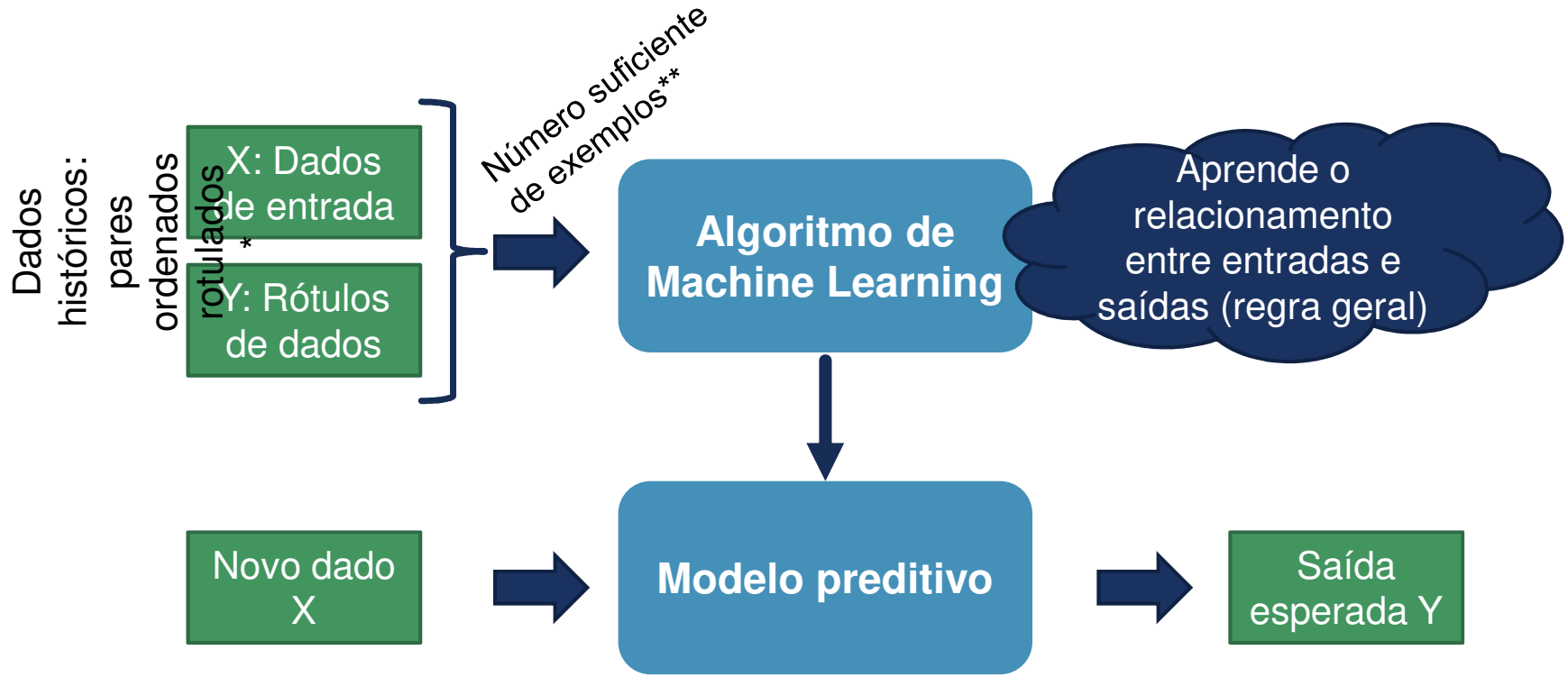
ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS – PUC-RIO
MACHINE LEARNING

AULA 6: PROBLEMAS DE CLUSTERIZAÇÃO

Tatiana Escovedo, PhD.
tatiana@inf.puc-rio.br

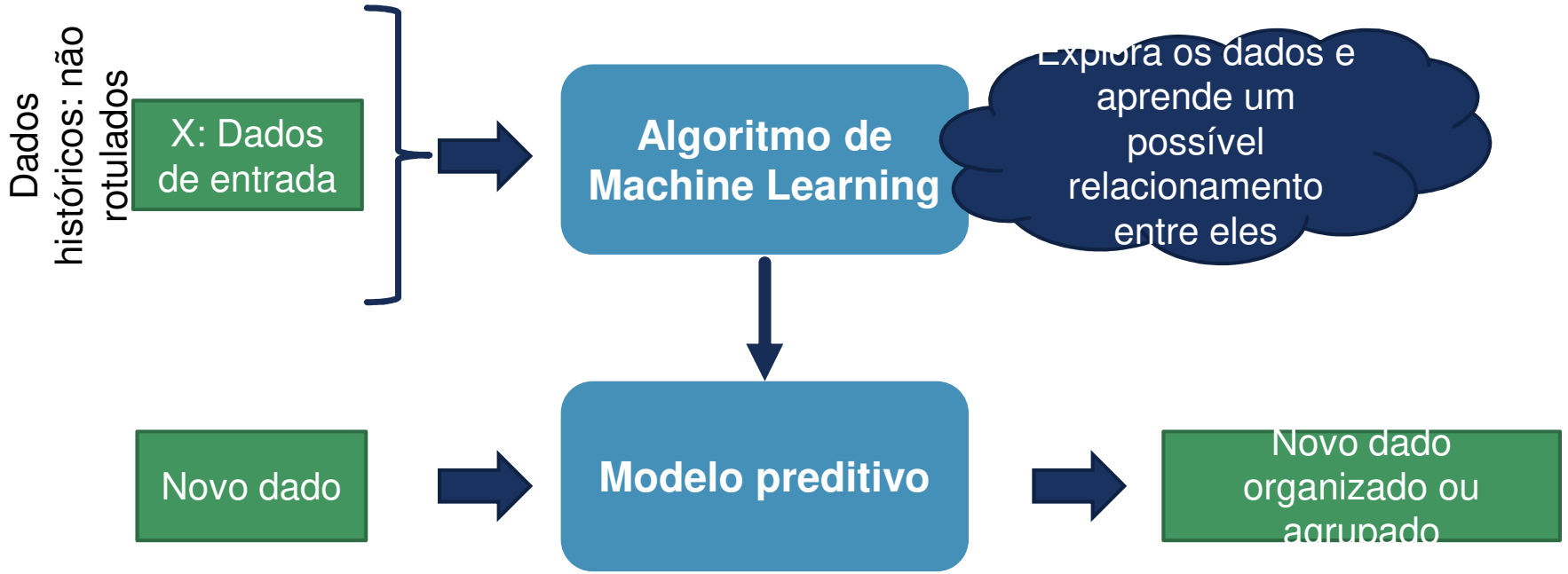
APRENDIZADO SUPERVISIONADO



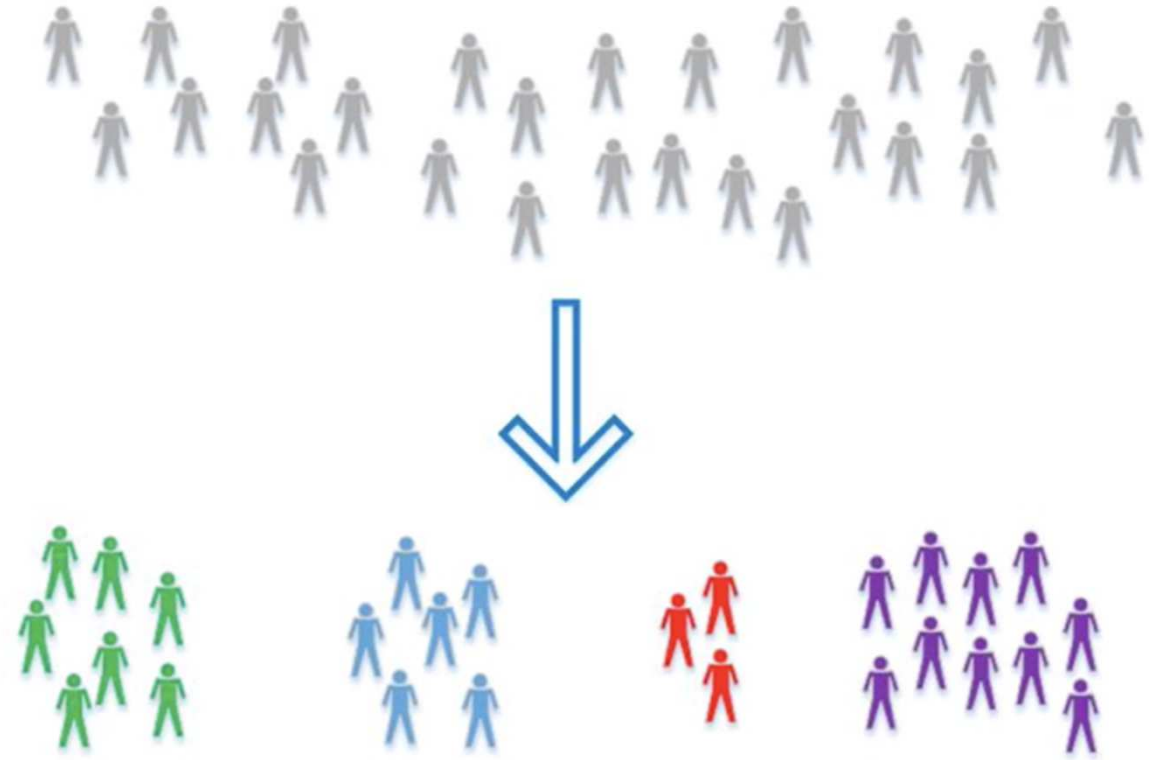
* X: atributos, características, atributos previsores ou de predição...
Y: atributo-alvo, target

**Instâncias, registros...

APRENDIZADO NÃO-SUPERVISIONADO



APRENDIZADO NÃO-SUPERVISIONADO - EXEMPLO



CLUSTERIZAÇÃO

- Também chamada de **Agrupamento**
- Separa os registros de um conjunto de dados em subconjuntos ou **clusters** (clusters) semelhantes e homogêneos, de tal forma que elementos em um cluster **compartilhem** um conjunto de **propriedades comuns** que os diferencie dos elementos de outros clusters.
- É um problema de otimização, em que o objetivo é **maximizar** a similaridade **intracluster** e **minimizar** a similaridade **intercluster**.
- Os objetos de entrada **não possuem rótulos** associados (não-supervisionado).

CLUSTERIZAÇÃO

- Os clusters são formados de acordo com alguma medida de similaridade: objetos pertencentes a um dado cluster devem ser **muito similares entre si** (compartilham um conjunto maior de propriedades comuns) e muito diferentes dos a objetos pertencentes a outros clusters.
- Alguns algoritmos requerem que o usuário forneça o **número de clusters a formar**, e há algoritmos de que tentam detectar a quantidade de clusters naturais existentes no conjunto de dados de entrada.
- A presença de dados distribuídos em um espaço de **grande dimensionalidade** dificulta a detecção de clusters.

CLUSTERIZAÇÃO - EXEMPLOS

- Segmentação de clientes para identificar padrões de compra para campanhas de marketing direcionadas.
- Detecção de comportamento anômalo, como intrusões de rede não autorizadas, identificando padrões de uso fora dos clusters conhecidos.
- Simplificação de grandes conjuntos de dados agrupando características com valores semelhantes em um número menor de categorias homogêneas.

ALGORITMOS DE CLUSTERIZAÇÃO MAIS POPULARES

The 5 Clustering Algorithms Data Scientists Need to Know:
https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

- **K-means Clustering**
  - https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- **Agglomerative Hierarchical Clustering**
  - https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

ALGORITMOS DE CLUSTERIZAÇÃO MAIS POPULARES

- **Mean-Shift Clustering**
  - https://spin.atomicobject.com/2515/05/26/mean-shift-clustering/
  - https://towardsdatascience.com/machine-learning-algorithms-part-13-mean-shift-clustering-example-in-python-4d6452720b00
  - https://scikit-learn.org/stable/modules/generated/sklearn.cluster.MeanShift.html
- **Density-Based Spatial Clustering of Applications with Noise (DBSCAN)**
  - https://towardsdatascience.com/how-dbscan-works-and-why-should-i-use-it-443b4a191c80
  - https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html
- **Expectation-Maximization (EM) Clustering using Gaussian Mixture Models (GMM)**
  - https://towardsdatascience.com/gaussian-mixture-models-d13a5e915c8e
  - https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html

FAMÍLIAS DE MÉTODOS

- **Classificação 1:**
  - Baseados em distâncias
    - *K-means, K-modes, K-medoids*
  - Baseados em densidade
    - *DBSCAN, Mean-Shift*
  - Baseados em distribuições de probabilidades
    - *EM*

- **Classificação 2:**
  - Partitivos
  - Hierárquicos
    - Aglomerativos
    - Divisivos

ALGORITMOS PARTITIVOS

Dividem o conjunto de dados em ***k*** clusters.

Produzem agrupamentos simples, tentando fazer os ***k*** clusters tão compactos e separados quanto possível.

Funcionam bem quando os clusters são compactos, densos e bastante separados uns dos outros.

São efetivos se o valor de ***k*** puder ser razoavelmente estimado e se os clusters possuírem forma convexa e tamanho e densidade similares.

Quando existem grandes diferenças nos tamanhos e geometrias dos diferentes clusters, podem dividir desnecessariamente grandes clusters para minimizar a distância total calculada.

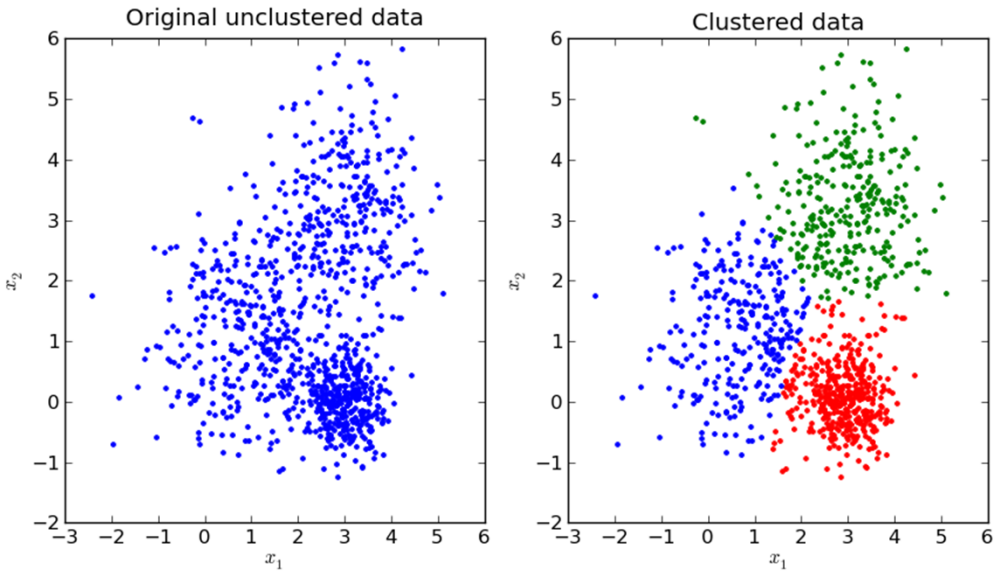
ALGORITMOS PARTITIVOS

Funcionamento:

- Inicialmente, ***k*** objetos são escolhidos como os centros dos ***k*** clusters.
- Os objetos são divididos entre os ***k*** clusters de acordo com a medida de similaridade adotada, de modo que cada objeto fique no cluster que forneça o menor valor de distância entre o objeto e o centro do cluster.
- Uma estratégia iterativa determina se os objetos devem mudar de cluster, fazendo com que cada cluster contenha somente os elementos mais similares entre si.
- Após a divisão inicial, há duas possibilidades na escolha do “elemento” que vai representar o centro do cluster, e que será a referência para o cálculo da medida de similaridade:
  - Média dos objetos que pertencem ao cluster (centróide ou centro de gravidade do cluster).
  - O objeto que se encontra mais próximo ao centro de gravidade daquele cluster (medoide).

K-MEANS

- Considera que os registros do conjunto de dados correspondem a pontos no R^n, em que cada atributo corresponde a uma dimensão deste espaço.
- Parâmetro de entrada: ***k*** (quantidade de clusters a ser identificados)

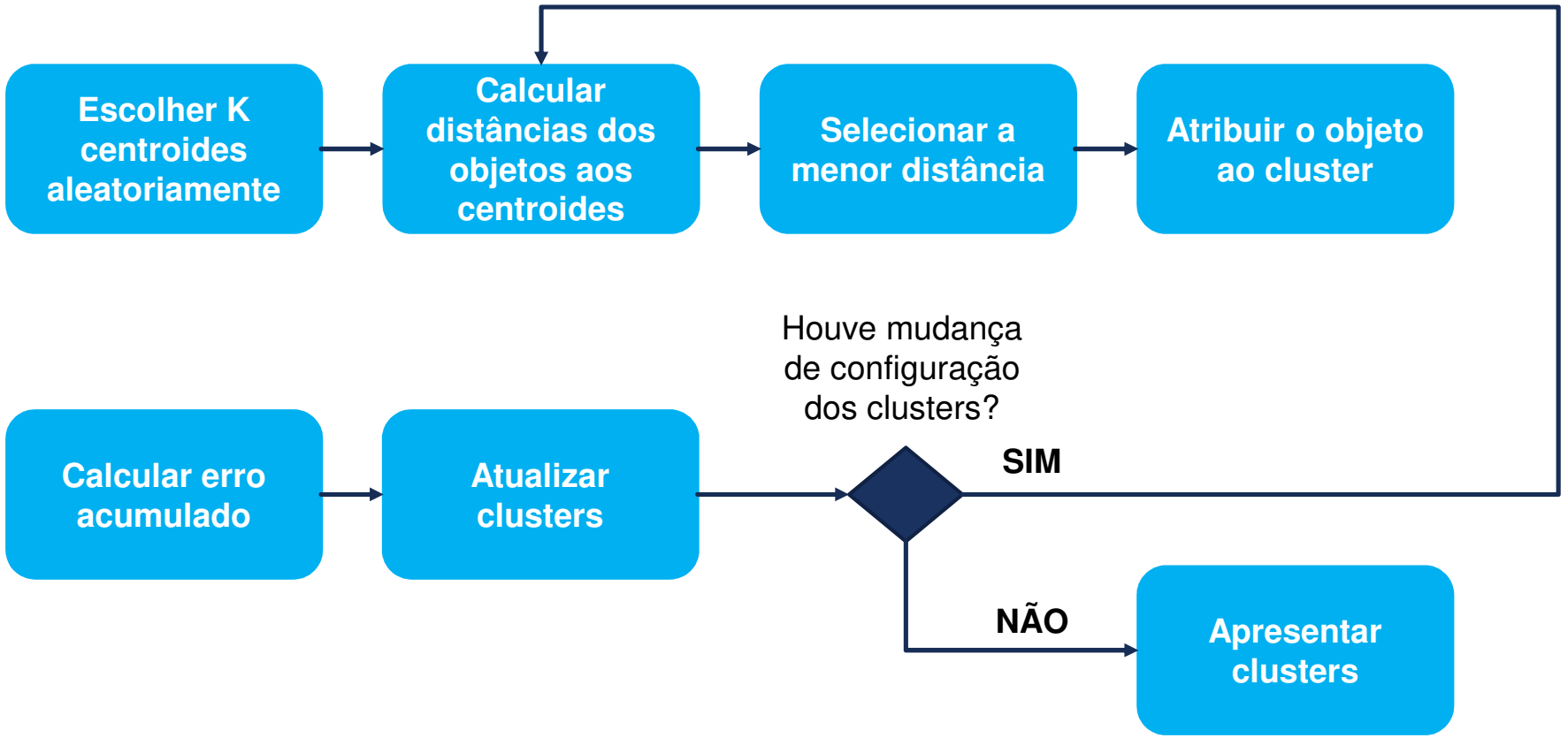


K-MEANS

Funcionamento:

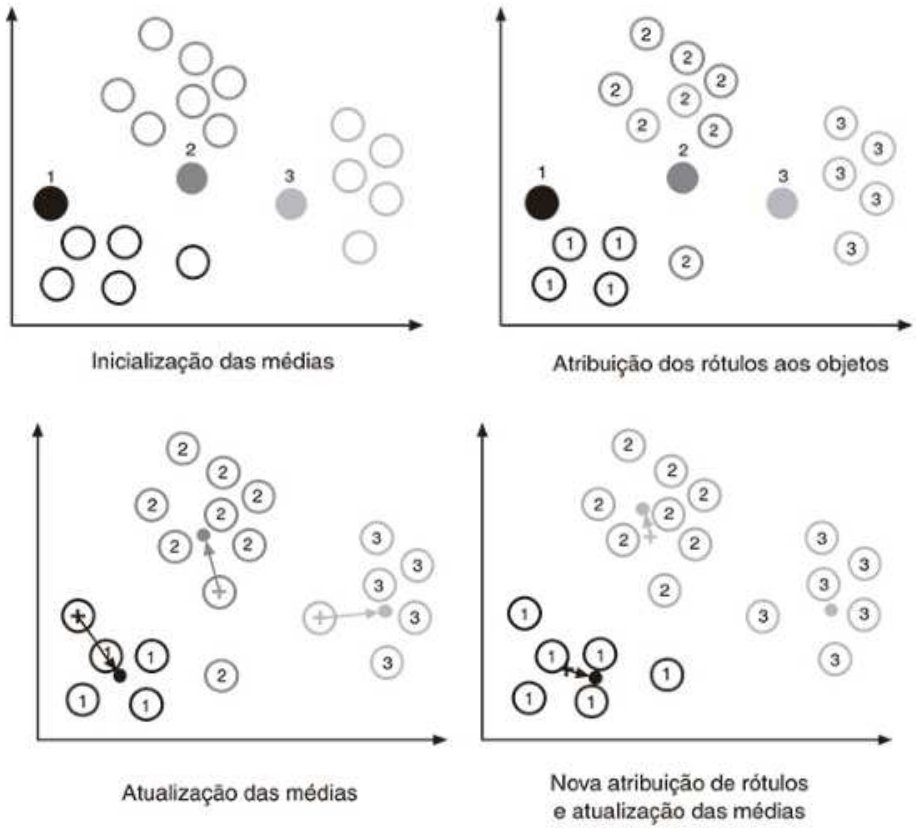
- Seleciona ***k*** pontos do conjunto de dados (**sementes**), que são os representantes iniciais, ou centroides, dos ***k*** clusters a ser formados.
- Para cada ponto, calcula-se a distância euclidiana deste ponto a cada um dos centroides e atribui-se este ponto ao cluster representado pelo centroide cuja distância é a menor entre todas as calculadas. Cada ponto do conjunto de dados fica associado a um e apenas um dos ***k*** clusters.
- Após a alocação inicial, o método segue iterativamente, por meio da atualização dos centroides de cada cluster e da realocação dos pontos ao centroide mais próximo. O novo centroide de cada cluster ***G*** é calculado pela média dos pontos alocados a ***G***.
- O processo iterativo termina quando os centroides dos clusters param de se modificar, ou após um número preestabelecido de iterações ter sido realizado.

K-MEANS


Escolher K centroides aleatoriamente
Calcular distâncias dos objetos aos centroides
Selecionar a menor distância
Atribuir o objeto ao cluster
Houve mudança de configuração dos clusters?
SIM
NÃO
Calcular erro acumulado
Atualizar clusters
Apresentar clusters

K-MEANS

1. Inicialmente, a sementes são selecionadas de forma aleatória.
2. Cada ponto restante é alocado a algum cluster, em função de sua distância a cada um dos centroides.
3. Os centroides são atualizados.
4. Ocorre nova realocação de pontos.
5. O processo continua até a convergência.



K-MEANS

- O K-means divide um conjunto de ***n*** objetos em ***k*** clusters tal que a similaridade ***intraclusters*** resultante seja alta, mas a similaridade ***interclusters*** seja baixa.
  - A similaridade em um cluster é a média dos pontos alocados neste cluster (seu centro de gravidade).
- Isso é equivalente a determinar uma partição de tamanho ***k*** que minimize a função do erro quadrático médio (*mean squared error,* ***MSE***).
  - Dado o conjunto de clusters ***{Gi}*** de uma determinada iteração do K-means, e seu conjunto de centroides ***{mi}***, o K-means calcula o MSE.
- O **propósito** do K-means é determinar um agrupamento para o qual o valor de MSE seja mínimo.

K-MEANS

Vantagens:

- Utiliza princípios simples, que podem ser explicados em termos não-estatísticos.
- Apresenta bom desempenho quando os clusters são densos, compactos e bem separados uns dos outros.
- Rápido e fácil de implementar, podendo ser customizado.

K-MEANS

Desvantagens:

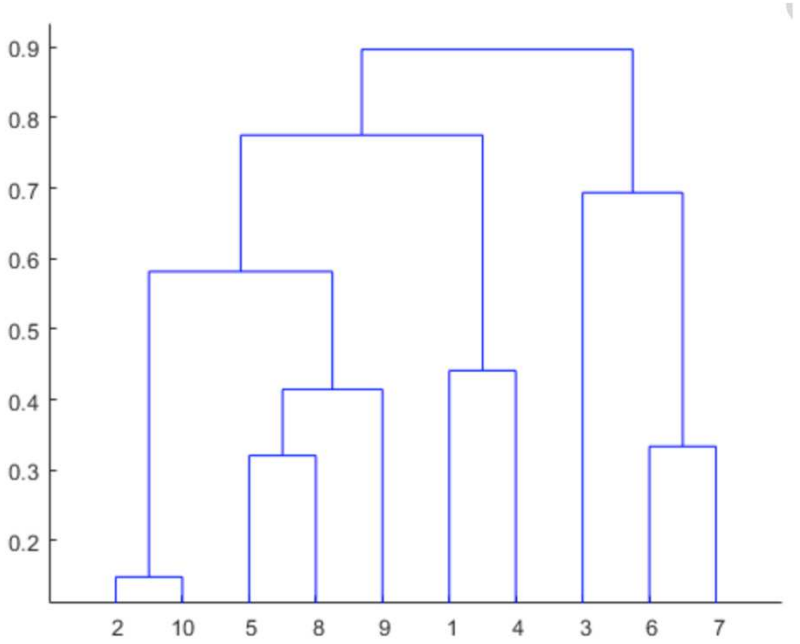
- Necessidade de especificar previamente ***k*** (número de clusters). Diversos experimentos variando o valor de k devem ser realizados para determinar o valor adequado.
- Não é adequado para descobrir clusters com formas não convexas, de tamanhos muito diferentes e com sobreposição.
- É muito sensível à existência de ruídos no conjunto de dados, visto que pequeno número de dados ruidosos pode influenciar substancialmente os valores médios dos clusters.
- Não é garantido encontrar o conjunto ideal de clusters, por utilizar elementos aleatórios.

K-MODES E K-MEDOIDS

- Também têm como objetivo particionar os registros de dados, agrupando-os por similaridade.
- O **K-Modes** é uma variação do método K-means, com a diferença de ser utilizado para agrupamento de dados **categóricos** (variáveis nominais). Em geral, no lugar do cálculo da média, calcula-se a moda dos objetos. Usa medidas de similaridade para tratar objetos categóricos, além de usar técnicas baseadas em frequência para atualizar as modas dos clusters.
- O **K-Medoids** concentra-se, primeiramente, em encontrar o ***medoid*** (mediana). Os objetos restantes são agrupados com o medoid ao qual eles são mais similares. Há uma troca iterativa, de um medoid por um não medoid, visando à melhoria do agrupamento. A qualidade é estimada usando uma função custo que mede a similaridade média entre os objetos e o medoid de seu cluster.

ALGORITMOS HIERÁRQUICOS

- Criam uma decomposição hierárquica do conjunto de dados, representada por um **dendrograma:**
  - Uma árvore que iterativamente divide o conjunto de dados em subconjuntos menores até que cada subconjunto consista de somente um objeto.
- Cada **nó** da árvore representa um **cluster** do conjunto de dados e a **união** dos clusters em um determinado nível da árvore corresponde ao cluster no nível exatamente acima.



ALGORITMOS HIERÁRQUICOS ALGOMERATIVOS

Abordagem aglomerativa (bottom-up): parte-se das folhas para a raiz.

- Inicialmente, coloca-se cada um dos ***n*** objetos em seu próprio cluster (cada objeto representa um cluster separado), totalizando ***n*** clusters.
- Em cada etapa, calcula-se a distância entre cada par de clusters, e as distâncias são armazenadas em uma **matriz de similaridade**.
- São escolhidos dois clusters com a distância mínima, juntando-os para formar um único cluster.
- Este processo continua até que todos os objetos estejam em um único cluster (o nível mais alto da hierarquia), ou até que uma condição de término ocorra (por exemplo, o número de clusters desejado tenha sido alcançado).

ALGORITMOS HIERÁRQUICOS DIVISIVOS

Abordagem divisiva (top-down): parte-se da raiz para as folhas.

- Inverte-se o processo, começando com todos os objetos em um **único** cluster.
- Em cada etapa, um cluster é escolhido e dividido em dois menores.
- Este processo continua até que se tenha ***n*** clusters, ou até que uma condição de término seja satisfeita.

ESTRUTURAS DE DADOS

Algoritmos de Agrupamento normalmente utilizam uma das seguintes estruturas de dados no seu processamento:

- **Matriz de Dados:** as linhas representam cada um dos objetos a serem agrupados; as colunas representam os atributos (ou características) de cada objeto.
- Considerando n objetos, cada qual com p atributos, obtém-se uma matriz ***n x p***:

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2p} \\ x_{31} & x_{32} & x_{33} & \cdots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{bmatrix}$$

ESTRUTURAS DE DADOS

- **Matriz de Similaridade:** cada elemento da matriz representa a distância entre pares de objetos. Como a distância entre i e j é igual à distância entre j e i, não é necessário armazenar todas as distâncias entre os objetos.
- Considerando ***n*** objetos a serem agrupados, obtém-se uma matriz quadrada de tamanho ***n x n***, onde o ponto ***d(i, j)*** representa a distância ou similaridade entre o objeto ***i*** e o ***j***.
- Como as medidas de similaridade expressam o conceito de distância, estas são sempre números positivos. Quanto mais próximo de zero for ***d(i, j)***, mais similares serão os objetos.

$$D = \begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \\ d(n,1) & d(n,2) & d(n,3) & \dots & 0 \end{bmatrix}$$

Quando um algoritmo que trabalha com matrizes de similaridade recebe uma matriz de dados, ele primeiro a transforma em uma matriz de similaridade antes de iniciar o processo de Agrupamento.