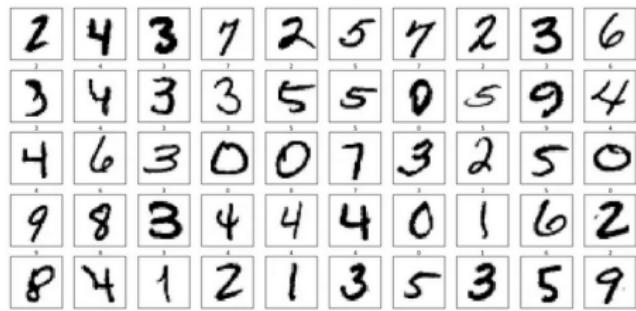
Classificação de Flores



Detecção e Classificação de Objetos



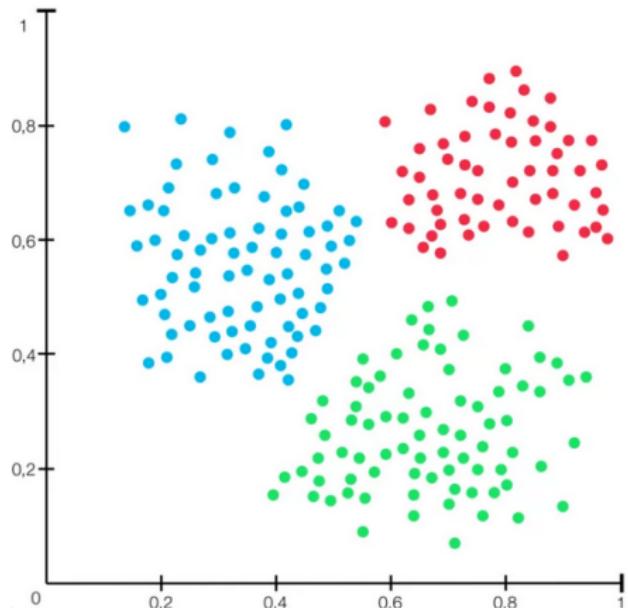
Reconhecimento de Dígitos



Clusterização de Músicas no Spotify



Clusterização e Segmentação de Clientes



Análise de Sentimentos



Sentiment Analysis



IA jogando jogos



State:247

R:	:	:	G
:			
:			
:			
Y	:	:	B

State:99

R:	:	:	G
:			
:			
:			
Y	:	:	B

State:475

R:	:	:	G
:			
:			
:			
Y	:	:	B

Projetos

Teremos muitos projetos clássicos e outros que são pouco abordado, alguns projetos:

- Classificação de Flores (Iris Flowers Classification)
- Predição da Qualidade do Vinho (Wine Quality Dataset).
- Reconhecimento de Dígitos (AMNIST Handwritten Digit)
- Recomendação de Filmes (Movie Recommender System Dataset)
- Predição de Preço de Imoveis (Boston House Pricing)
- Analise de Sentimentos utilizando o dataset do Twitter
- Predição de Churn.
- Segmentação/Clusterização de Clientes.
- Atrasos e voos cancelados.
- Detecção de Imagens e Objetos
- Reconhecimento de Emoções através da fala.
- Predição de Ações com Séries Temporais.

Tecnologias

Teremos muitos projetos clássicos e outros que são pouco abordado

- Python, Pandas, NumPy, SciPy, Spark, SQL
- Scikit-learn, Seaborn, matplotlib, Optuna, SHAP
- OpenCV, PyTorch, TensorFlow
- AWS, Cloud

Público Alvo

- Cientista de dados, engenheiros
- Ciências exatas
- Estatísticos
- Migrar para carreira de dados

Obs: Recomendo um Mestrado ou MBA, isso é muito valorizado.

Pré-requisitos

- Lógica de programação (Python principalmente)
- Cálculo, álgebra linear, matrizes
- Estatística, probabilidade, distribuições
- Inglês Técnico

Obs: será um curso com uma matemática 'densa', mas essa que é a parte divertida :)

- **Dados:** Matéria-prima da Ciência de Dados.

- **Dados Estruturados:**

- Organizados em formato tabular (linhas e colunas).
 - Exemplos: Bancos de dados relacionais, planilhas Excel.
 - Fáceis de processar e analisar com ferramentas tradicionais.

- **Dados Não Estruturados:**

- Não possuem um formato definido.
 - Exemplos: Textos, imagens, vídeos, áudios.
 - Requerem técnicas avançadas para análise (e.g., NLP, visão computacional).

- **Dados Semi Estruturados:**

- Possuem alguma organização, mas não são totalmente estruturados.
 - Exemplos: JSON, XML, logs de sistemas.
 - Flexíveis, mas podem exigir pré-processamento para análise.

- **Análise de Dados:**

- Processo de inspeção, limpeza e transformação de dados.
- Objetivo: Extrair insights e suportar a tomada de decisões.
- Estatística descritiva, visualização de dados, análise exploratória.

- **Machine Learning:**

- Algoritmos que aprendem padrões a partir de dados.
- Aplicações: Classificação, regressão, clustering, recomendação.
- Exemplos: Redes neurais, árvores de decisão, SVM.

- **Big Data:**

- Dados em grande volume, variedade e velocidade.
- Desafios: Armazenamento, processamento e análise em tempo real.
- Tecnologias: Hadoop, Spark, NoSQL (e.g., MongoDB, Cassandra).

- **Cientista de Dados:**

- Responsável por extrair insights e construir modelos preditivos.
- Habilidades: Estatística, machine learning, programação (Python, R).

- **Engenheiro de Dados:**

- Foco na coleta, armazenamento e processamento de dados em escala.
- Habilidades: Python, SQL, Spark, Airflow, Git, Cloud.
- Responsável por criar e dar manutenção nos pipelines de dados.

- **Analista de Dados:**

- Responsável por transformar dados em insights para tomada de decisões.
- Habilidades: SQL, Excel, ferramentas de visualização (e.g., Power BI, Tableau).

- **Engenheiro de Machine Learning:**

- Responsável por implementar, otimizar e escalar modelos de ML.
- Habilidades: Python, frameworks (TensorFlow, PyTorch), pipelines de dados, Cloud, Docker

Função do time de Dados

A função principal de um time de dados é resolver problemas e entregar valor ao time de negócios. Por meio da utilização de dados, análises e tecnologias para gerar insights e soluções que impulsionem decisões estratégicas e operacionais

Definição de Aprendizado de Máquina

"Campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados."

A. L. Samuel

**Some Studies in Machine Learning
Using the Game of Checkers**



Figura: Arthur Samuel (1959)

Definição de Aprendizado de Máquina

"Diz-se que um programa de computador aprende pela experiência \mathcal{E} , com respeito a algum tipo de tarefa, \mathcal{T} e performance, \mathcal{P} , se sua performance nas tarefas, melhoraram com a experiência \mathcal{E} ".

Experiência → Dados

Performance → Métrica

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



Figura: Tom M. Mitchell (1998)

Teste de Turing

Em 1950 Alan Turing propôs o teste de Turing que consiste em responder a pergunta.

"Será que uma máquina consegue pensar?"

"Um computador passa no teste se um interrogador humano, depois de fazer algumas perguntas por escrito, não consegue distinguir se as respostas foram feitas por uma pessoa ou por um computador." .



Figura: Alan Turing (1950)

O que o computador precisa para passar no teste?

- **Processamento de Linguagem Natural:** para se comunicar com sucesso em uma linguagem humana.
- **Reconhecimento de Fala:** para reconhecimento de fala.
- **Representação de Conhecimento:** para armazenar o que sabe ou ouve.
- **Raciocínio Automatizado:** para responder perguntas e tirar novas conclusões.
- **Aprendizado de Máquina:** para se adaptar a novas circunstâncias e detectar e extrapolar padrões
- **Visão Computacional:** para reconhecer imagens com precisão.
- **Robótica:** para manipular objetos e se locomover.

Comparativo Computador × Cérebro Humano

	Supercomputer	Personal Computer	Human Brain
Computational units	10^6 GPUs + CPUs	8 CPU cores	10^6 columns
	10^{15} transistors	10^{10} transistors	10^{11} neurons
Storage units	10^{16} bytes RAM	10^{10} bytes RAM	10^{11} neurons
	10^{17} bytes disk	10^{12} bytes disk	10^{14} synapses
Cycle time	10^{-9} sec	10^{-9} sec	10^{-3} sec
Operations/sec	10^{18}	10^{10}	10^{17}

Figura: Comparativo de Computadores e o Cérebro Humano (Stuart J. Russell and Peter Norvig).

Visão Geral dos Problemas de Machine Learning

Objetivo: achar uma função que descrevam algum comportamento dado um conjunto de dados $(\mathbf{x}^{(i)}, y^{(i)})$, em que $y^{(i)} \in \mathbb{R}$ é a saída desejada

Dado um **conjunto de treino** com N pares de entrada-saída

$$\{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \dots, (\mathbf{x}^{(N)}, y^{(N)})\}$$

$\underbrace{\text{tamanho, cidade}}_{\text{input/features}} \rightarrow \underbrace{\text{valor}}_{\text{target/saída/variável resposta}}$

$$\text{input} \rightarrow \mathbf{x}^{(i)} \in \mathbb{R}$$

$$\text{output} \rightarrow y^{(i)} \in \mathbb{R}$$

$$f'(\mathbf{x}) \rightarrow y$$

Visão Geral dos Problemas de Machine Learning



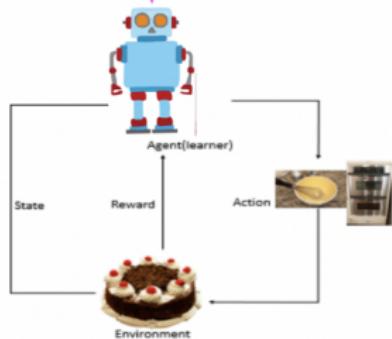
Machine Learning



$$V = \frac{4}{3}\pi r^3$$



Supervised Learning



Reinforcement Learning

Diferentes tipos de Aprendizagem

Aprendizado Supervisionado

Modelo aprende a partir de dados rotulados, ou seja as saídas são bem definidas com uma variável resposta é como se tivesse um 'professor'. Exemplos: prever o preço de imóvel, classificar imagens

Aprendizado Não Supervisionado

Envolve uma exploração intrínsecas nos dados, como agrupamentos ou distribuições, aqui não temos uma variável resposta.
Exemplos: agrupamento de tipos musicais, segmentação de clientes

Aprendizado por Reforço

Já o aprendizado por reforço é uma abordagem na qual um agente 'aprende' a realizar ações em um ambiente por meio de um estímulo (recompensa) Exemplos: robótica, jogos

Aprendizado Supervisionado: Regressão × Classificação

Regressão: a variável resposta é contínua e $y \in \mathbb{R}$

Exemplos: predição de preço de imóveis

Classificação: a variável resposta é discreta, sendo dividida em 'classes'.

Geralmente, abordamos a classificação binária (0 ou 1)

Exemplos: classificação de flores, classificação de objetos em visão computacional, reconhecimento de dígitos

Métodos e ferramentas para resolver problema



Referencias



Mitchell T. M, (1997)

Machine Learning



Trevor Hastie and Robert Tibshirani and Jerome Friedman (2009)

The Elements of Statistical Learning



Stuart J. Russell and Peter Norvig (2021)

Artificial Intelligence: A Modern Approach, Global Edition



Ian Goodfellow, (2016)

Deep Learning



Sutton, Richard S. and Barto, Andrew G, (2018)

Reinforcement Learning: An Introduction