UNIPAC - CENTRO UNIVERSITÁRIO PRESIDENTE ANTÔNIO CARLOS
CAMPUS BARBACENA

Bacharelado em Ciência da Computação

Mineração de dados

Material de Apoio

Parte VII – Clusterização

Prof. Felipe Roncalli de Paula Carneiro
felipecarneiro@unipac.br

1º sem / 2022

Material cedido pela Profª Lívia e Profº Osvano

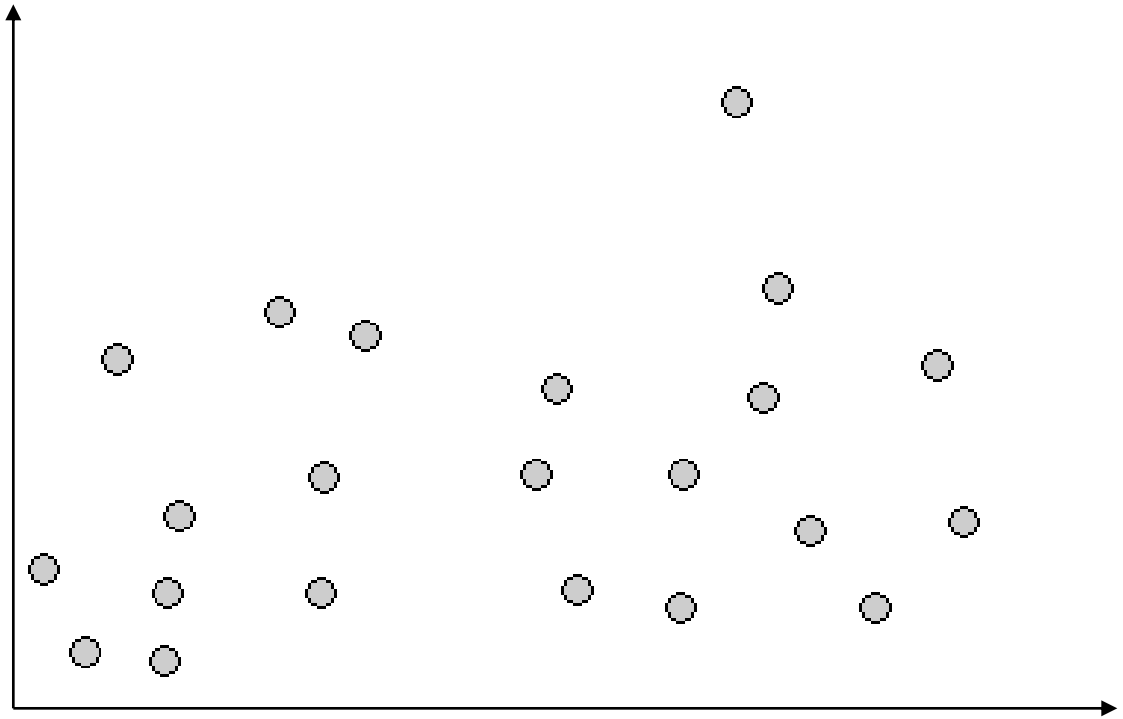
Sumário

- Aprendizagem Não-Supervisionada
- Clusterização *(Clustering)*
- *Cluster*
- Algoritmos

Aprendizagem Não-Supervisionada

- O que pode ser feito quando se tem um conjunto de exemplos mas não se conhece as categorias envolvidas?

Como ‘‘classificar’’ esses pontos?



Por que estudar esse tipo de problema?

Aprendizagem Não-Supervisionada

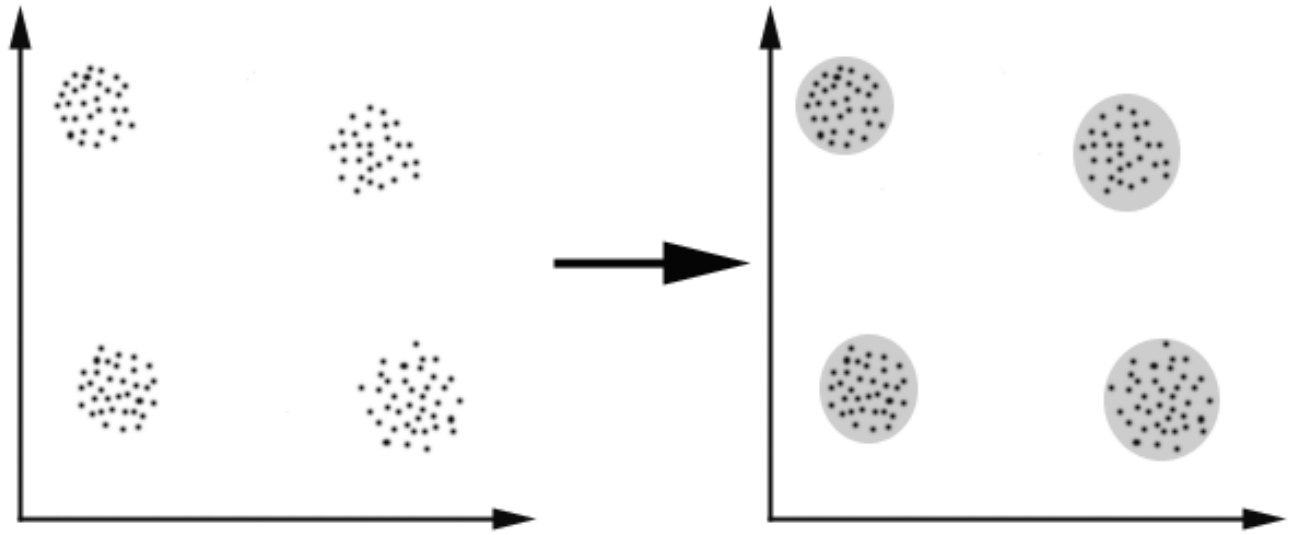
- Primeiramente, coletar e rotular bases de dados pode ser extremamente caro.
  - Gravar voz é barato, mas rotular todo o material gravado é caro.
  - Rotular TODA uma grande base de imagens é muito caro, mas... alguns elementos de cada classe não
- Segundo, muitas vezes não se tem conhecimento das classes envolvidas.
  - Trabalho exploratório nos dados (ex. *Data Mining*.)

Aprendizagem Não-Supervisionada

- Pré-classificação:
  - Suponha que as categorias envolvidas são conhecidas, mas a base não está rotulada.
  - Pode-se utilizar a aprendizagem não-supervisionada para fazer uma pré-classificação, e então treinar um classificador de maneira supervisionada

Clusterização *(Clustering)*

- É a organização dos objetos similares (em algum aspecto) em grupos.



Quatro grupos (clusters)

Cluster

- Uma coleção de objetos que são similares entre si, e diferentes dos objetos pertencentes a outros *clusters*.
- Isso requer uma medida de similaridade.
- No exemplo anterior, a similaridade utilizada foi a *distância.*
  - *Distance-based Clustering*

Algoritmos

- Clusters formados por diferentes métodos de agrupamento podem ter características diferentes.
- Os clusters podem ter diferentes formas, tamanhos e densidades.
- Os clusters podem formar uma hierarquia.
- Os clusters podem ser desconexos, tocantes ou sobrepostos.
- Em particular, fornecemos uma visão geral de três métodos de agrupamento:
  - agrupamento k-Means;
  - agrupamento hierárquico;
  - DBSCAN.

k-Means

- É um algoritmo de aprendizagem não supervisionada.
- Ele não classifica, mas agrupa vetores de atributos similares, isto é, coloca em um mesmo agrupamento vetores similares.
- Por ser um bastante simples e funcionar bem na prática, ele é um dos principais e mais usados métodos de agrupamento.

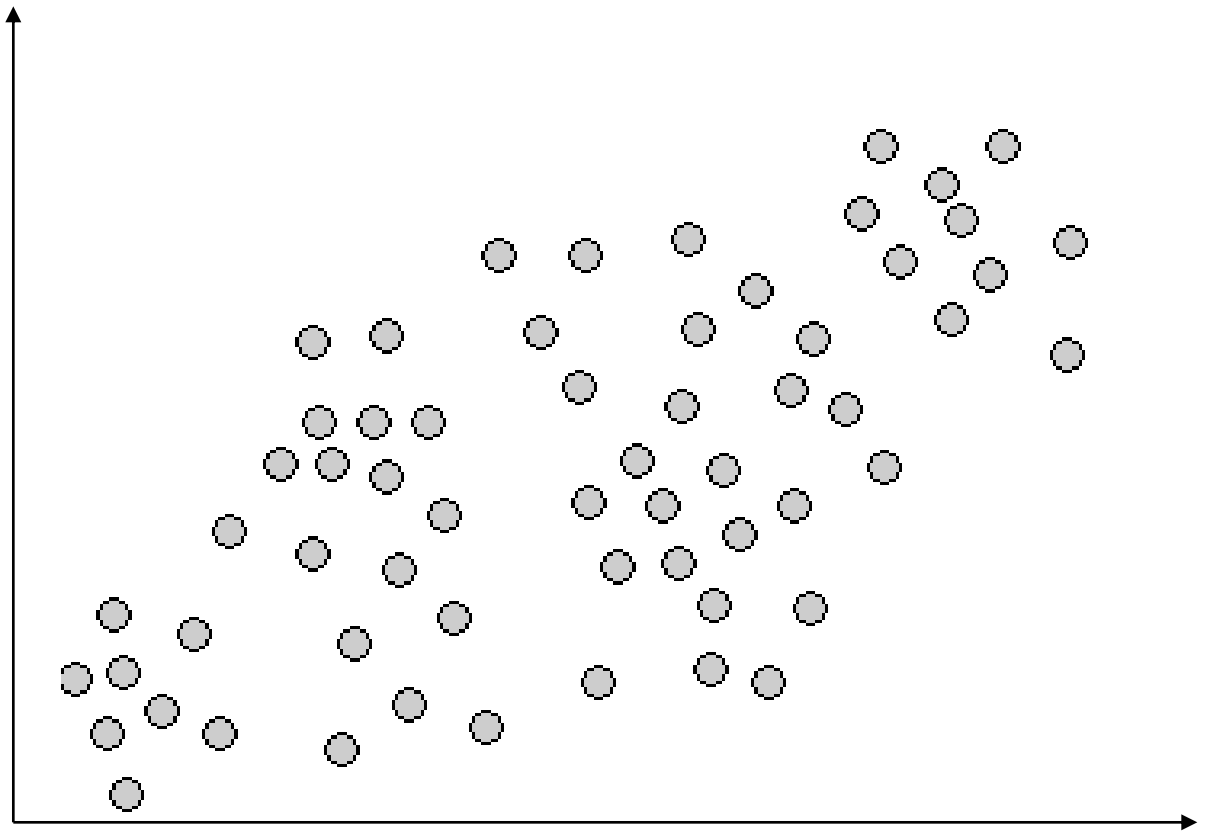
k-Means

- É a técnica mais simples de aprendizagem não supervisionada.
- Consiste em fixar *k* centróides (de maneira aleatória), um para cada grupo (clusters).
- Associar cada indivíduo ao seu centróide mais próximo.
- Recalcular os centróides com base nos indivíduos classificados.

Algoritmo *k-Means*

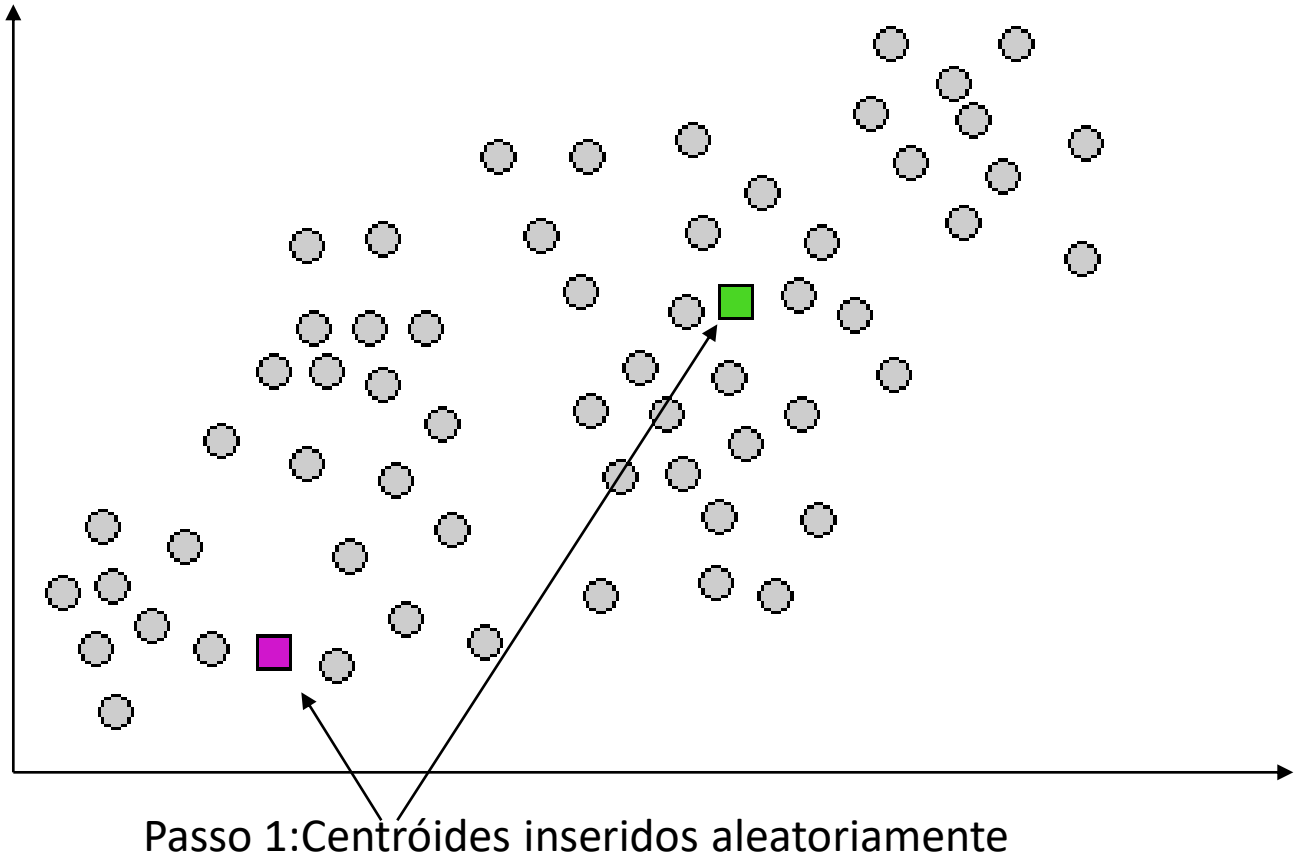
1. Determinar os centróides.
2. Atribuir a cada objeto do grupo o centróide mais próximo.
3. Após atribuir um centróide a cada objeto, recalcular os centróides.
4. Repetir os passos 2 e 3 até que os centróides não sejam modificados.

k-Means – Um Exemplo



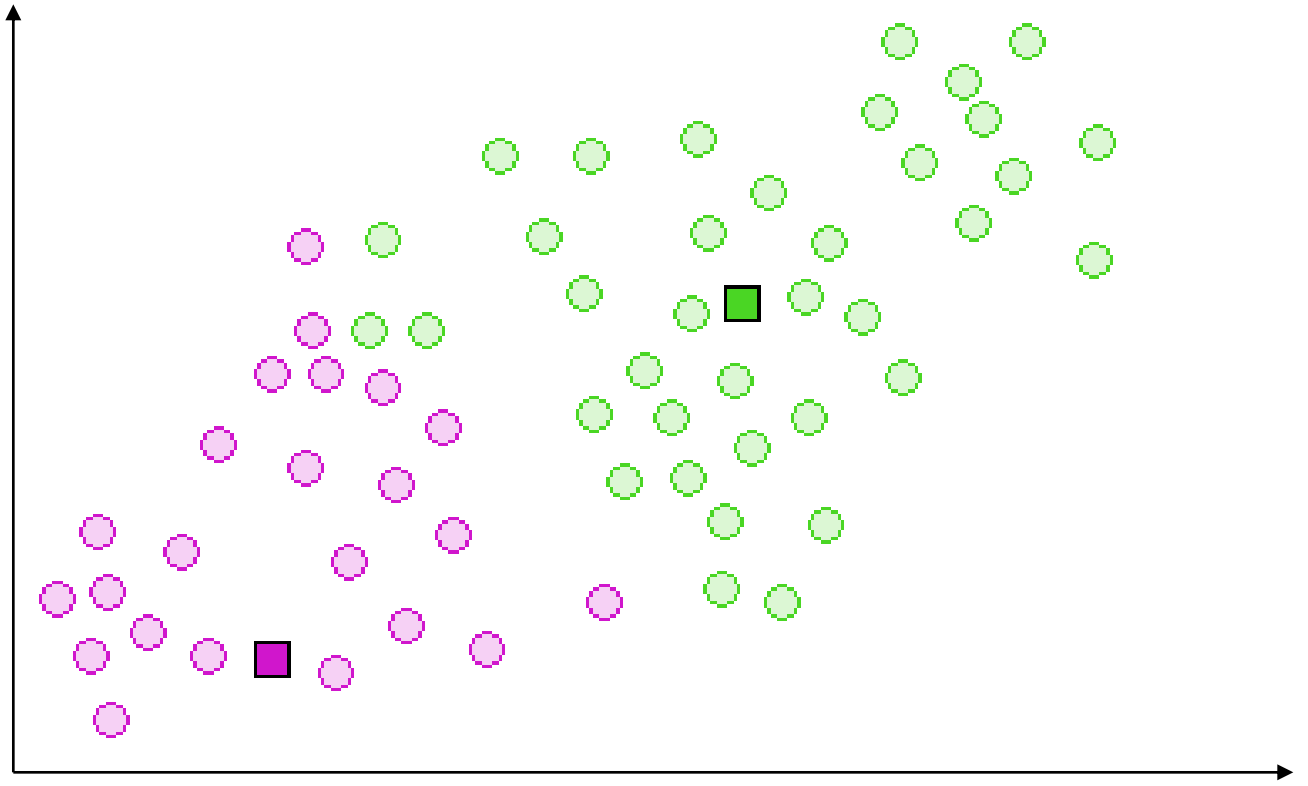
Objetos em um plano 2D

k-Means – Um Exemplo



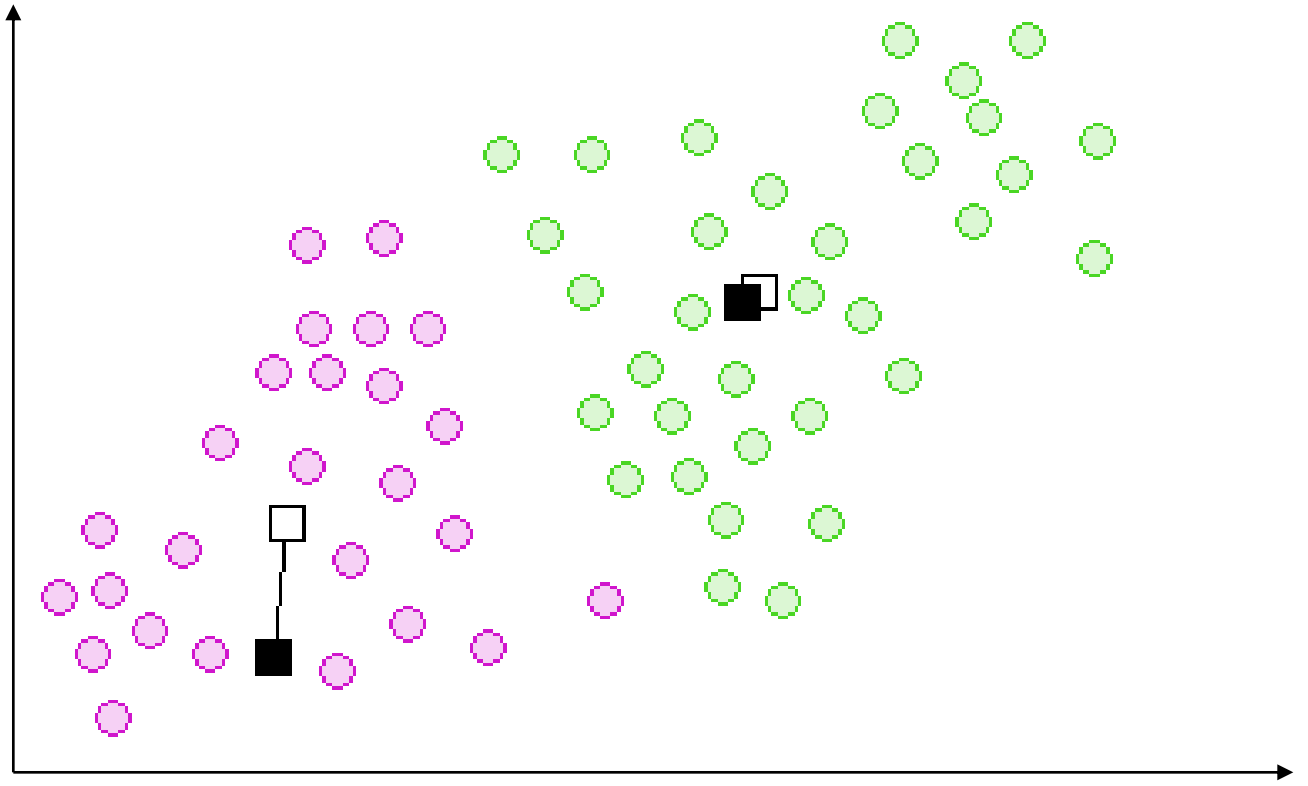
Passo 1:Centróides inseridos aleatoriamente

k-Means – Um Exemplo



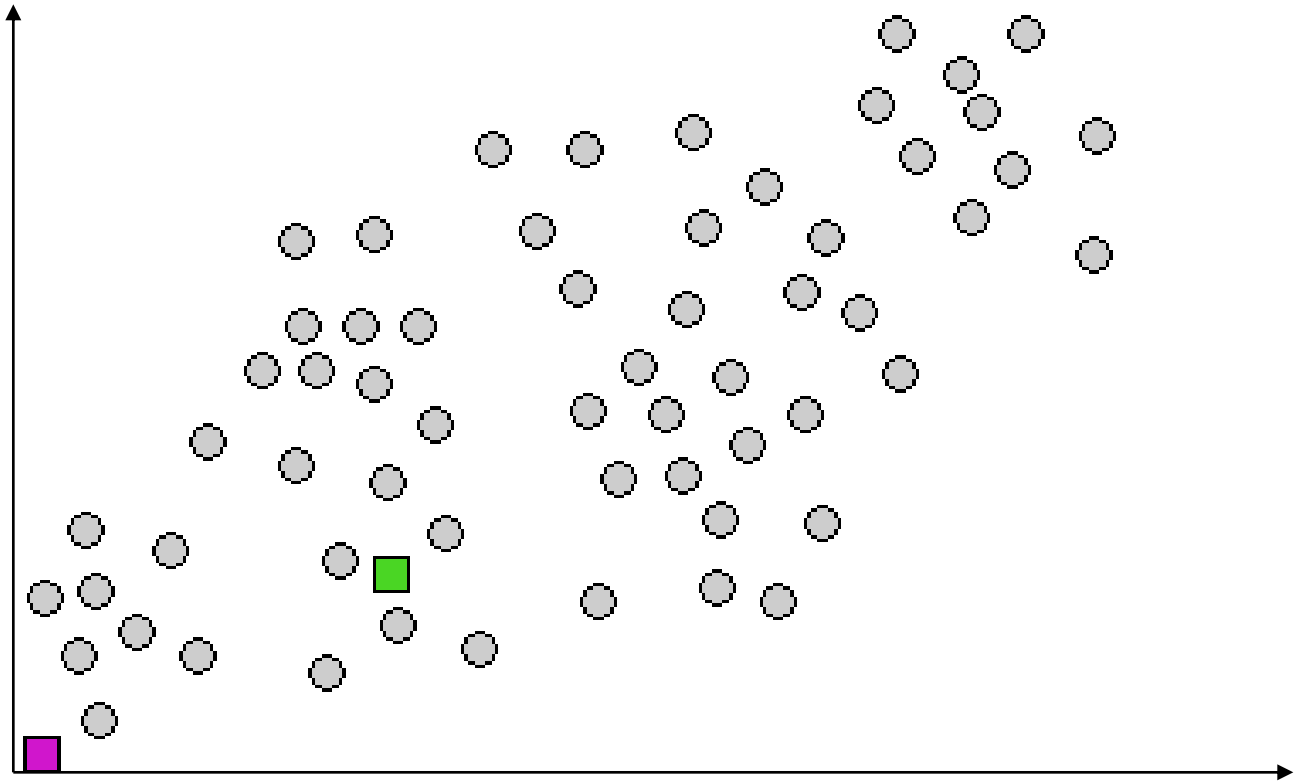
Passo 2: Atribuir a cada objeto o centróide mais próximo

k-Means – Um Exemplo



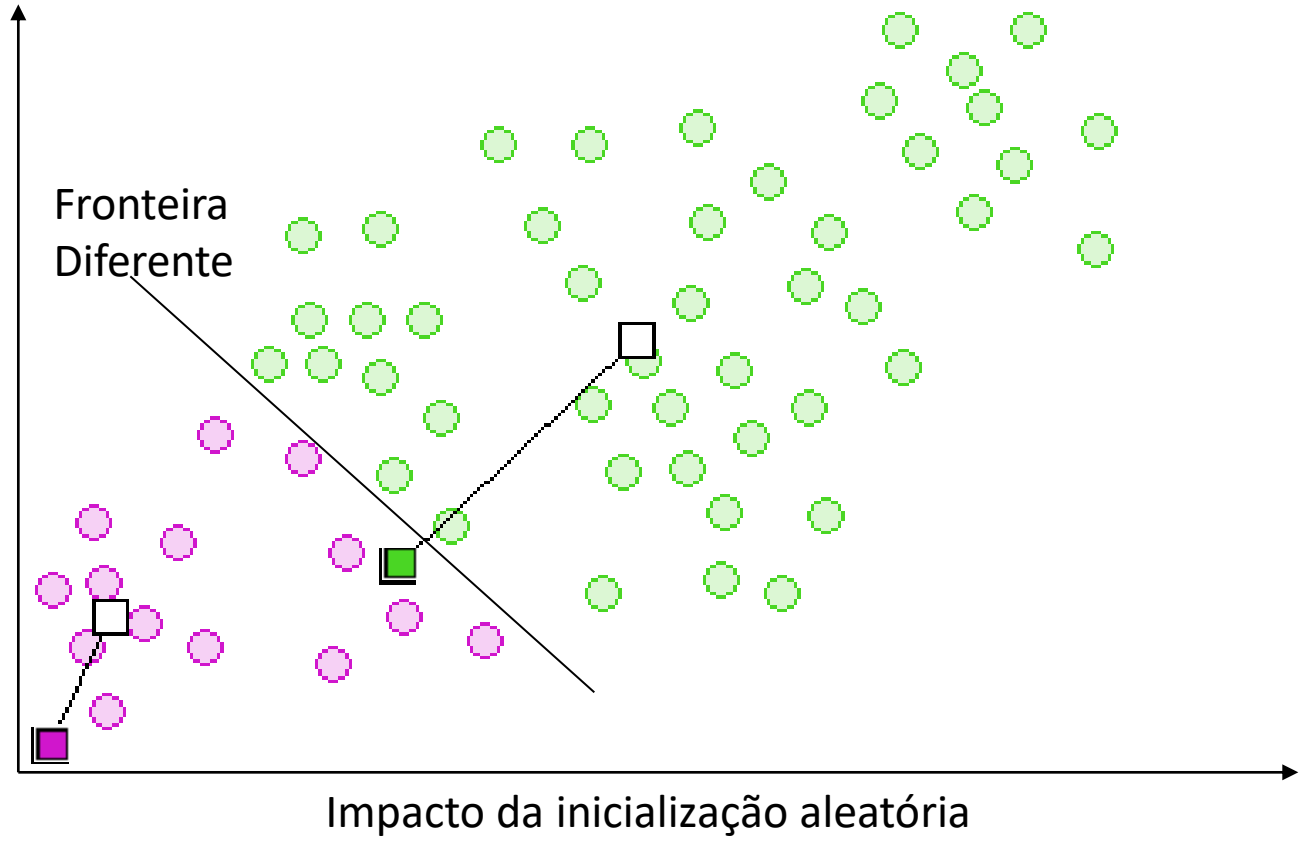
Passo 3: Recalcular os centróides

k-Means – Um Exemplo



Impacto da inicialização aleatória.

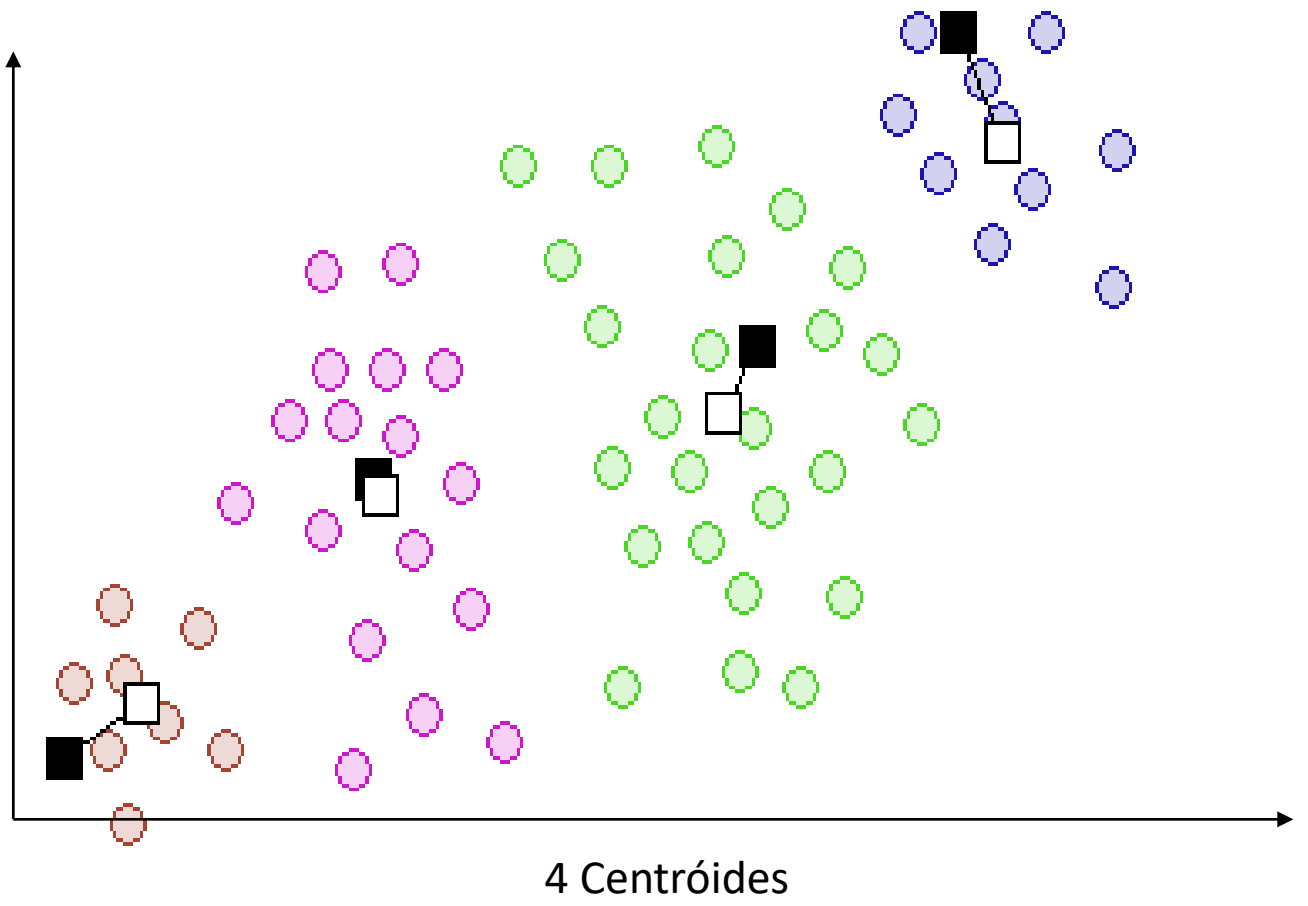
k-Means – Um Exemplo



k-Means – Inicialização

- Importância da inicialização.
- Quando se têm noção dos centróides, pode-se melhorar a convergência do algoritmo.
- Execução do algoritmo várias vezes, permite reduzir impacto da inicialização aleatória.

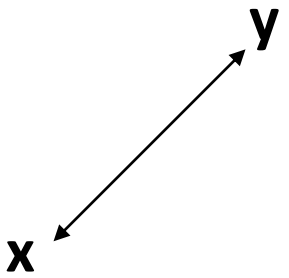
k-Means – Um Exemplo



Calculando Distâncias

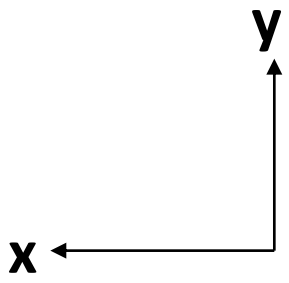
- Distância Euclidiana

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$



- Manhattan (City Block)

$$d = \sum_{i=1}^{n} |x_i - y_i|$$



Calculando Distâncias

- Minkowski
  - Parâmetro r
    - r = 2, distância Euclidiana
    - r = 1, City Block

$$d = \left(\sum_{i=1}^{n} \left(x_i - y_i \right)^r \right)^{1/r}$$

Normalização

- Considerando a distância Euclidiana, mais utilizada nas aplicações, um problema ocorre quando um dos atributos assume valores em um intervalo relativamente grande, podendo sobrepujar os demais atributos
  - V1 = {200, 0.5, 0.002}
  - V2 = {220, 0.9, 0.050}

  Se calcularmos a distância Euclidiana, veremos que a primeira característica dominará o resultado.

Normalização

- Portanto, as distâncias são frequentemente normalizadas dividindo a distância de cada atributo pelo intervalo de variação (i.e. diferença entre valores máximo e mínimo) daquele atributo.

- Assim, a distância para cada atributo é normalizada para o intervalo [0,1].

Normalização

- Diferentes técnicas de normalização

Min-Max

$$n_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Z-Score

$$n_i = \frac{x_i - mean(x)}{std(x)}$$

Tanh

$$n_i = \frac{1}{2}\left[\tanh\left(001\frac{x_i - mean(x)}{std(x)}\right)+1\right]$$

Soma

$$n_i = \frac{x_i}{\sum x}$$

k-means - Vantagens

- Simples e intuitivo
- Eficaz em muitos cenários de aplicação e produz resultados de interpretação simples
- Considerado um dos 10 mais influentes algoritmos em *DataMining*

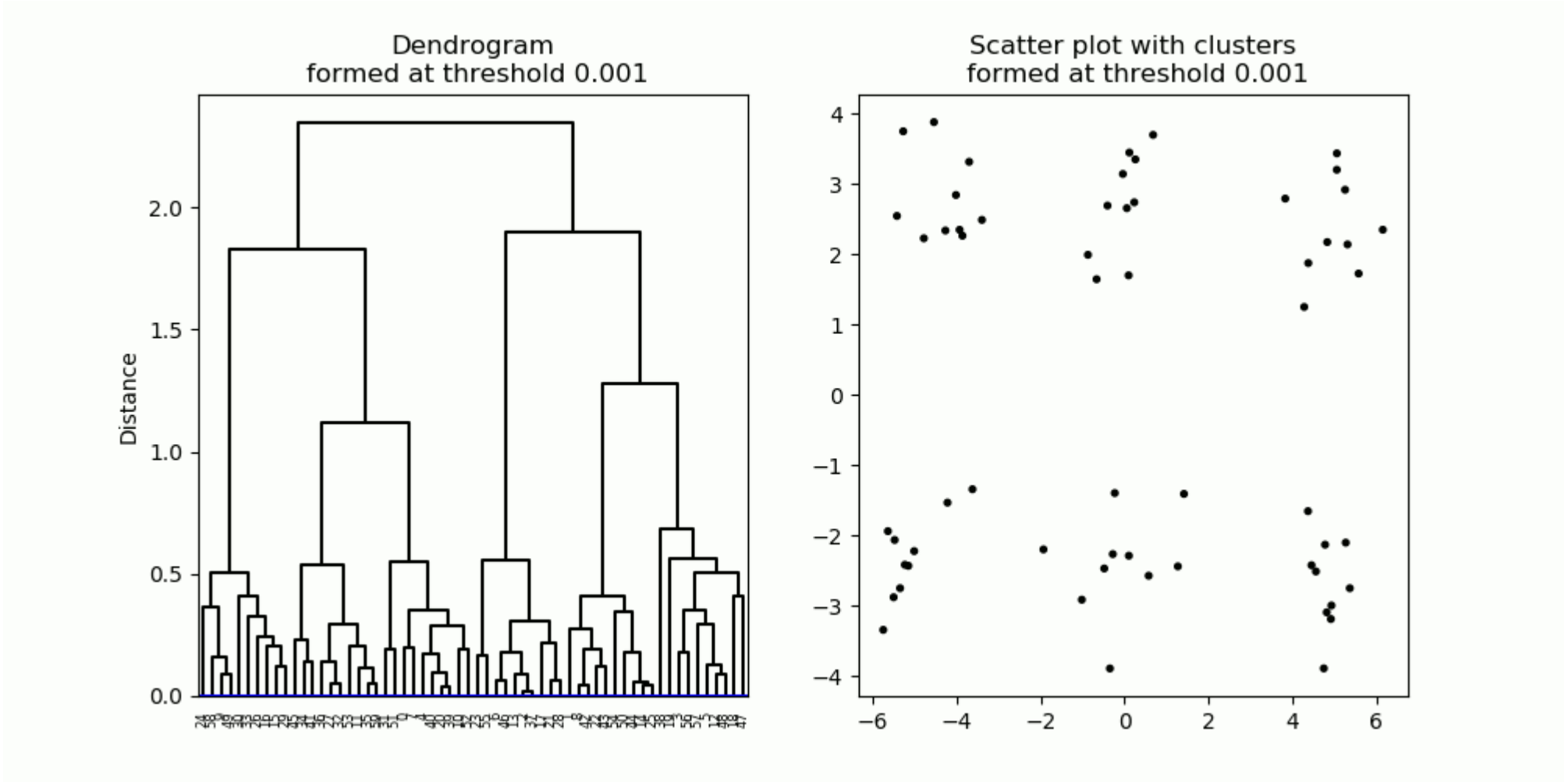
k-means - Desvantagens

- k = ?
  - *tem que saber a priori a quantidade de clusters*
- Sensível à inicialização dos centróides
- Limita-se a encontrar clusters volumétricos / globulares
- Cada item deve pertencer a um único cluster (partição rígida, ou seja, sem sobreposição)
- Limitado a atributos numéricos
- Sensível a *outliers* (valor atípico ou inconsistente)

Agrupamento hierárquico

- O algoritmo de clustering hierárquico funciona conectando iterativamente os pontos de dados mais próximos para formar clusters.
- Inicialmente todos os pontos de dados são desconectados uns dos outros; cada ponto de dados é tratado como seu próprio cluster.
- Em seguida, os dois pontos de dados mais próximos são conectados, formando um cluster.
- Em seguida, os dois pontos de dados mais próximos (ou clusters) são conectados para formar um cluster maior.
- E assim por diante. O processo é repetido para formar clusters progressivamente maiores e continua até que todos os pontos de dados estejam conectados em um único cluster.

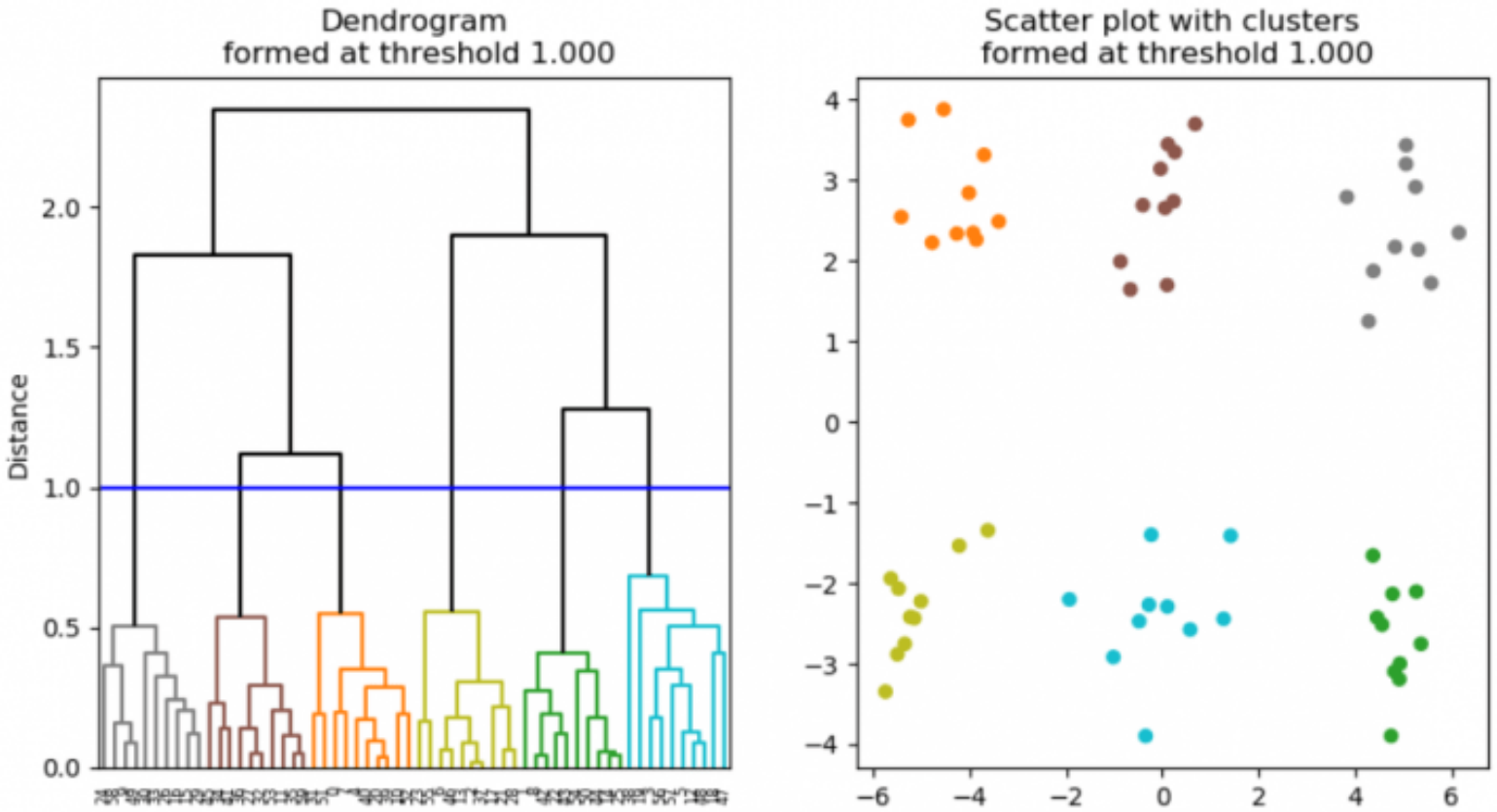
Agrupamento hierárquico



Agrupamento hierárquico

- O agrupamento hierárquico forma uma hierarquia de agrupamentos, descrita em um diagrama conhecido como dendrograma .
- Um dendrograma descreve quais pontos de dados / clusters estão conectados a que distância, começando de pontos de dados individuais na parte inferior até o único grande cluster na parte superior.
- Para obter uma partição de cluster com um determinado número de clusters, pode-se simplesmente aplicar um limite de corte a uma determinada distância no dendrograma, produzindo o número desejado de clusters.

Agrupamento hierárquico



DBSCAN

- DBSCAN significa Agrupamento Espacial Baseado em Densidade de Aplicativos com Ruído.
- É um método de agrupamento baseado em densidade, agrupando nuvens densas de pontos de dados em agrupamentos.
- Quaisquer pontos isolados não são considerados parte de clusters e são tratados como ruídos.

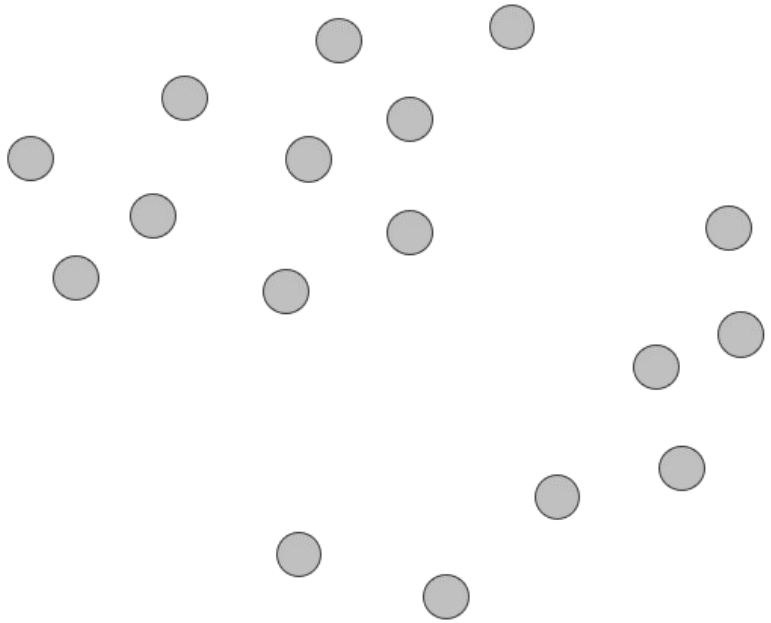
DBSCAN

- O algoritmo DBSCAN começa selecionando aleatoriamente um ponto de partida.
- Se houver um número suficientemente grande de pontos dentro da vizinhança ao redor desse ponto, então esses pontos são considerados como parte do mesmo cluster que o ponto de partida.
- As vizinhanças dos pontos recém-adicionados são então examinadas.
- Se houver pontos de dados dentro dessas vizinhanças, esses pontos também serão adicionados ao cluster.
- Esse processo é repetido até que não seja possível adicionar mais pontos a esse cluster específico.

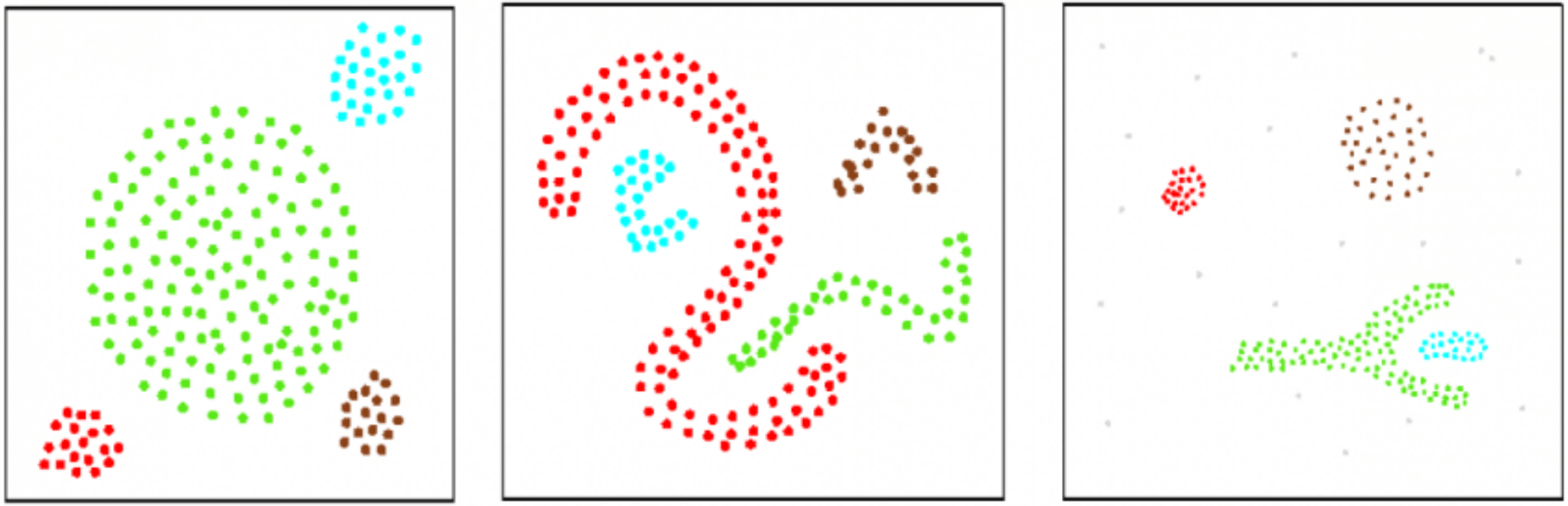
DBSCAN

- Em seguida, outro ponto é selecionado aleatoriamente como ponto de partida para outro cluster;
- O processo de formação do cluster é repetido até que não haja mais pontos de dados disponíveis para serem atribuídos aos clusters .
- Se os pontos de dados não estiverem na vizinhança de quaisquer outros pontos de dados, esses pontos de dados são considerados ruídos.
- Clusters de qualquer formato podem ser formados pelo algoritmo DBSCAN

DBSCAN



DBSCAN



Referências

- RUSSEL, S., NORVIG, P. *Inteligência Artificial*, Editora Campus, 2ª. edição.
- Knime: https://www.knime.com/blog/what-is-clustering-how-does-it-work