Machine Learning 🤖

Prof. Me. Alexandre Henrick

Sistemas de Informação - 8º P

Introdução

- Até o momento vimos exemplos de **IAs projetadas por pessoas**
- E que atuam para resolver problemas usando algum "comportamento" programado pelo programador
- A qualidade dessas IAs está totalmente atrelada a como o programador projeta esses "comportamentos"

Introdução

- Visando trazer mais **autonomia** para essas IAs, foram criados os agentes baseados em **experiência**
- Essa é uma classe de IA que consegue aprender com a experiência e através de **observações**
- Essa característica permite com que essas IAs consigam **melhorar seu desempenho através da experiência**
- A medida que essas IA recebem mais **observações, e observações de qualidade**, consegue interpretar e melhorar sua atuação

Introdução

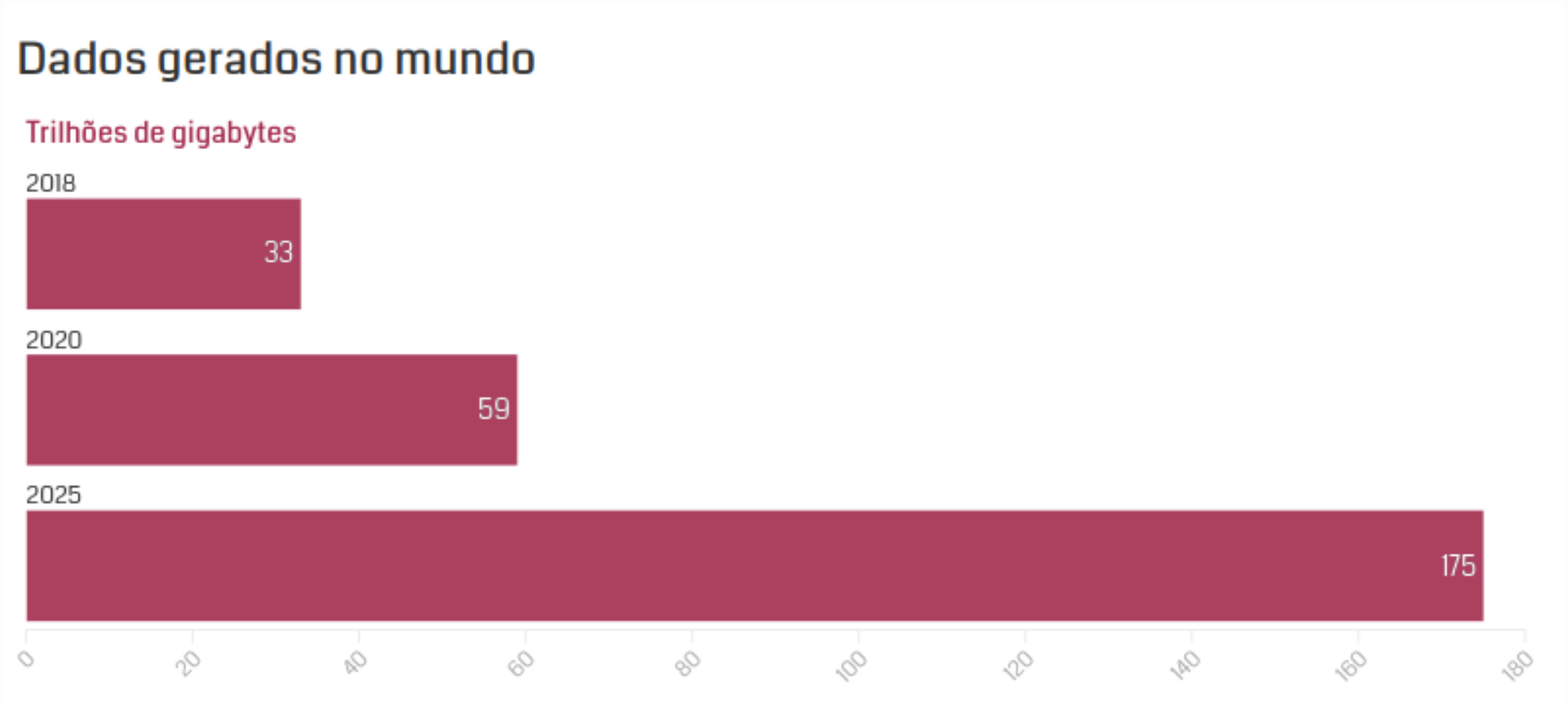
- Essas observações que os **modelos (IAs/algoritmos de machine learning)** recebem são os **dados que representam o contexto do problema que desejamos resolver**
- Portanto, os **dados são parte fundamental em todo o processo de implementação desses modelos**

A importância dos dados

- Vivemos na **era dos dados**. Nunca se produziu tanta informação como atualmente
- Isso se deve principalmente a democratização de ferramentas na Internet e redes sociais
- Qualquer tipo de interação de qualquer usuário em qualquer ferramenta online gera dados

A importância dos dados

- Segundo relatório Data Age 2025, do IDC, em 2018 geramos cerca de **33 zettabytes (33 trilhões de gigabytes)**. Em 2020 foi para **59 ZB**


Dados gerados no mundo
Trilhões de gigabytes
2018
33
2020
59
2025
175
0
20
40
60
80
100
120
140
160
180

A importância dos dados

- Mas o que isso significa?
- Os dados se transformaram no **ativo mais valioso do mundo!**
- Empresas competem para criar as **melhores soluções que consigam extrair padrões desses dados**
- Isto é, soluções que consigam lidar com grandes massas de dados

Estatística ou Machine Learning?

- Os métodos mais tradicionais de estatísticas conseguem extrair padrões valiosos em grandes volumes de dados
- Mas de qualquer maneira, o volume pode se tornar um gargalo
- Outro ponto, e o mais importante, **é a necessidade de automação** de processos
- Além de procurar padrões em dados e entregar insights, precisamos que essas soluções sejam **autônomas**

Estatística ou Machine Learning?

- Para atender essas necessidades, foram desenvolvidas técnicas avançadas de estatística para capturar e aprender padrões em dados
- O conjunto dessas técnicas foi chamado de **Machine Learning**
- É um ramo da IA, que também é conhecido como **Mineração de Dados**
- Leva esse nome pelo fato de utilizar algoritmos que **"aprendem" através de observações (dados)**

Machine Learning

- Em Machine Learning (ML) existem **dois tipos principais de técnicas**
  - Classificação;
  - Clusterização (Agrupamento)

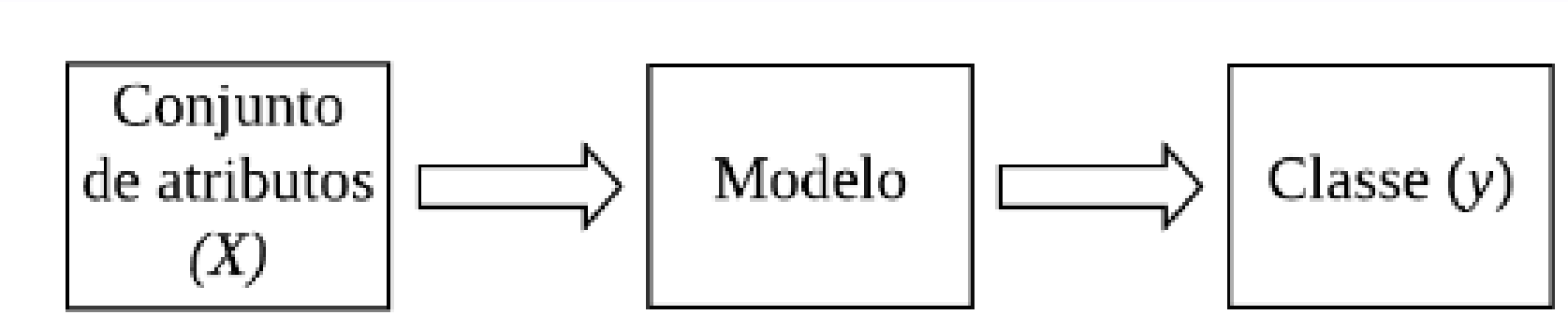
Classificação

- A classificação é uma das tarefas onde usamos algoritmos de ML para **rotular observações**
- Dado uma observação, a qual classe ela pertence?
- Ex: Dado um conjunto de informação sobre características de uma flor, qual é a sua espécie?

Registro	*sepal length*	*sepal width*	*petal length*	*petal width*	Classe
1	5,1	3,5	1,4	0,2	Iris-setosa
2	5,0	3,3	1,4	0,2	Iris-setosa
3	7,0	3,2	4,7	1,4	Iris-versicolor
4	5,7	2,8	4,1	1,3	Iris-versicolor
5	6,3	3,3	6,0	2,5	Iris-virginica
6	5,9	3,0	5,1	1,8	Iris-virginica

Classificação

- Ou seja, dados as **features (atributos) de uma determinada flor, qual a sua espécie (classe)**?
- Podemos dizer então, que a tarefa de classificação faz um **mapeamento dos atributos para uma classe**
- É comum em ML usarmos X para fazer referência aos atributos e y para fazer referência as classes


Conjunto de atributos (X)
Modelo
Classe (y)

Classificação

- Esse mapeamento dos atributos para uma classe é feita pelo modelo de ML
- Para isso existem **diversas técnicas disponíveis e sendo desenvolvidas**
- Para exemplificar, podemos imaginar um modelo que utiliza simples regras **SE-ENTÃO**

SE $petal\ length > 5$ E $petal\ width > 2$ ENTÃO $Iris\text{-}virginica$

Classificação - Treinamento

- Para "aprender" a fazer esse mapeamento, os algoritmos de ML precisam passar por um processo de **treinamento**
- Esse treinamento é uma etapa onde o modelo irá receber uma **amostra** dos dados e tentar classificá-los
- Esse processo envolve a avaliação de desempenho desse modelo, como uma **medida de erro**
- O objetivo do modelo e **diminuir esse erro** ao longo das observações dessa amostra

Classificação - Treinamento

- O modelo consegue medir o erro por que oferecemos as observações **juntamente com o rótulo**
- Portanto, o modelo consegue **comparar sua resposta com o rótulo real (ground truth)**
- Essa amostra chamamos de **conjunto de treino**
- Esse conjunto de treinamento será dividido outros dois, **treinamento e validação**
- A validação tem o objeto de **validar o treinamento, verificar a assertividade do modelo**

Classificação - Treinamento

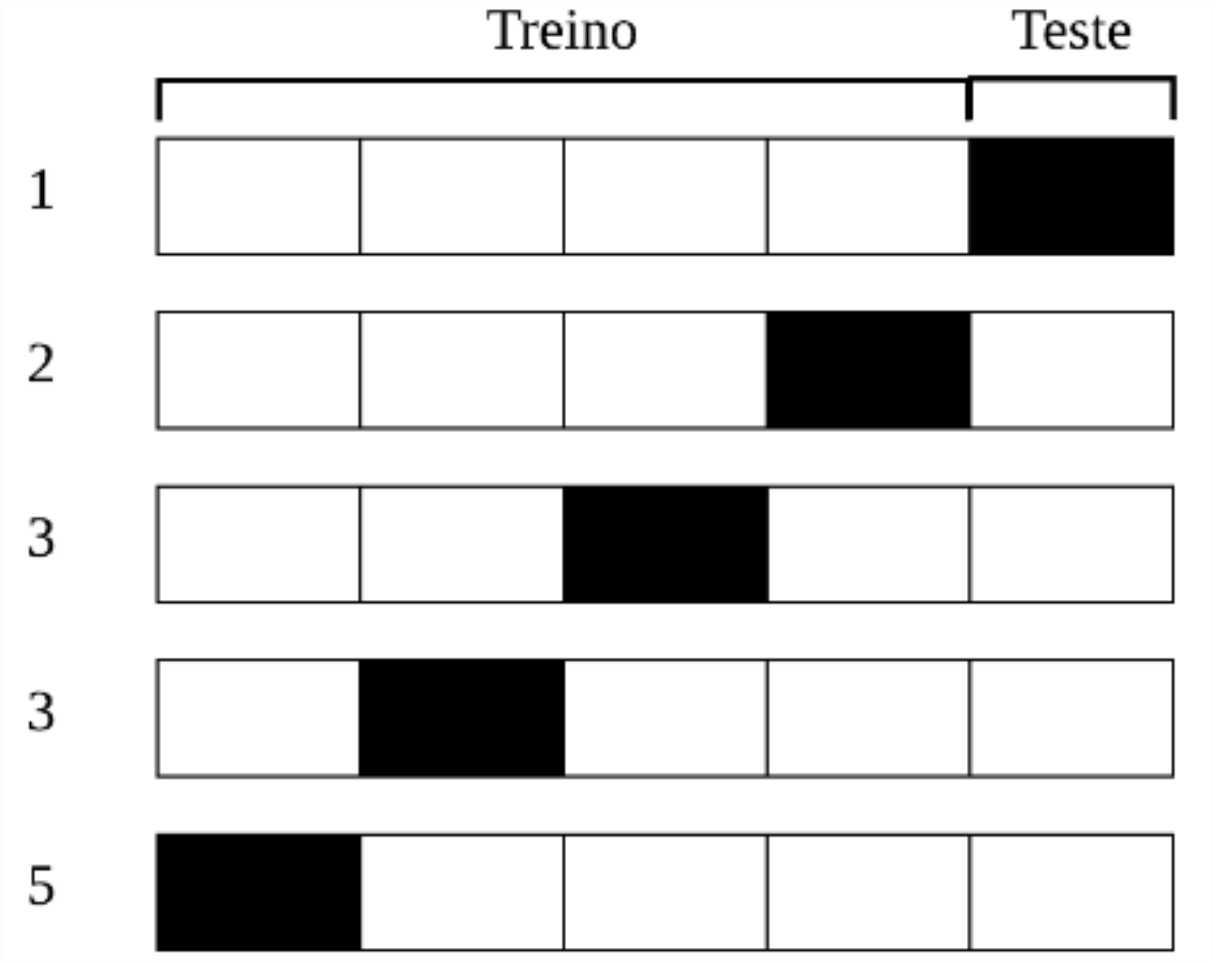
- Perceba que essa é uma abordagem em que **oferecemos a resposta verdadeira** para que o modelo consiga diminuir seu erro
- Isto é, existe uma espécie de **Supervisão**. Por isso chamamos essa abordagem de **Aprendizado Supervisionado**

Classificação - Treinamento e Amostragem

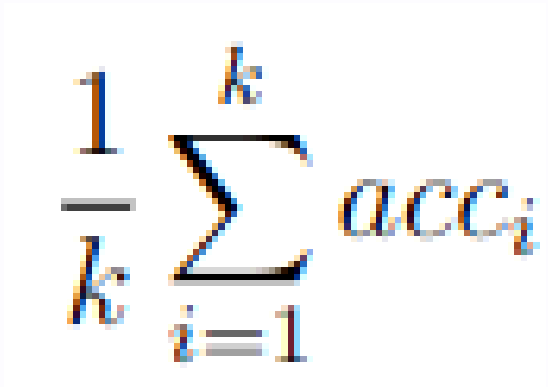
- Amostragem é uma parte fundamental no treinamento. Como já sabemos, precisamos dividir nossa base entre treinamento e validação. Existem algumas convenções, como, 80% treino e 20% validação, ou até mesmo 90/10
- Essa divisão será encontrada através de experimentos

Classificação - Treinamento e Amostragem

- Existem técnicas mais sofisticadas de amostragem, como por exemplo o **k-fold**
- O **k-fold** divide o conjunto de dados em k partições proporcionais, onde k geralmente é 5 ou 10, e o algoritmo é treinado em $k-1$ partições e validado na partição restante


Treino
Teste
1
2
3
3
5

- Após a execução das k partições a média dos resultados é calculado



Classificação - Medidas de Avaliação

- Mas como medimos os resultados de um algoritmo de classificação?
- Para cada observação que o algoritmo tenta classificar, podemos ter 4 resultados diferentes

Classificação - Medidas de Avaliação

- Verdadeiro Positivo (VP): O registro é classificado como **sendo** da class C e realmente **pertence** a classe C
- Verdadeiro Negativo (VN): O registro é classificado como **não sendo** da classe C e realmente **não pertence** a classe C

Classificação - Medidas de Avaliação

- Falso Positivo (FP): O registro é classificado como **sendo** da classe C mas **não pertence** a classe C
- Falso Negativo (FN): O registro é classificado como **não sendo** da classe C mas pertence a classe C

Classificação - Medidas de Avaliação

- A partir desse valores citados anteriormente, conseguimos calcular algumas medidas de avaliação, alguns exemplos são:
- Acurácia
- Sensibilidade
- Especificidade

Acurácia

- Considerando todos os resultados, quais foram classificados corretamente?

$$acc = \frac{VP + VN}{VP + FN + FP + VN}$$

Sensibilidade

- Proporção de VPs. De todos que **pertencem a classe** C, quantos realmente foram **classificados como pertencentes a essa classe**?

$$se = \frac{VP}{VP + FN}$$

Especificidade

- Proporção de VNs. De todos que **não pertencem a classe** C, quantos foram classificados **como não pertencente**?

$$sp = \frac{VN}{VN + FP}$$

Referências

- https://www.insper.edu.br/noticias/o-mar-de-dados-virou-um-oceano-e-nao-para-de-crescer-mas-nem-tudo-e-aproveitado/
- Russell, S. J. 1., & Norvig, P. (1995). Artificial intelligence: a modern approach. Englewood Cliffs, N.J., Prentice Hall.

- ALVES, Alexandre Henrick da Silva. Análise de novas abordagens para mineração de regras de classificação utilizando algoritmos genéticos. 2020. 134 f. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Uberlândia, Uberlândia, 2020. DOI http://doi.org/10.14393/ufu.di.2020.260.