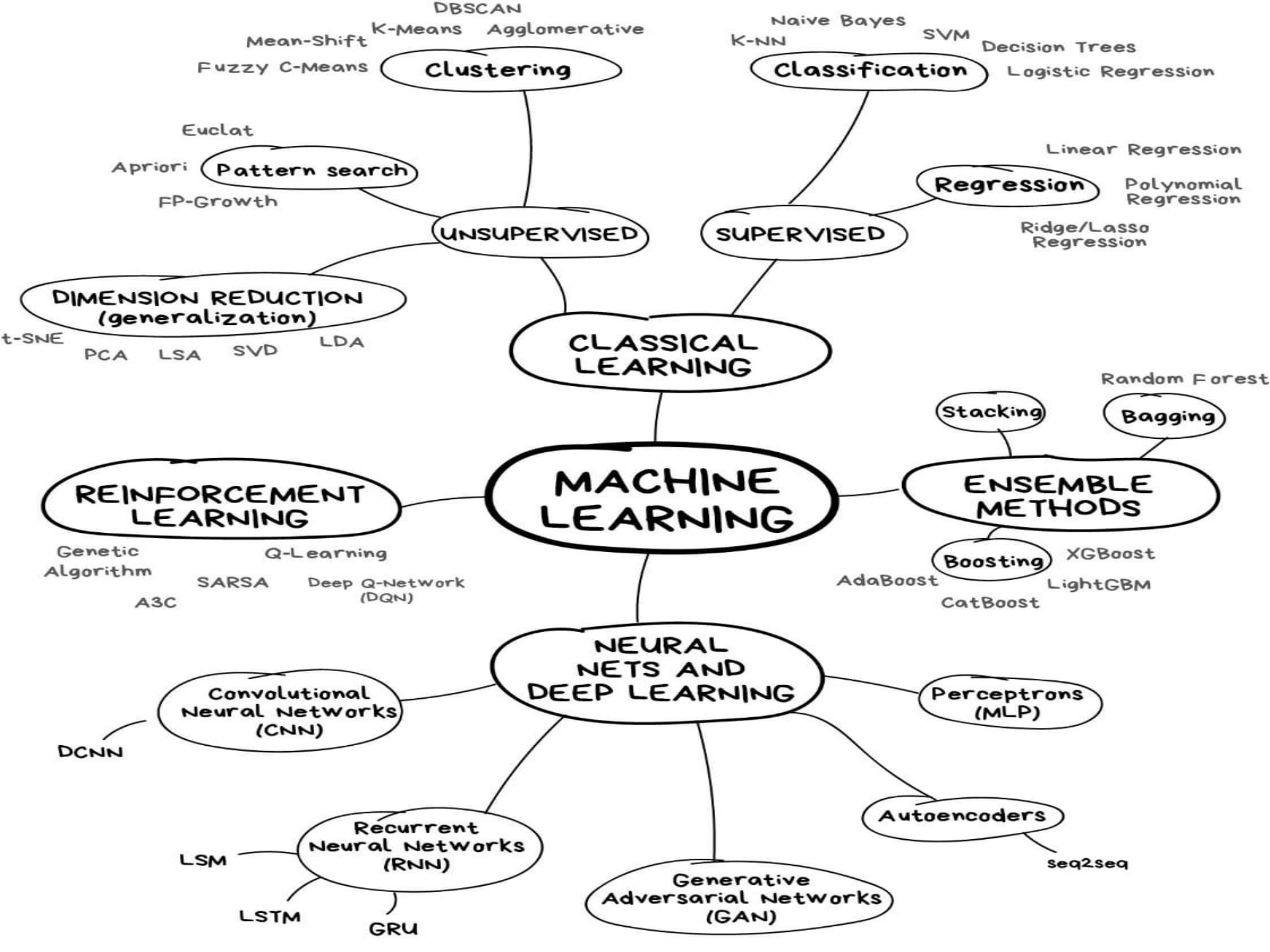
Aula 16/09/2020

Discussão sobre as diversas ideias sobre aprendizado supervisionado e não-supervisionado.


MACHINE LEARNING
CLASSICAL LEARNING
UNSUPERVISED
Clustering
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Pattern search
Euclat
Apriori
FP-Growth
DIMENSION REDUCTION (generalization)
t-SNE
PCA
LSA
SVD
LDA
SUPERVISED
Classification
Naive Bayes
K-NN
SVM
Decision Trees
Logistic Regression
Regression
Linear Regression
Polynomial Regression
Ridge/Lasso Regression
REINFORCEMENT LEARNING
Genetic Algorithm
A3C
SARSA
Q-Learning
Deep Q-Network (DQN)
ENSEMBLE METHODS
Stacking
Bagging
Random Forest
Boosting
AdaBoost
CatBoost
XGBoost
LightGBM
NEURAL NETS AND DEEP LEARNING
Convolutional Neural Networks (CNN)
DCNN
Perceptrons (MLP)
Recurrent Neural Networks (RNN)
LSM
LSTM
GRU
Generative Adversarial Networks (GAN)
Autoencoders
seq2seq

“o único objetivo do aprendizado de máquina é prever os resultados com base nos dados recebidos. “

Mas várias áreas da Estatística, como Regressão, já não fazem isso?

Quanto maior a variedade de amostras, mais fácil será encontrar **padrões relevantes** para prever um resultado.

Ensinar a máquina a reconhecer padrões

Classificação tem a mesma ideia.

Contudo, podemos pensar que o reconhecimento de padrões permite a Classificação dos dados em categorias diferentes.

Quanto mais diversificados forem os dados, melhor será o resultado.

Existem duas maneiras principais de obter os dados: a manual e a automática.

É extremamente difícil coletar uma boa coleção de dados, ao qual geralmente denominamos conjunto de dados, ou datasets. Eles são tão importantes que as empresas podem até revelar seus algoritmos, mas raramente seus conjuntos de dados.

O Twitter fornece gratuitamente um conjunto de tweets.

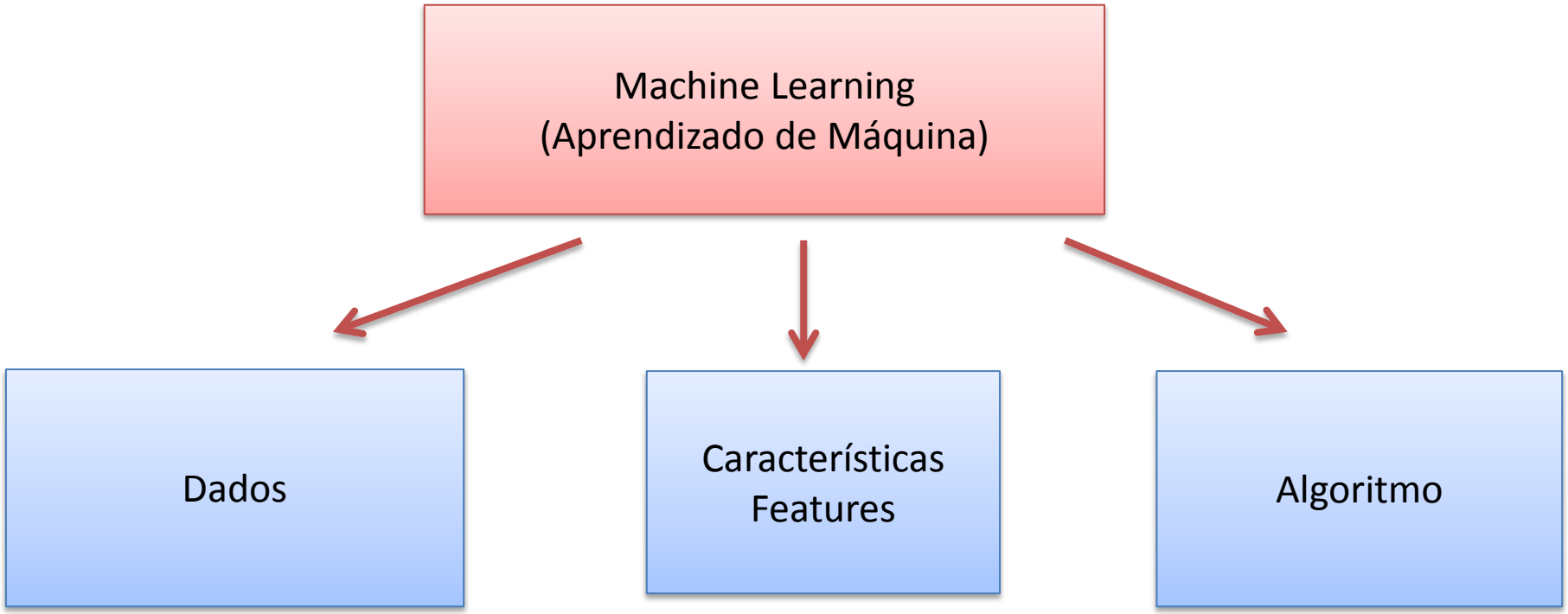
Características/Features

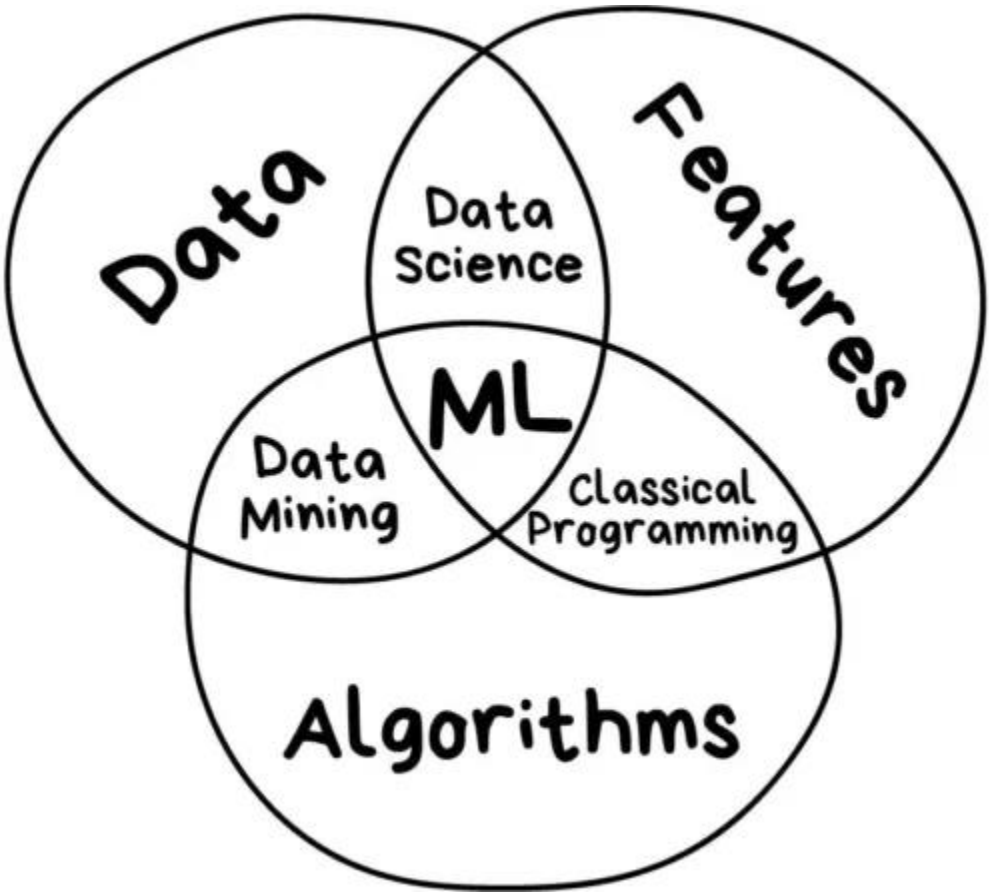
são os fatores que a máquina vai analisar para classificar os dados.

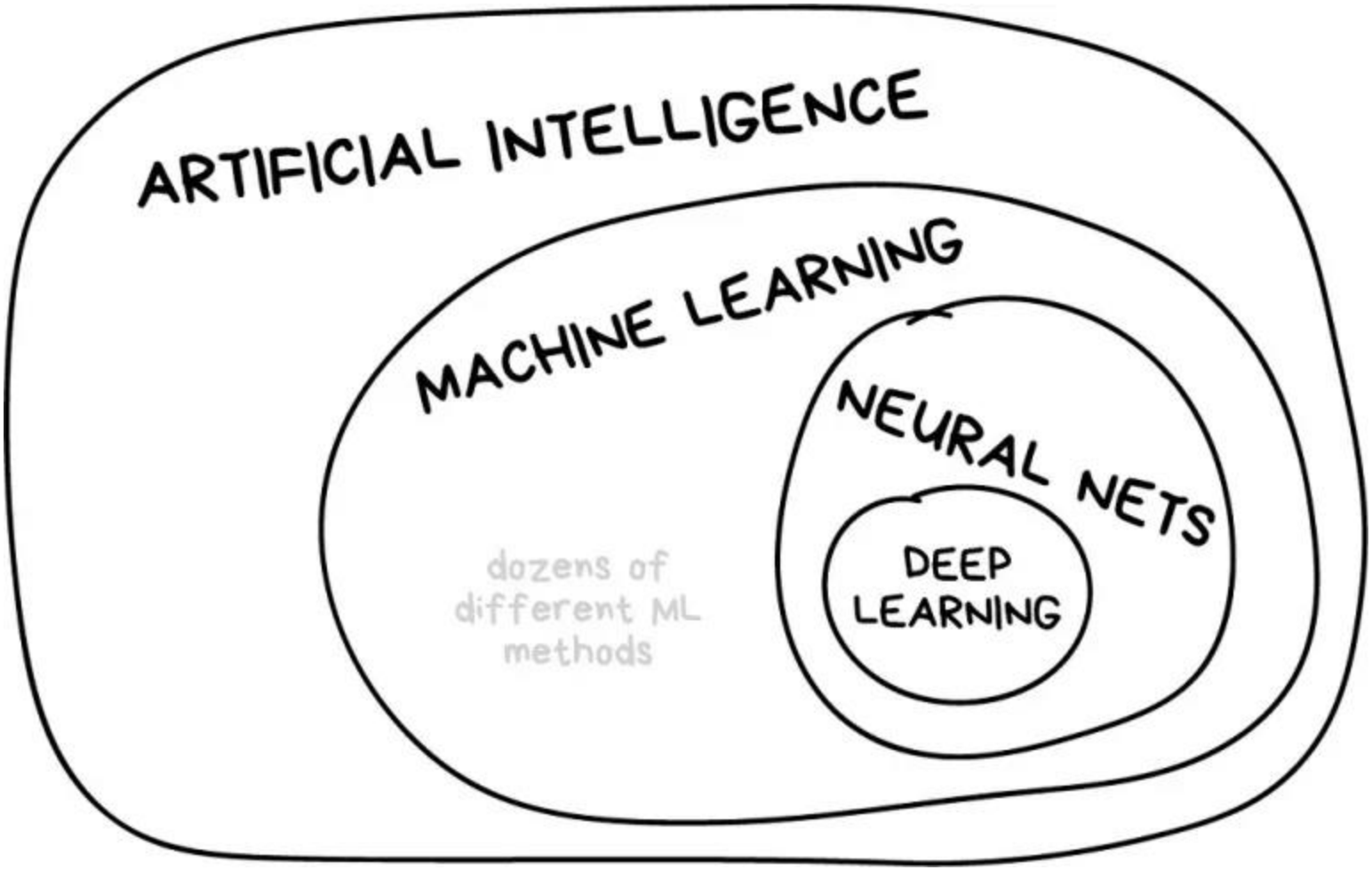
Exemplos

- Um conjunto de tweets sobre o novo bolo do Starbucks. Como serão classificados? Quais as características serão consideradas para classificar os tweets?
- Um conjunto de imagens de manchas na pele.
- Um conjunto de produtos que uma família compra.

Selecionar as características certas sobre seu conjunto de dados, para poder classificá-los é muito importante.


Machine Learning
(Aprendizado de Máquina)
Dados
Características
Features
Algoritmo


Data
Features
Algorithms
Data Science
ML
Data Mining
Classical Programming


ARTIFICIAL INTELLIGENCE
MACHINE LEARNING
dozens of different ML methods
NEURAL NETS
DEEP LEARNING

CLASSICAL MACHINE LEARNING

Data is pre-categorized or numerical

SUPERVISED

Predict a category

CLASSIFICATION

«Divide the socks by color»

Predict a number

REGRESSION

«Divide the ties by length»

Data is not labeled in any way

UNSUPERVISED

Divide by similarity

CLUSTERING

«Split up similar clothing into stacks»

Find hidden dependencies

DIMENSION REDUCTION (generalization)

«Make the best outfits from the given clothes»

Identify sequences

ASSOCIATION

«Find what clothes I often wear together»

No Aprendizado Supervisionado , a máquina tem um “supervisor” ou um “professor”, que dá à máquina todas as respostas, como por exemplo para identificar se é um gato ou um cachorro em uma foto. Neste caso, o “professor” já dividiu (rotulou) os dados em gatos e cães e a máquina está usando esses exemplos para aprender.

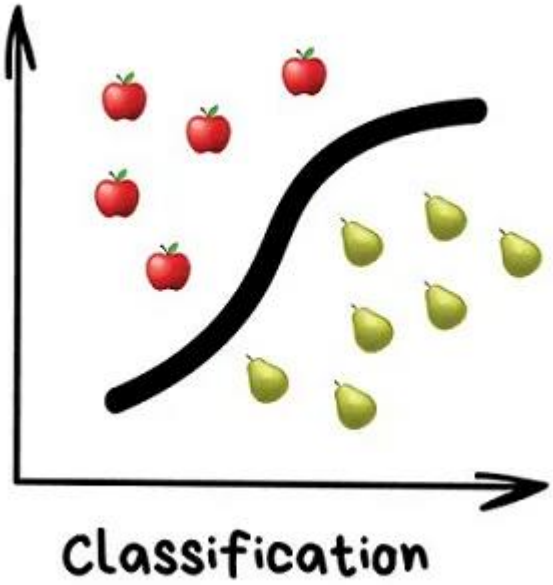
Aprendizado não supervisionado significa que a máquina é deixada sozinha com uma pilha de fotos de animais e uma tarefa: descobrir quem é quem.

- Os dados não são rotulados,
- não há “professor”,
- e a máquina está tentando encontrar padrões por conta própria.

Classificação

Os algoritmos de classificação dividem os objetos com base em um dos atributos conhecidos de antemão.

- Separa as meias com base na cor,
- documentos baseados na linguagem,
- e as músicas por gênero.



Usada nos dias de hoje para:

- Filtragem de spam;
- detecção de idioma;
- Pesquisa por documentos semelhantes;
- Análise de sentimentos;
- Reconhecimento de caracteres e números manuscritos;
- Detecção de fraude.

Quais são os algoritmos mais usados?

Exemplo prático de classificação:

- Vamos dizer que você precisa de algum dinheiro a crédito.
- Como o banco saberá se você vai pagá-lo de volta ou não?
- Não há como saber com certeza.
- Mas o banco tem muitos perfis de pessoas das quais recebeu de volta o dinheiro emprestadado.
- O banco tem dados sobre idade, educação, ocupação, salário e – o mais importante – o fato de pagar o dinheiro de volta. Ou não.

Usando esses dados, podemos ensinar a máquina a encontrar padrões e obter a resposta. Entretanto, o banco não pode confiar cegamente na resposta da máquina. E se houver uma falha no sistema ou ataque hacker?

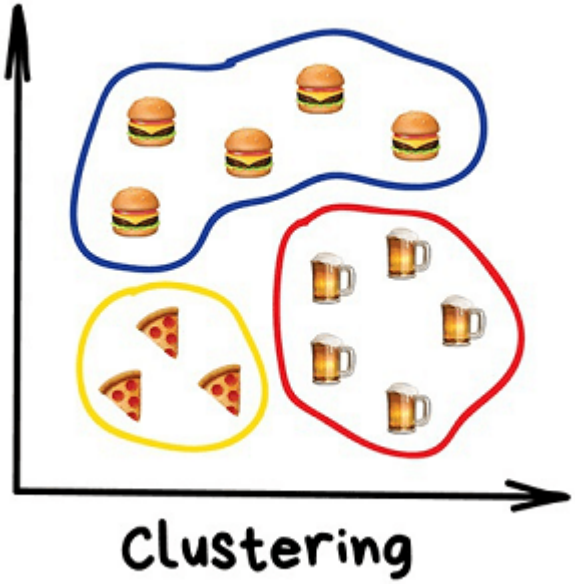
“Para lidar com isso, temos as **árvores de decisão. “**

Mas o que são essas árvores e por que garantiriam a liberação de crédito para um bom pagador?

Árvores de decisão são amplamente usadas em esferas de alta responsabilidade: diagnósticos, medicina e finanças.

Aprendizagem não supervisionada

Clusterização: é uma classificação sem classes predefinidas. É como dividir meias por cor quando você não se lembra de todas as cores que você tem.



O algoritmo tenta encontrar objetos semelhantes (por características) e agrupá-los em clusters (ou classes). Aqueles que têm muitas características semelhantes são unidos em uma classe.

Algumas aplicações

- Segmentação de mercado (tipos de clientes, fidelidade).
- Detectar um comportamento anormal.

Algoritmos populares: **K-means_clustering**, Mean-Shift, DBSCAN.

Nesta disciplina vamos estudar o **algoritmo K-means_clustering**