INSIGHT

Data Science Laboratory
Federal University of Ceará

Aprendizado Não Supervisionado

Francisco Carlos

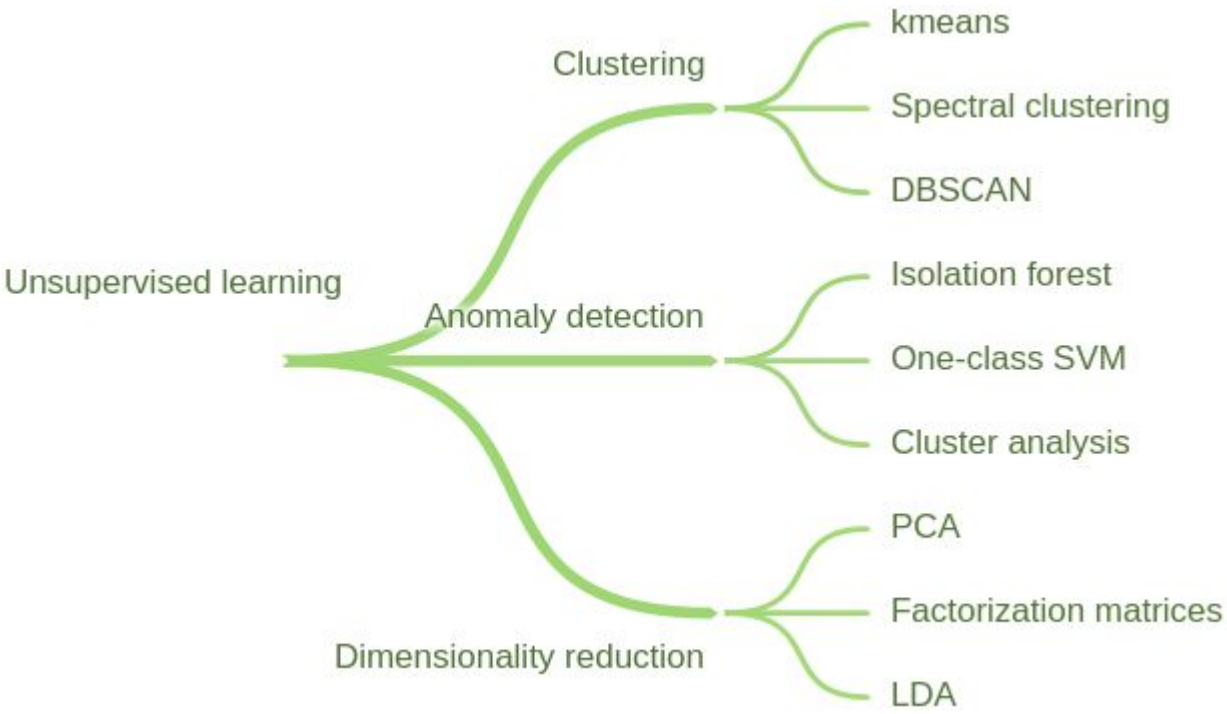
AGENDA

1. Aprendizado não supervisionado
2. Clusterização
3. K-means

1. Aprendizado não supervisionado

Como aprender sobre dados sem rótulos?

Aprendizado não supervisionado



Mapa completo: https://storage.ning.com/topology/rest/1.0/file/get/135091853?profile=original

Aprendizado não supervisionado

No aprendizado **supervisionado**, os dados de treinamento **possuem rótulos**.

Exemplo:

- Classificação:
  [0.50, 0.78, 0.32, 0.89, 0.41] **[“Alto”]**

- Regressão:
  [0.34, 0.76, 0.48, 0.12, 0.43] **[257]**

Em muitas situações reais temos que lidar com dados **não supervisionados**, ou seja, que **não possuem rótulos.**

Aprendizado não supervisionado

Por que os dados não possuem rótulos?

- Rotular um grande conjunto de dados pode custar muito **tempo**, **esforço** e **dinheiro**;
- Em muitas situações podemos querer descobrir as **similaridades** ou **diferenças** entre os padrões existentes nos dados.

Aprendizado não supervisionado

Exemplos:

- Seguro: identificar grupo de clientes que acionam sinistros com alta frequência;
- Classificação de documentos;
- Planejamento urbano: identificar grupos de casas conforme valor, tipo e localização;
- Organizar produtos em lojas;
- Detecção de fraudes.

2. Clusterização

Criando grupos de dados

Clusterização

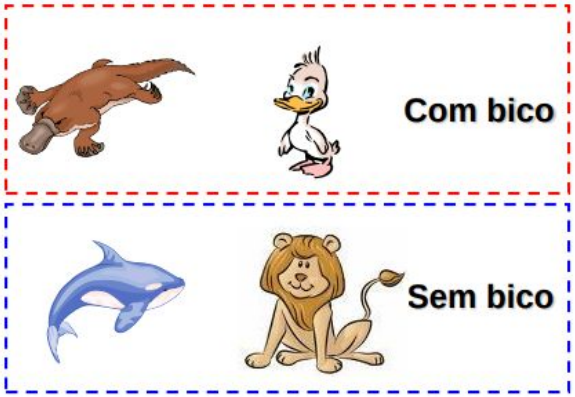
Clusterização é o **agrupamento** em conjuntos de dados, utilizando **similaridades** baseadas nas **características**.

Exemplo, como separar esse conjunto de animais?

Clusterização

Clusterização é o **agrupamento** em conjuntos de dados, utilizando **similaridades** baseadas nas **características**.

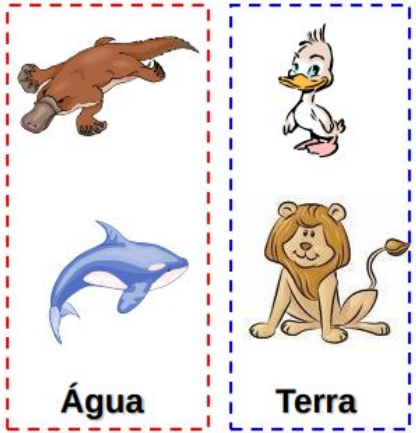
Exemplo, como separar esse conjunto de animais?



Clusterização

Clusterização é o **agrupamento** em conjuntos de dados, utilizando **similaridades** baseadas nas **características**.

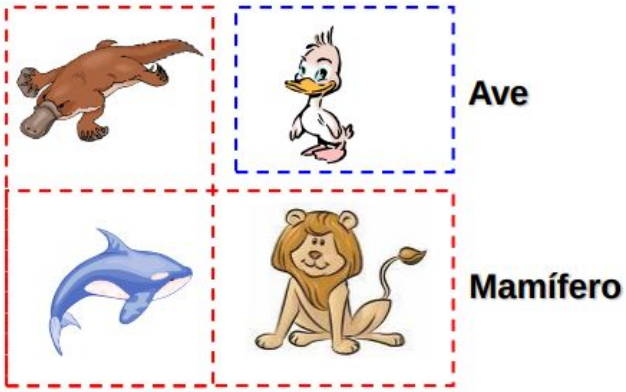
Exemplo, como separar esse conjunto de animais?



Clusterização

Clusterização é o **agrupamento** em conjuntos de dados, utilizando **similaridades** baseadas nas **características**.

Exemplo, como separar esse conjunto de animais?



3.
K-means

clusterização em k partições

ETAPAS PRINCIPAIS

1. Inicialização
2. Agrupamento
3. Atualização
4. Convergência

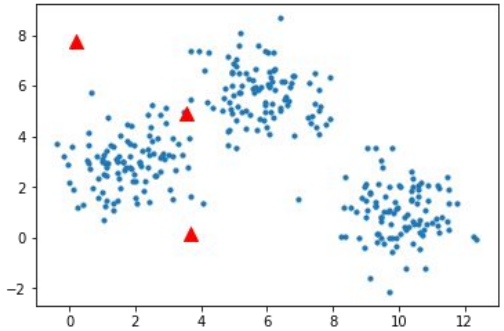
1. INICIALIZAÇÃO

Cada grupo será representado por uma amostra com a mesma dimensão dos dados, intitulado de **centróide**.

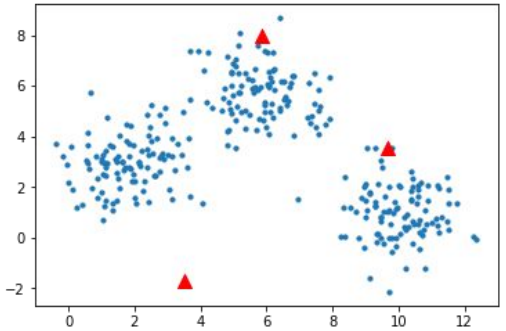
A primeira etapa consiste em escolher **K dados randomicamente** para representar os centróides iniciais.

Uma boa inicialização dos centróides possui as seguintes características:

1. Centróides esparsos entre si;
2. Centróides próximos do conjunto de dados.



Exemplo 1



Exemplo 2

ETAPAS PRINCIPAIS

1 Inicialização

2 Agrupamento

3 Atualização

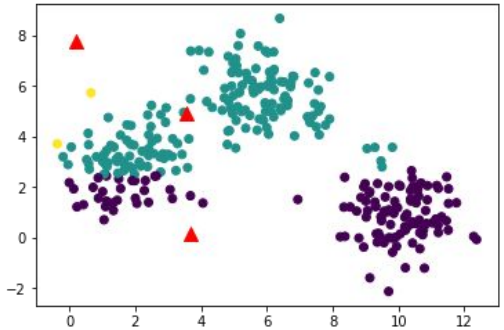
4 Convergência

2. ATRIBUIÇÃO AOS CLUSTERS

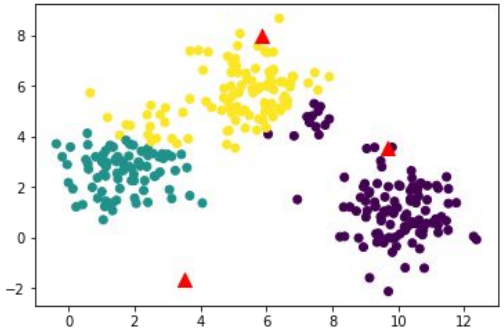
Cada **centróide** será responsável pela formação de um **cluster**.

Na segunda etapa, **cada amostra** do conjunto de dados será atribuído ao **centróide mais próximo**, utilizando a função de distância euclidiana.

Um cluster será o agrupamento de todas as amostras próximas a um centróide.



Exemplo 1



Exemplo 2

ETAPAS PRINCIPAIS

1 Inicialização

2 Agrupamento

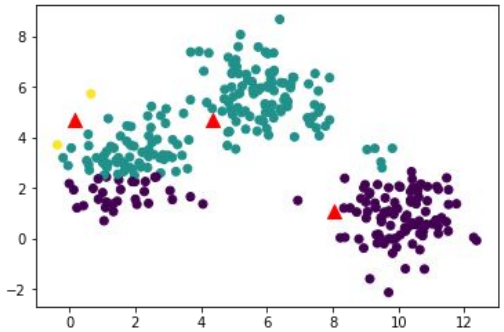
3 Atualização

4 Convergência

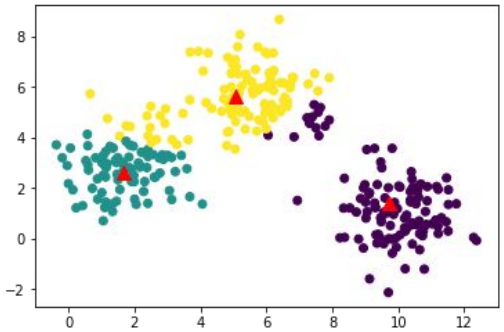
3. ATUALIZAÇÃO DOS CENTRÓIDES

Após a atribuição dos dados aos respectivos clusters, a etapa de atualização consiste em **recalcular os centróides**.

O novo valor de cada centróide será a média de todos os dados pertencentes ao cluster.



Exemplo 1



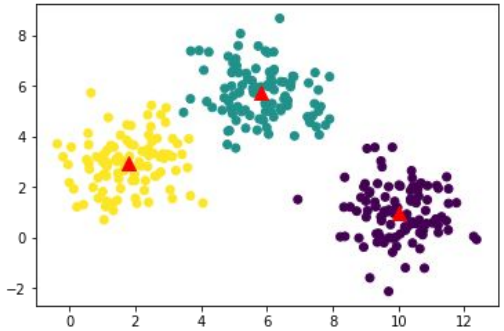
Exemplo 2

4. CONVERGÊNCIA

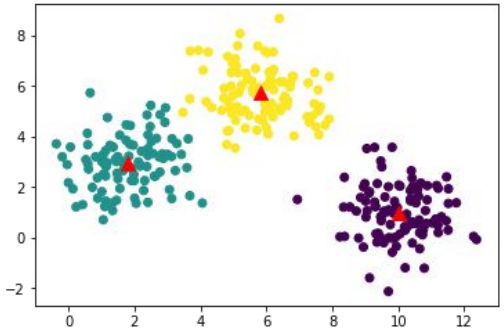
O algoritmo repete os passos de **agrupamento (2)** e **atualização (3)** até que uma das seguintes proposições seja verdadeira:

- Os centróides não se modificar na etapa de atualização;
- A modificação dos centróides ser menor que um limiar estabelecido;
- Terminar o número de épocas do algoritmo.

Os cluster resultantes do K-means podem ser representados pelos centróides gerados.

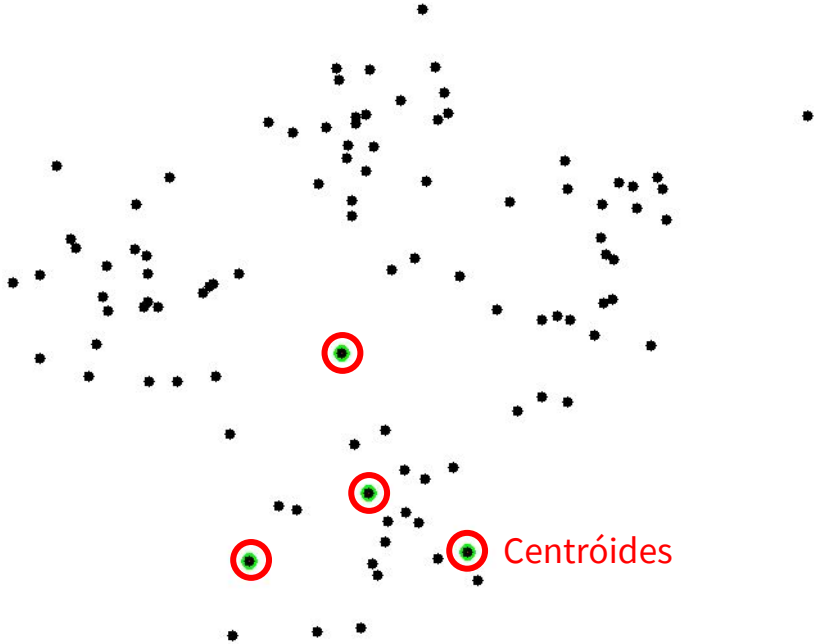


Exemplo 1



Exemplo 2

Exemplo



Fonte: http://shabal.in/visuals/kmeans/6.html

Exemplo

Fonte: http://shabal.in/visuals/kmeans/6.html

Número de clusters

Como escolher o valor de K?

A princípio o algoritmo do K-means parece ser um pouco ingênuo, pois ele divide os dados em K clusters, mesmo que não existam K clusters. Alguns métodos podem ajudar na escolha do valor de K.

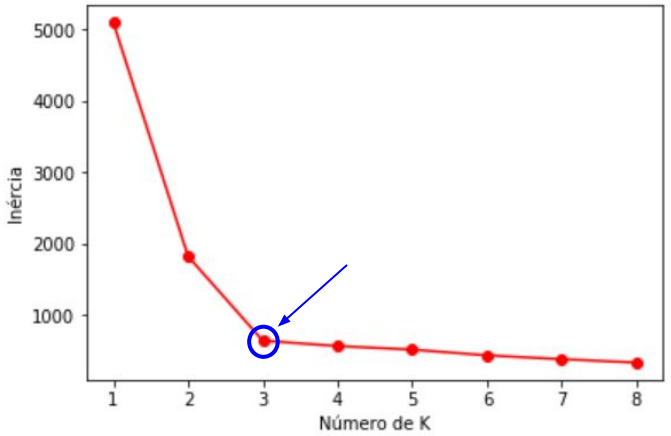
Exemplo:

- Método do cotovelo

Método do cotovelo

Executar o algoritmo K-means para um intervalo de valores de K ($1 \leq K \leq 20$, por exemplo), para cada valor de K é calculado a soma dos quadrados das distâncias dos dados para o centróide do cluster.

A ideia é analisar a variação intra-cluster para diferentes valores de K, buscando o número ideal da quantidade de clusters.



Colocar a mão na massa!

Regras:

- Codificação Individual
- Pode pesquisar na internet a vontade

Pontuação:

- Inicializar os centróides ---------- (1 ponto)
- Função de distância ---------- (1 ponto)
- Calcular o centróide mais próximo ---------- (1 ponto)
- Centróide mais próximo para todos os dados -- (1 ponto)
- Métrica de avaliação ---------- (1 ponto)
- Atualizar os clusters ---------- (2 pontos)
- Algoritmo completo ---------- (2 pontos)
- Método do cotovelo ---------- (1 ponto)

https://github.com/InsightLab/data-science-cookbook/blob/master/2020/trabalho-03/Construindo%20KMeans.ipynb

K-means
Complexidade

Complexidade de espaço: o espaço necessário para armazenar os dados e os centróides.

Complexidade de espaço = O((m+k)*n), no qual **m** é a quantidade de dados, **k** é o número de centróides e **n** é o número de atributos.

Complexidade de tempo: é um problema NP-difícil, porém executando um número fixo de iterações, o algoritmo padrão apenas faz uma aproximação do ótimo local.

Complexidade de tempo = para um número fixo de **t** iterações, O(t*k*m*n), no qual **m** é a quantidade de dados, **k** é o número de centróides e **n** é o número de atributos.

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

K-means
Vantagens e Desvantagens

Vantagens

1. Fácil de implementar;
2. Com grande número de atributos, o K-means é computacionalmente mais rápido que a clusterização hierárquica;
3. K-means pode produzir clusters mais concêntricos;
4. Uma amostra pode mudar de cluster, quando os centróides são recalculados.

Desvantagens

1. Inicialização dos centróides tem um grande impacto no resultado final;
2. Sensível a escala dos dados;
3. Todos os dados pertencem a um grupo;
4. É necessário definir o número de **k**.

OBRIGADO!

Dúvidas?

Você pode me encontrar em

- carlos@insightlab.ufc.br
- Telegram: @CarlosJun