INSTITUTO FEDERAL
Sergipe

Scikit-Learn
Classificação de Dados

João Paulo Dias de Almeida
joao.almeida@academico.ifs.edu.br

Instituto Federal de Sergipe

O que vamos aprender hoje?

- Identificar fontes com grandes volumes de dados
- Entender aplicações de algoritmos de aprendizado de máquina
- Descobrir os passos necessários para projetar um sistema de aprendizado
- Utilizar o scikit-learn para criar um algoritmo de aprendizado
- Aprender formas de avaliar um sistema de aprendizado

Qual o volume de dados?

Qual o volume de dados?

- Estima-se que foram gerados 40 trilhões de gigabytes, entre 2005 e 2020
  - Equivale a 40 000 exabytes
    - 1 Exabyte (EB) = 1 024 PB = 1 048 576 TB
  - Essa quantidade corresponde a 5.2 gigabytes por pessoa
- Computadores estão por toda parte e permitem armazenar muitas coisas que, inclusive, poderiam ser enviadas para o lixo
- Mecanismos de armazenamento local e remoto parecem ilimitados

"Preciso mesmo deletar esse arquivo agora?
Vou guardar, depois tomo uma decisão..."

Volume de dados

- Dados são produzidos, coletados e armazenados

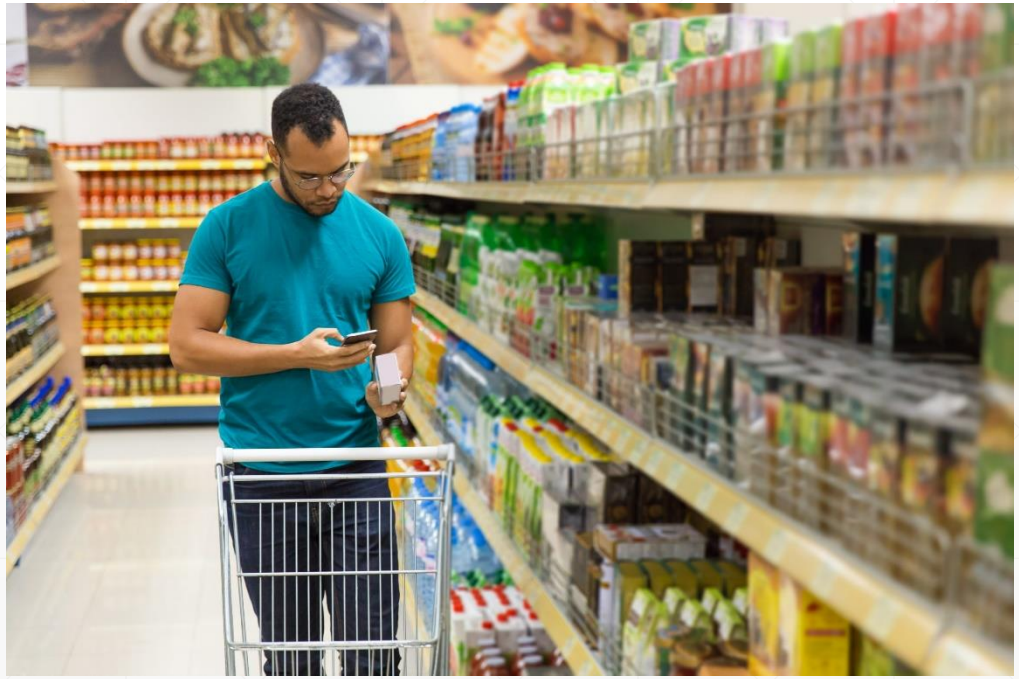
Introdução

- Atualmente, Inteligência Artificial (IA) é utilizado em aplicações aplicadas a vários contextos:
  - Marketing
  - Banco
  - Finanças
  - Agricultura
  - Saúde
  - Jogos
  - Exploração espacial
  - Chatbots

- Inclusive, The Online Privacy Foundation, patrocinou uma competição para verificar a possibilidade de classificar alguém como psicopata a partir de seu uso do Twitter

Introdução

- Como extrair informações úteis a partir de dados armazenados?



Estudo de caso

- **Cenário:**
  - Homem adulto comprando fraldas após o trabalho à caminho de casa. Sendo sexta-feira a noite, ele também pega um pack de cervejas
- **Análise:**
  - Diz a lenda, que um estudo foi realizado e encontrou que homens entre 30-40 anos, realizando compras entre 17–19 horas nas sextas, tem o perfil de comprar cervejas junto com fraldas
- **Resultado:**
  - Rede de supermercado aumentou em 35% a venda dos dois produtos ao posicioná-los juntos na loja

Introdução

- Estratégias atuais de mercado:
  - Identificar vendas de produtos com maior similaridade
  - Recomendar novos produtos a partir de compras recentes
  - Indicar estratégias de marketing de acordo com o perfil dos clientes

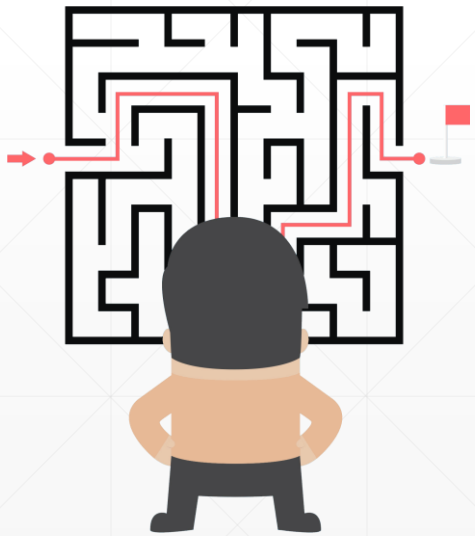
- As estratégias acima são facilmente realizadas por humanos
  - Apenas em pequenos conjuntos de dados!

- Solução:

  Técnicas de Inteligência Artificial → Aprendizado de Máquina

Aprendizado de Máquina

- Técnicas Autônomas
  - Baixa intervenção humana
  - Criar hipótese a partir de experiências passadas
  - Capaz de resolver problemas



IA e AM

- **Aprendizado de máquina:** É o processo de indução de hipóteses (aproximação de funções) a partir de experiências passadas

"Aprendizado de máquina é a capacidade de melhorar o desempenho na realização de alguma tarefa por meio de experiência." (Mitchell, 1997)

Aplicações

- Fertilização *in Vitro*
  - Coleta de óvulos
  - Fertilização
  - Seleção e transferência dos melhores
  - Embriões para ovário

Como selecionar embriões?

Aplicações

- Fertilização *in Vitro*
  - Existem cerca de 60 características para cada embrião. Como um especialista poderia avaliá-las simultaneamente para cada embrião?
  - Na Inglaterra, AM é utilizado para analisar essas características e dados históricos neste processo de **tomada de decisão**

Aprendizado de Máquina

- Desde que computadores foram criados, questiona-se se eles serão capazes de aprender
  - Imagine as vantagens:
    - Auxiliar no diagnóstico e tratamento de doenças com base em registros médicos
    - Otimizar custos residenciais com base no perfil dos moradores

Aprendizado de Máquina

- Computadores ainda não são capazes de aprender como as pessoas, mas alguns algoritmos já provaram ser muito eficientes para certos tipos de tarefas de aprendizado

Aplicações

- Reconhecimento facial

Aplicações

- Detectar grupos/padrões

Projetando Sistemas de Aprendizado

- Escolha do conjunto de treinamento
  - A experiência de treinamento afeta diretamente a qualidade do aprendizado
- Como determinar se um movimento no jogo de damas levou à vitória ou à derrota?
  - **Experiência direta:** exemplos de jogadas corretas
  - **Experiência indireta:** sequências aleatórias de jogadas e o resultado final
- Uma ótima jogada garante a vitória?

Projetando Sistemas de Aprendizado

- Representatividade

O aprendizado é mais confiável quando o conjunto de treinamento segue uma distribuição representativa do sistema real

Projetando Sistemas de Aprendizado

- O conjunto de treinamento pode ser formado por exemplos obtidos com experiência passada
  - O aprendizado com base em experiência passada utiliza o princípio da indução
    - Indução → conclusões genéricas a partir de um conjunto de exemplos

Indução de Hipótese

- O conjunto de treinamento contém exemplos (dados, padrão, registro, objetos)
- Cada exemplo é formado por atributos (campos, variáveis, características)
- Um atributo muito importante é o de saída
  - Rótulo ou alvo
  - Label ou target
- Quando não fornecido, esse atributo pode ser estimado dos demais (entrada)

Indução de Hipótese

- Objetivo técnicas de AM: aprender um modelo (hipótese), usando o conjunto de treinamento, para relacionar atributos de entrada a valores do atributo de saída

- Análise de dados: remoção de ruídos, inconsistências, dados ausentes, redundância, …

Indução de Hipótese

- Exemplo de diagnóstico de problema de coração

Matriz Atributo x Valor

Nome	CPF	Naturalidade	Fumante	Álcool	Time	Diag.
José	111111	Salvador	Sim	Sim	Bahia	Não
Pedro	222222	F. de Santana	Não	Não	Vitória	Sim

Indução de Hipótese

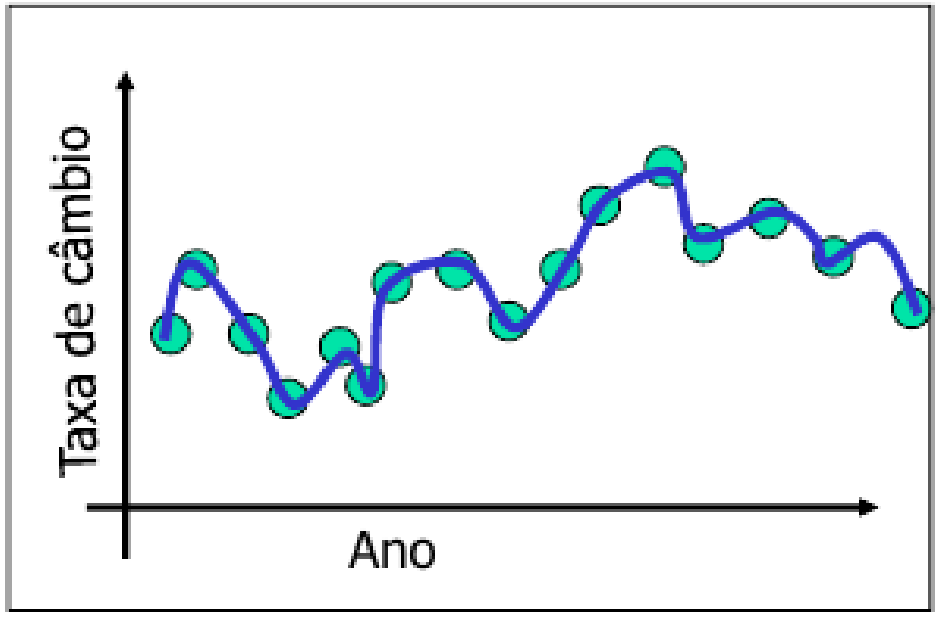
- Objetivo de técnicas de AM é induzir uma hipótese que seja capaz de fazer diagnósticos corretos para novos pacientes

Nome	CPF	Naturalidade	Fumante	Álcool	Time	Diag.
José	111111	Salvador	Sim	Sim	Bahia	Não
Pedro	222222	F. de Santana	Não	Não	Vitória	Sim
Marcos	333333	Recife	Sim	Não	Sport	???

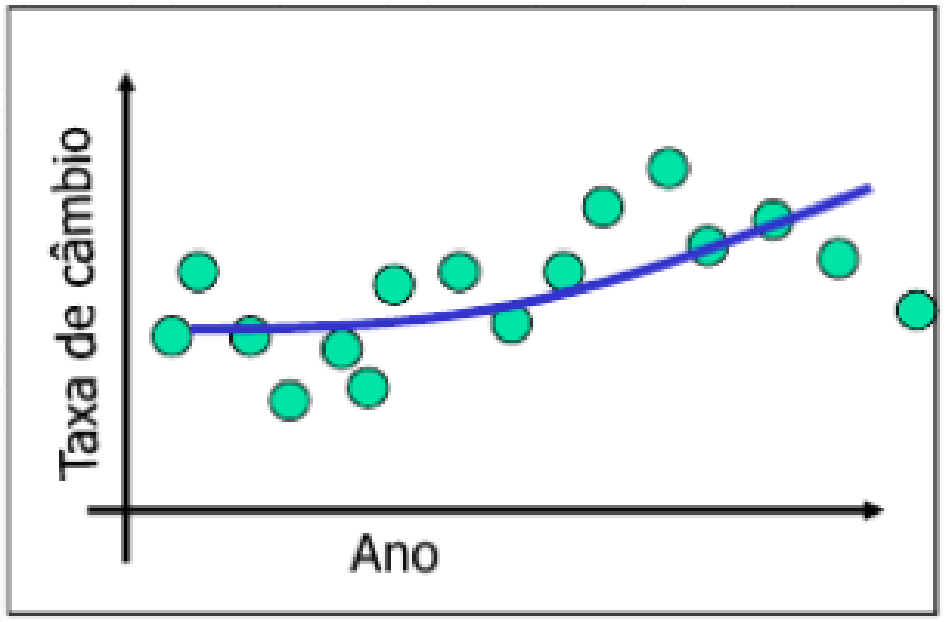
Generalização

Cuidados com a indução

Overfitting



Underfitting



Indução de Hipótese

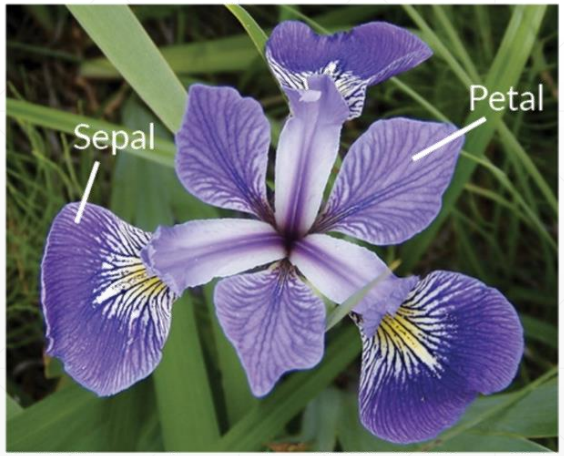
Aprender em AM significa procurar uma hipótese, no espaço de todas as possíveis hipóteses, capaz de descrever relações entre objetos do conjunto de treinamento

Kahoot

Prática: Classificação de flores

Iris dataset

- É um famoso dataset que contém características de três espécies de Iris
  - Essas características foram utilizadas pelo biólogo Ronald Fisher em 1936 para desenvolver um modelo linear capaz de discriminar as três espécies



Iris Versicolor

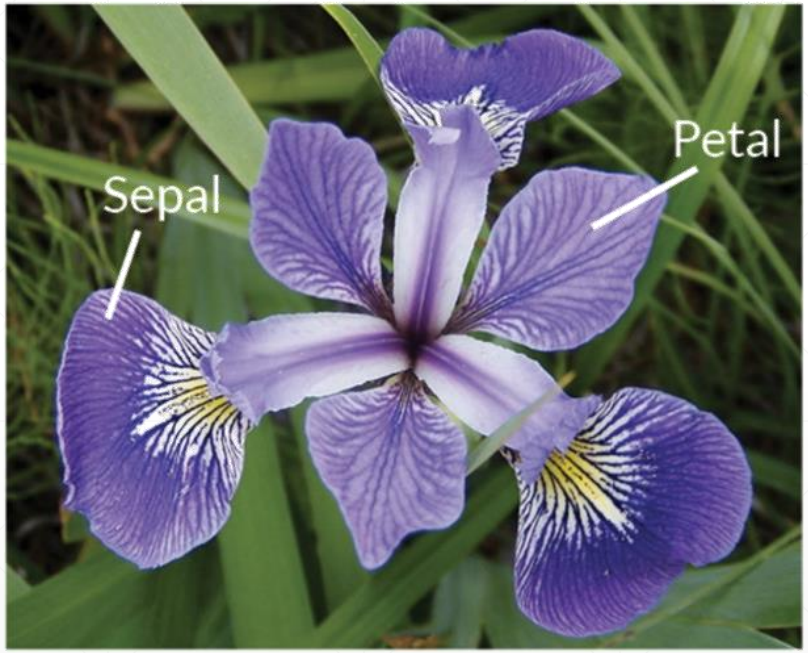
Iris Setosa

Iris Virginica

Iris dataset

- O dataset contém 50 amostras de cada espécie da Iris
  - Total = 150 amostras
  - Cada linha do dataset corresponde a uma amostra
- Para cada amostra foram coletadas características (features)
  - Comprimento e largura das sépalas e das pétalas em centímetros
  - Array 150x4

Iris dataset

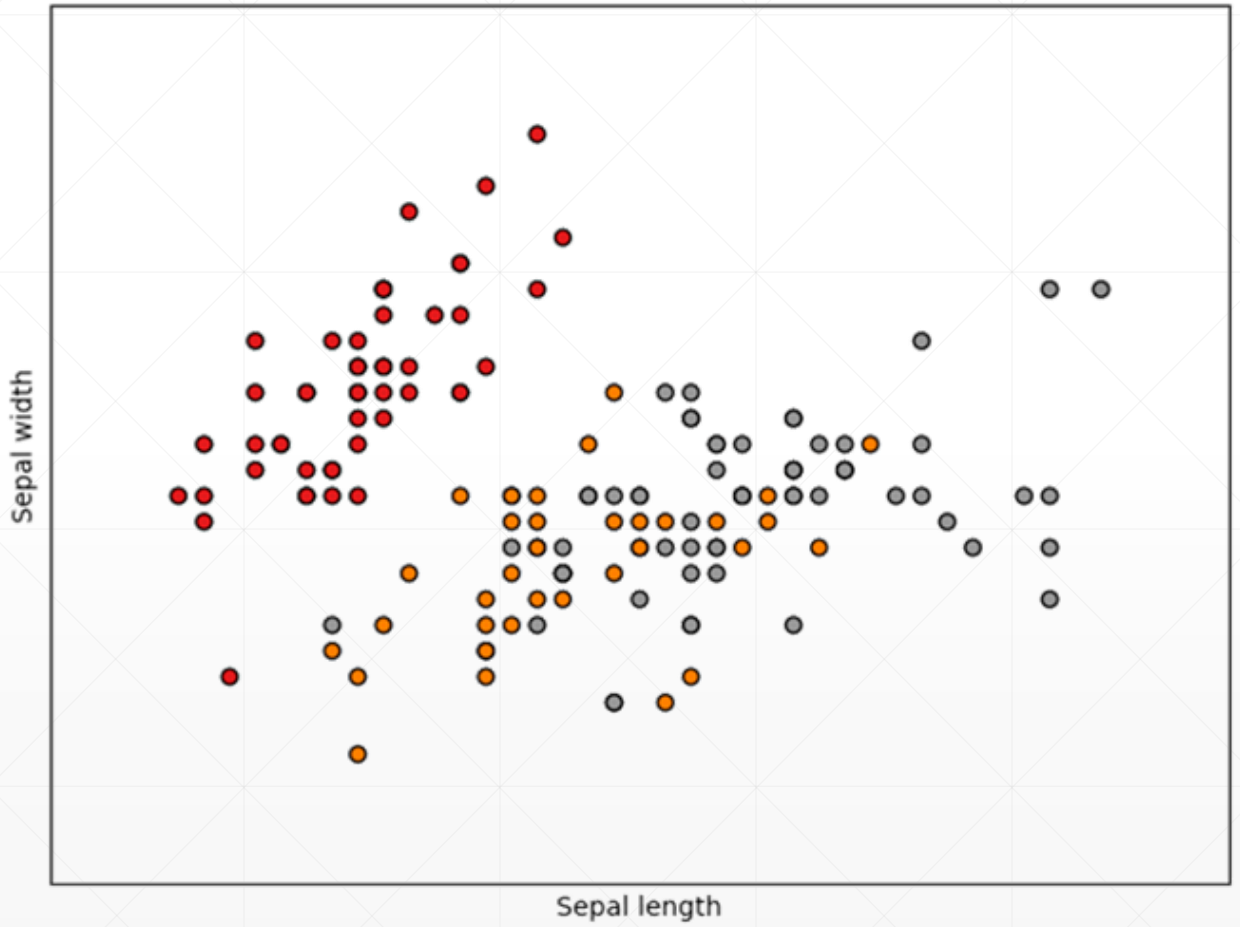


Iris Versicolor

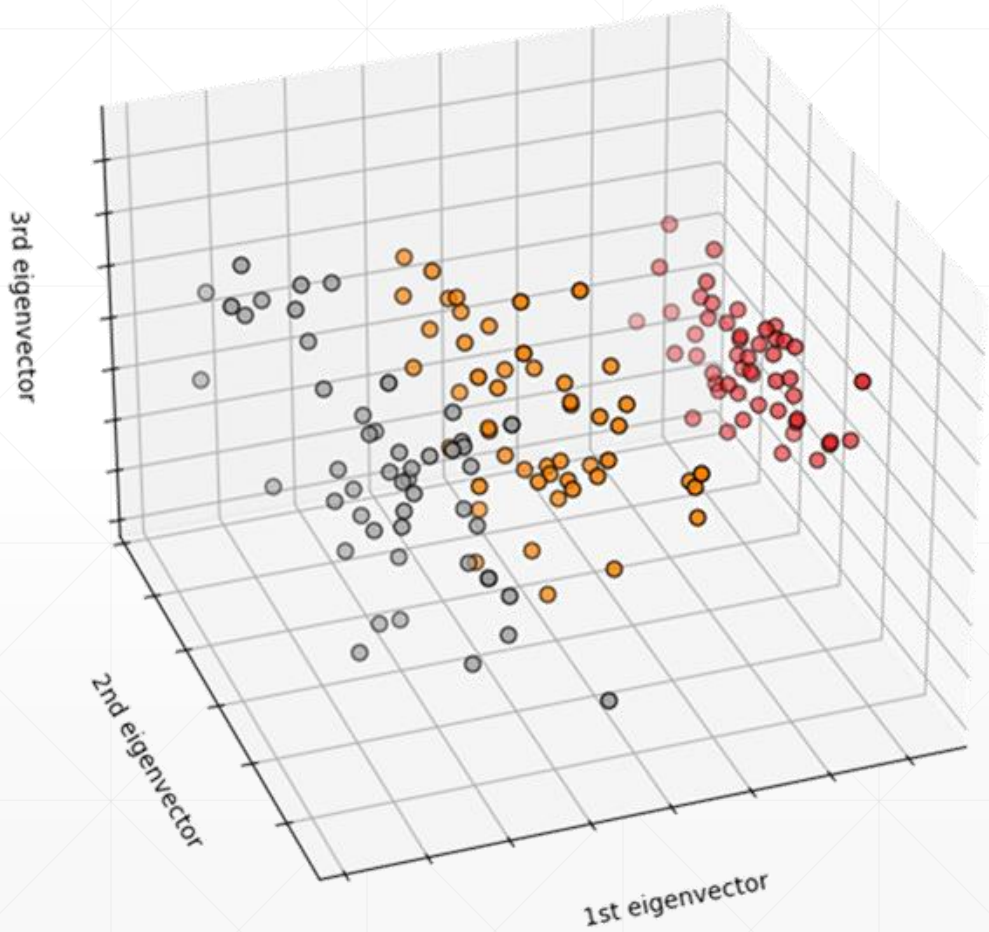
Iris Setosa

Iris Virginica

1. Analisando os dados

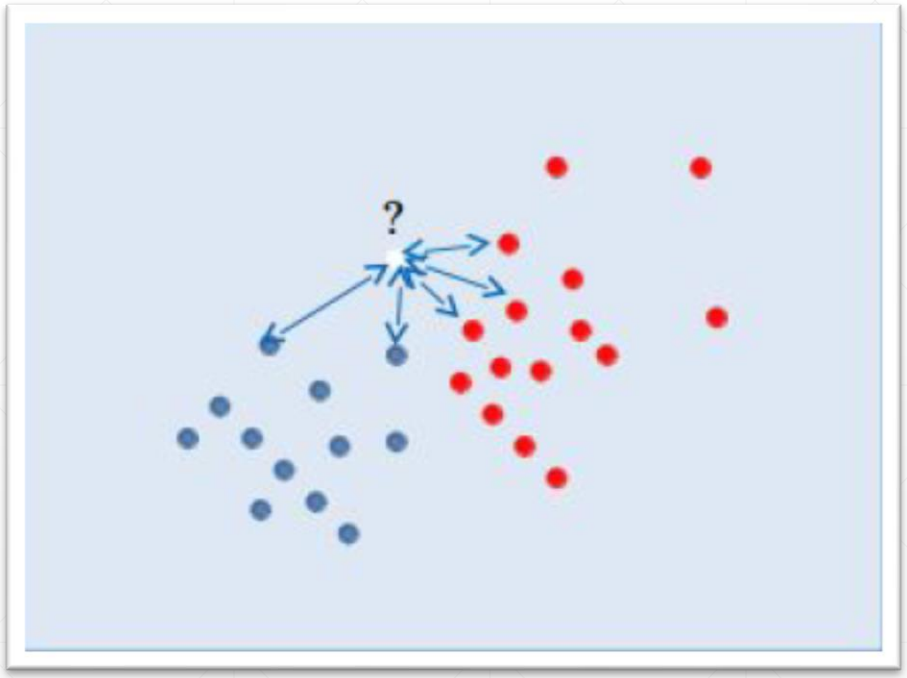


1. Analisando os dados



Aprendizado Baseado em Instâncias

- Como vamos ensinar ao computador o padrão que deve ser aprendido para identificar cada espécie da Iris?
  - Vamos utilizar um algoritmo de classificação
  - Este algoritmo considera a proximidade entre os dados para realizar predições



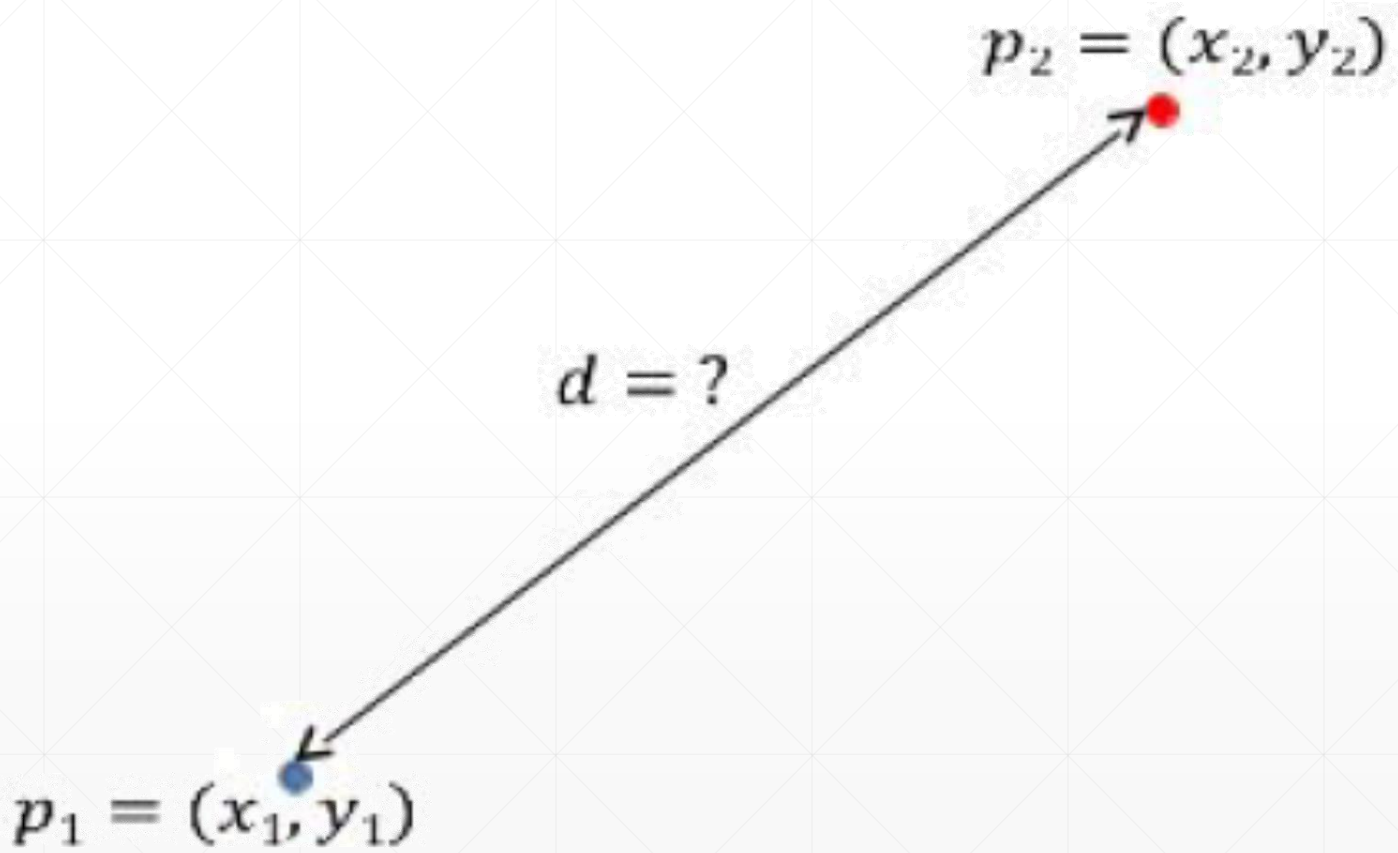
K-Nearest Neighbors (KNN)

- Algoritmo dos vizinhos mais próximos
- Objetos relacionados ao mesmo conceito são semelhantes entre si
- Algoritmo preguiçoso (lazy): não gera modelo, comparação com objetos usados no treinamento
- Variações podem ser definidas através do número de vizinhos (k)

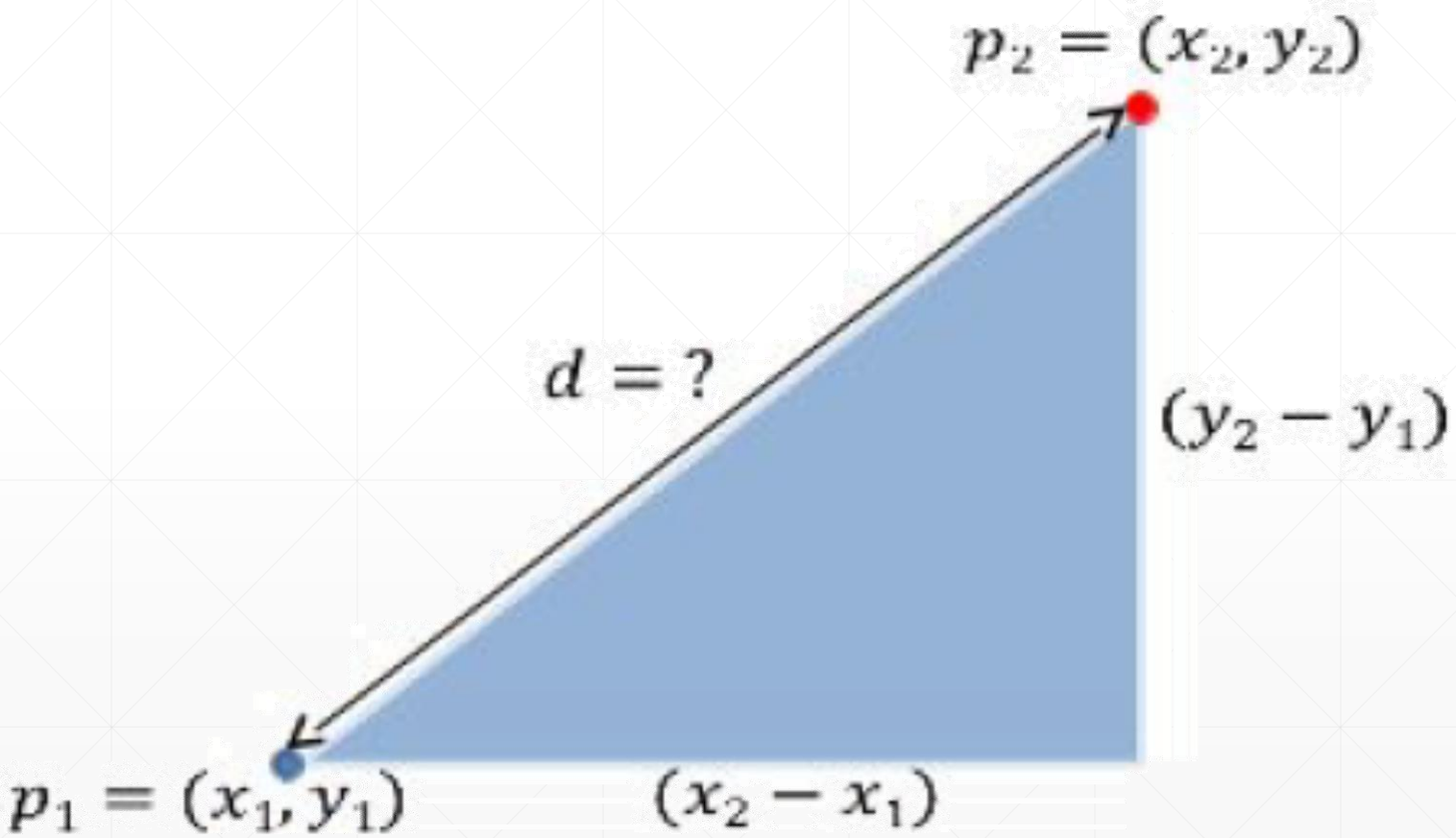
Algoritmo 1-NN

- Espaço de entrada
  - Cada amostra representa um ponto em um espaço definido pelos atributos
- Calcular as distâncias entre dois pontos
  - Métrica mais comum: distância euclidiana

Distância Euclidiana

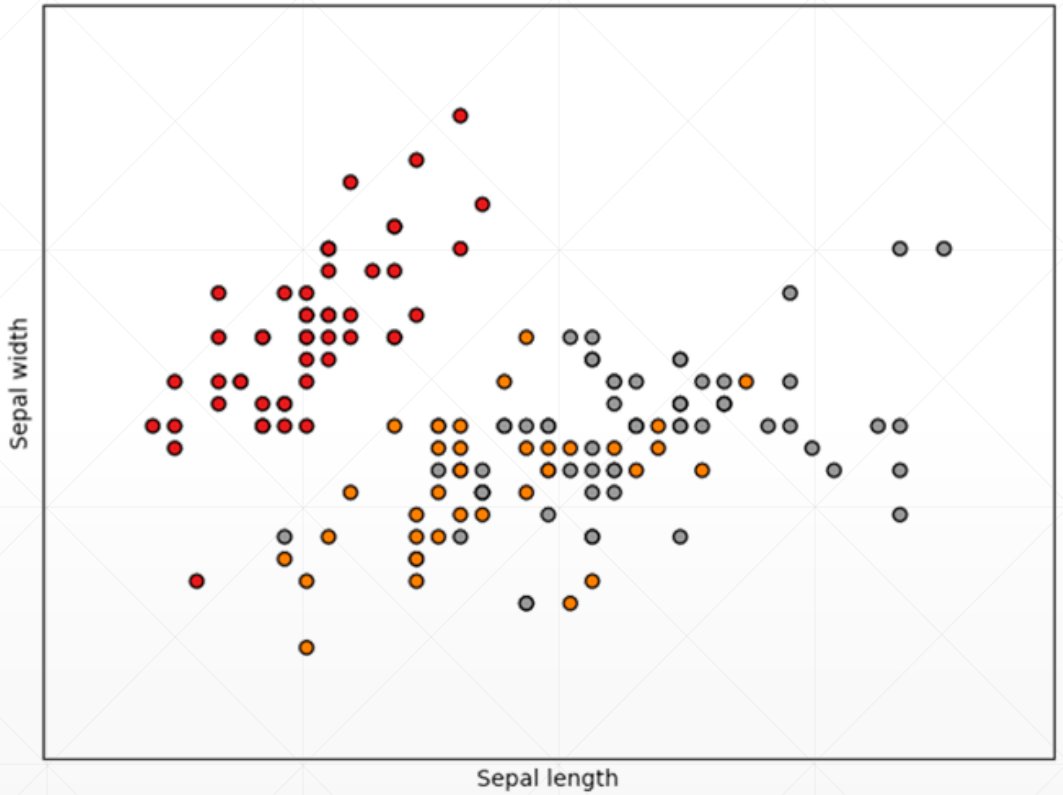


Distância Euclidiana



Aprendizado Baseado em Instâncias

- Dados correspondem a pontos no espaço d-dimensional, ou seja, seus atributos são numéricos
- Medidas de distância são afetadas pela escala
  - Solução: normalizar os atributos

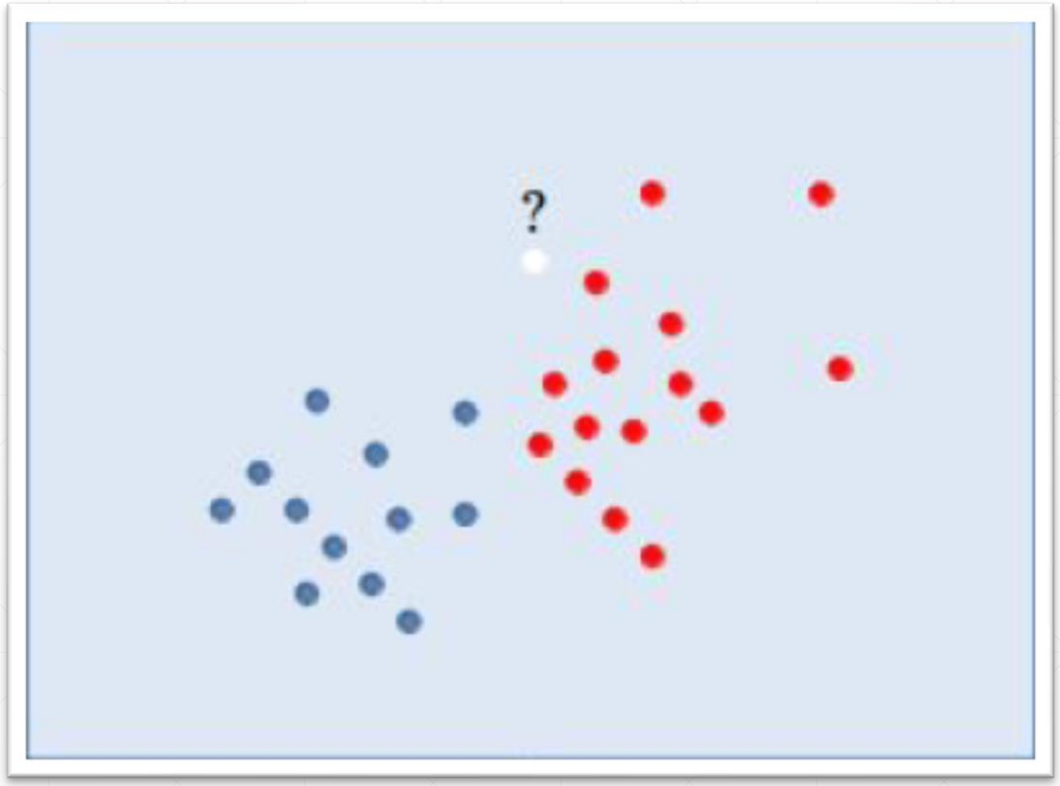


Algoritmo k-NN

- **Fase de treinamento:** memorização dos exemplos rotulados do conjunto de treinamento
- **Classificação:** cálculo da distância entre o vetor de valores do exemplo não rotulado e cada exemplo armazenado na memória
- **Resultado:** o rótulo do novo exemplo será o mesmo rótulo do vizinho mais próximo

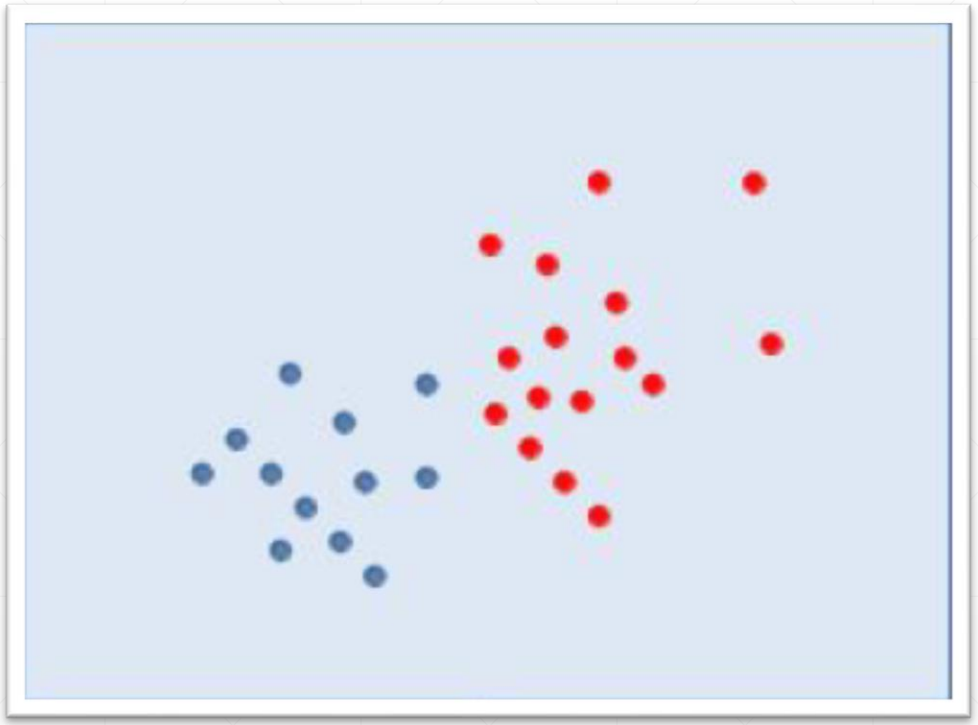
Algoritmo 1-NN

- Qual o rótulo do novo exemplo?



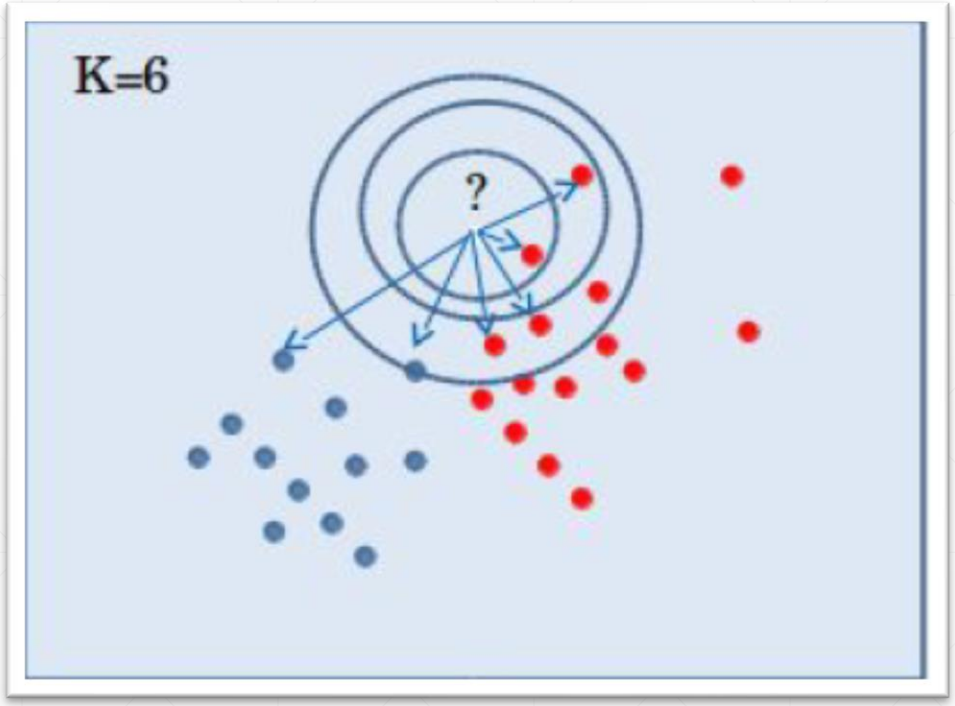
Algoritmo 1-NN

- Qual o rótulo do novo exemplo?



Algoritmo K-NN

- Considerar os k objetos do conjunto de treinamento mais próximos do ponto de teste



Algoritmo K-NN

- Qual o melhor valor de k?
  - O problema pode fornecer indícios do valor ideal
  - Validação
  - Geralmente, um valor pequeno e ímpar
  - Valores pares podem gerar empates

Vamos ao Código!

Repositório do Código

https://github.com/JoaoAlmeida/KNN-workshop

Avaliação do Algoritmo

- Matriz de confusão

		Valor Predito	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Classificação de vinhos

Faça você mesmo!

- Agora vamos tentar um novo dataset
  - Wine recognition dataset
- 178 instâncias
  - Três classes
  - Cada instância possui 13 atributos e uma classe
- As instâncias são resultantes de análises químicas feitas em amostras de vinhos de três produtores de regiões diferentes da Itália

Faça você mesmo!

- A partir da composição química de cada amostra de vinho é possível treinar o classificador para identificar quem é o produtor do vinho
- **Atributos disponíveis:**
  - Álcool
  - Ácido málico
  - Cinzas
  - Alcalinidade da cinza
  - Magnésio
  - Total de fenóis
  - Flavonoides
  - Fenóis não flavonoides
  - Proantocianinas
  - Intensidade da cor
  - Matiz
  - OD280/OD315 da diluição dos vinhos
  - Prolina

Resumo

Resumo

- É possível extrair informações úteis de grandes volumes de dados que a principio não fazem sentido
- Algoritmos de Aprendizado de Máquina são úteis para reconhecer padrões nos dados
  - Tarefas de classificação, predição, e regressão

Resumo

- Para utilizar um algoritmo de aprendizado, é necessário:
  - Realizar análise dos dados (limpeza, normalização, pré-processamento)
  - Realizar o treinamento utilizando o algoritmo adequado
  - Avaliar o algoritmo treinado
  - Analisar os resultados

Referências

- Rios, R., Rios, T. **Aprendizado de Máquina: Introdução**, 2017, 62 slides
- Rios, R., Rios, T. **Aprendizado Baseado em Instância**, 2017, 37 slides
- Johari, A. **AI Applications: Top 10 Real World Artificial Intelligence Applications**. Disponível em: https://www.edureka.co/blog/artificial-intelligence-applications/
- Gantz J., Reinsel D. **The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east**. IDC iView: IDC Analyze the future, 2012

Dúvidas?