
INSIGHT
Data Science Laboratory
Federal University of Ceará

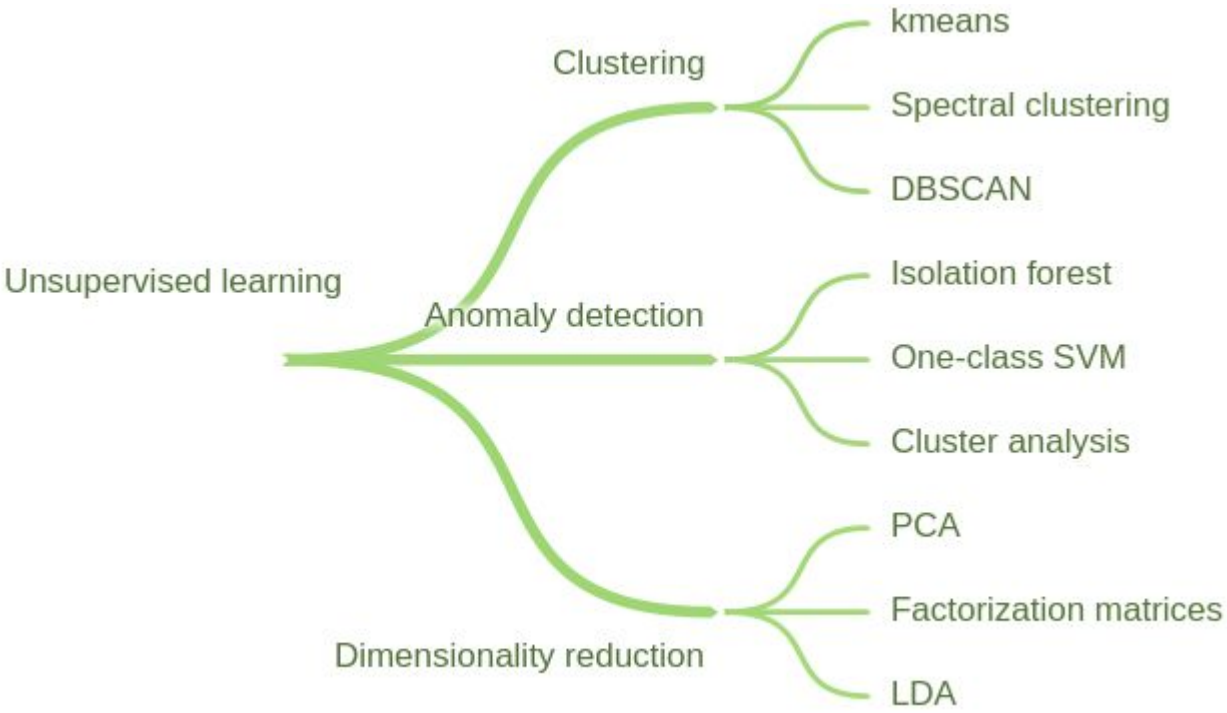
AGENDA

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

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] [“Bom”]

- 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

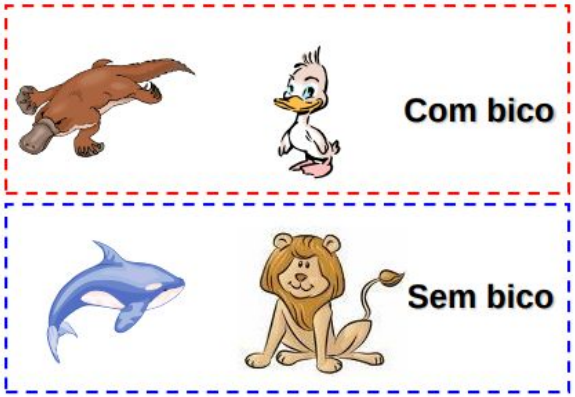
Clusterização é o **agrupamento** em conjuntos de dados, utilizando **similaridade** baseadas nas **características**.

Exemplo, como separar esse conjunto de animais?

Clusterização

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

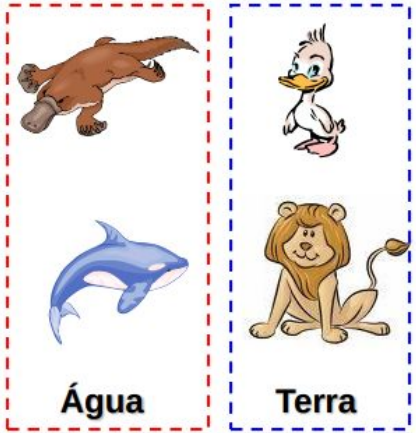
Exemplo, como separar esse conjunto de animais?



Clusterização

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

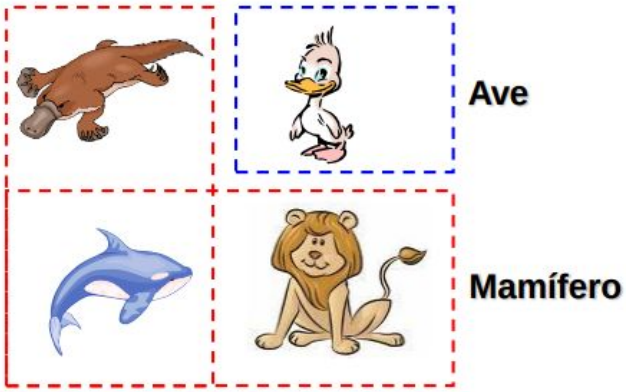
Exemplo, como separar esse conjunto de animais?



Clusterização

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

Exemplo, como separar esse conjunto de animais?



3.
K-means

Algoritmo de clusterização

ETAPAS PRINCIPAIS

1

2

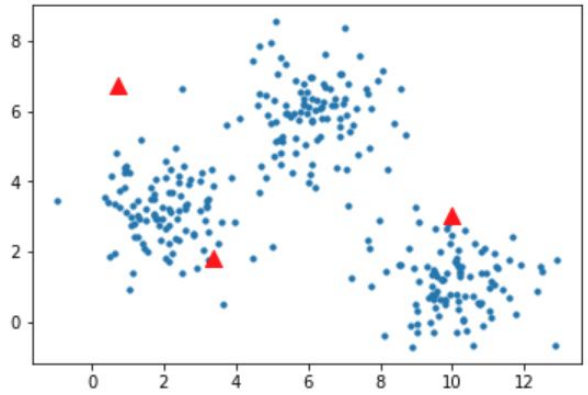
3

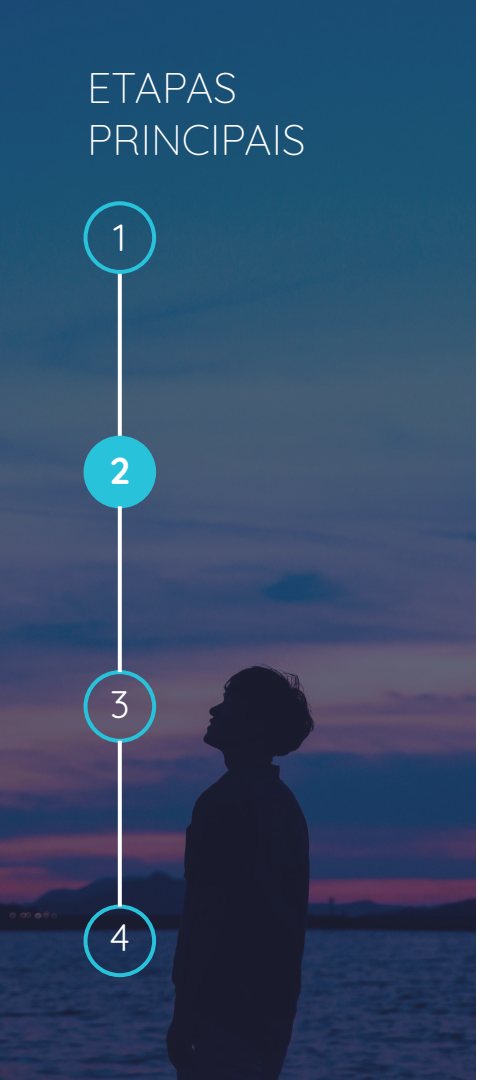
4

1. INICIALIZAÇÃO

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

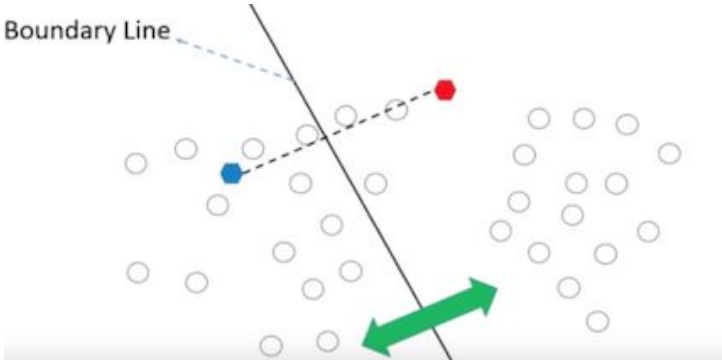
Uma boa maneira para inicializar os centróides, é utilizar as próprias amostras para criar pontos próximos ao conjunto de dados e esparsos entre si.





2. ATRIBUIÇÃO AOS CLUSTERS

Na segunda etapa, cada dado será atribuído a um cluster, que será o centróide mais próximo de acordo com uma função de distância.



ETAPAS PRINCIPAIS

1

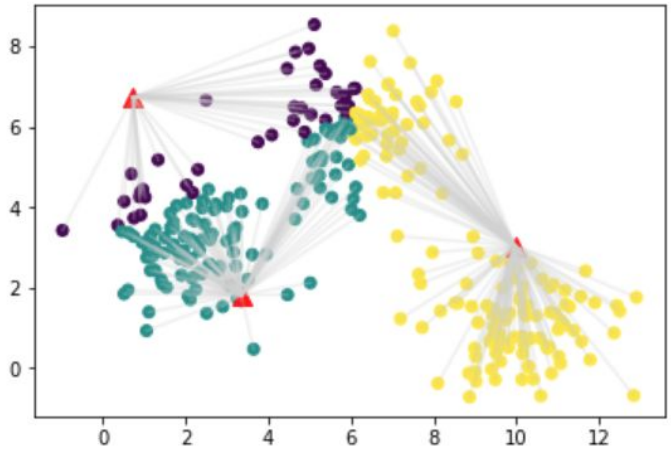
2

3

4

3. ATUALIZAR OS CENTRÓIDES

Após a atribuição dos dados aos clusters, a etapa de atualização consiste em calcular novos centróides. O novo valor de cada centróide será a média de todos os dados pertencentes ao cluster.



ETAPAS PRINCIPAIS

1

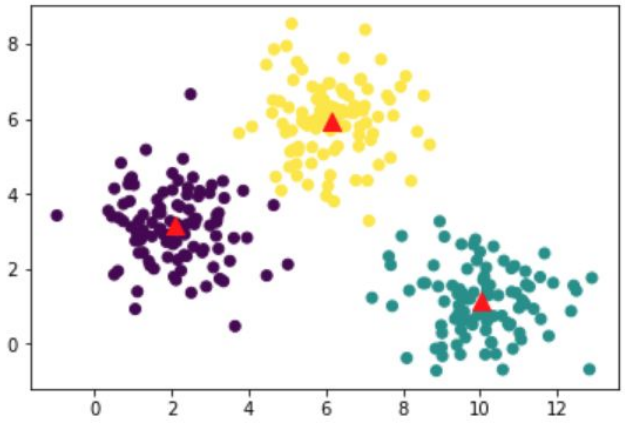
2

3

4

4. FINALIZAÇÃO

O algoritmo repete os passos 2 e 3 até não haver mais mudança na atualização dos centróides.



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.

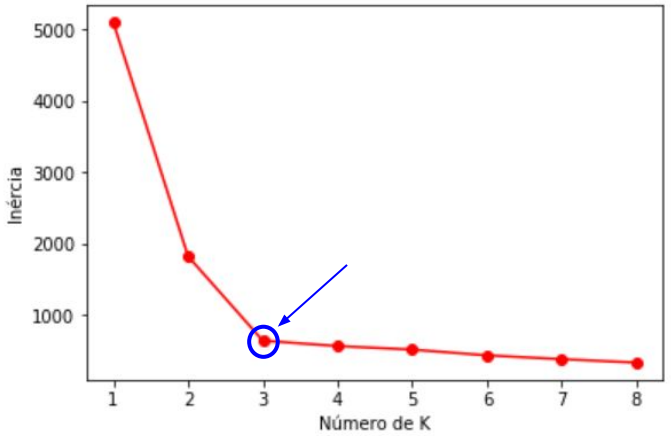
Exemplo:

- Método do cotovelo
- Dendrograma

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 ---------- (1 ponto)
- Algoritmo completo ---------- (2 pontos)
- Método do cotovelo ---------- (2 pontos)

REFERÊNCIAS

- Doutorando Lucas Cambuim - UFPE
  http://www.cin.ufpe.br/~lfsc/cursos/introducaoainteligenciaartificial/IA-Aula12-Clusterizacao.pdf
- Mestrando Felipe Zschornack R. Saraiva - UFC
  https://docs.google.com/presentation/d/10SnrYrevdnGF2JoYkIes2oBFa-Ttz7cZkgq_czH3AzI/edit#slide=id.g1726f05f0e_0_66
- Professor Edirlei Soares de Lima - UERJ
  http://edirlei.3dgb.com.br/aulas/ia_2011_2/IA_Aula_18_Aprendizado_Nao_Supervisionado.pdf
- Scikit-Learn - Machine Learning in Python
  https://scikit-learn.org/stable/modules/clustering.html

OBRIGADO!

Dúvidas?

Você pode me encontrar em

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