Big Data Analytics com R

Agrupamentos

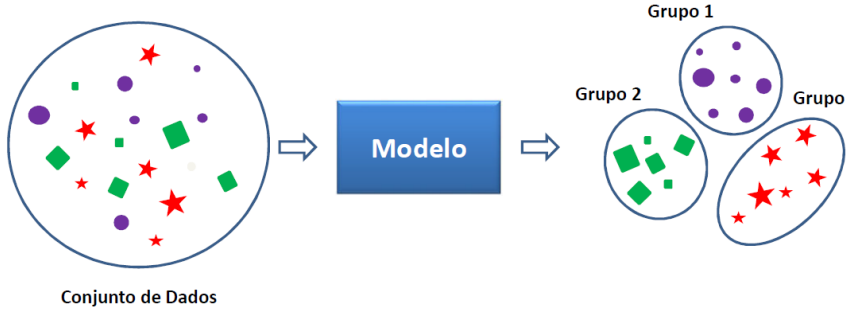
Agenda

- Implementar análise de segmentação de dados e análise de cluster utilizando R;
- Entender os resultados da clusterização utilizando R;
- Compreender os parâmetros opcionais para as análises de segmentação de dados e análise de cluster utilizando R;
- Implementar saída dos resultados da segmentação utilizando R.

Agrupamentos (Clustering)

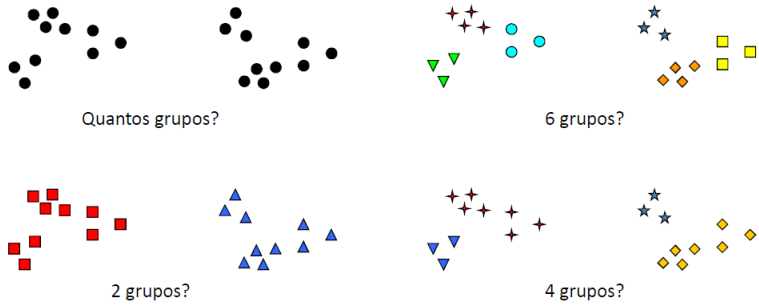
Definição

- O objetivo da análise de agrupamentos (ou segmentação) é encontrar grupos de objetos similares no conjunto de dados.



Análise de agrupamentos

- Um bom agrupamento tem grupos densos e separados entre si. Na prática, a noção de "grupo" é relativa e depende da aplicação.



Análise de agrupamentos

- O problema de análise de agrupamentos é diferente do problema de classificação:
  - No problema de classificação, a informação sobre as classes é externa.
  - Na análise de agrupamentos, a informação sobre os grupos é interna.

Métodos

- Métodos de Particionamento:
  - Os algoritmos constroem, a cada iteração, uma partição do conjunto de dados e a avaliam por um critério.
  - O critério mais comum é a soma das distâncias de todos os registros aos centros de grupos.
- Métodos Hierárquicos:
  - Os algoritmos geram uma decomposição hierárquica do conjunto de dados, através de estratégias “divisivas” (“top-down”) ou aglomerativas (“bottom-up”).
- Métodos a base de densidades:
  - Os grupos são gerados a partir da densidade dos registros.

Principais funções de distância

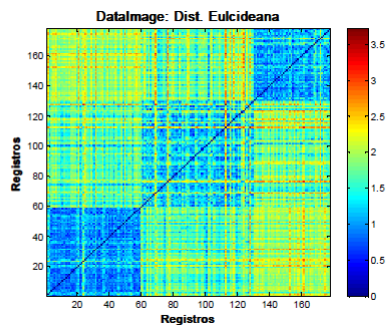
- Distância Euclidiana.
- Distância de Minkowski.
- Distância Manhattan.

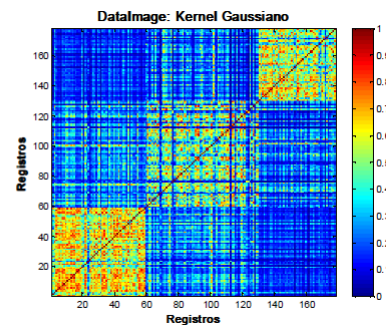
Principais funções de similaridade

- Similaridade Gaussiana.
- Similaridade do cosseno.

Exemplo

Wine





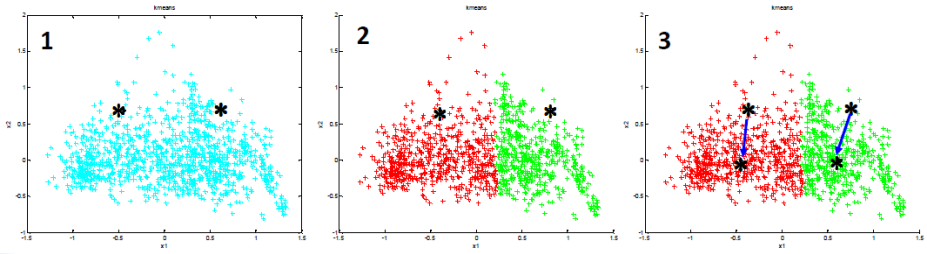
K-médias
(*K-means*)

K-médias

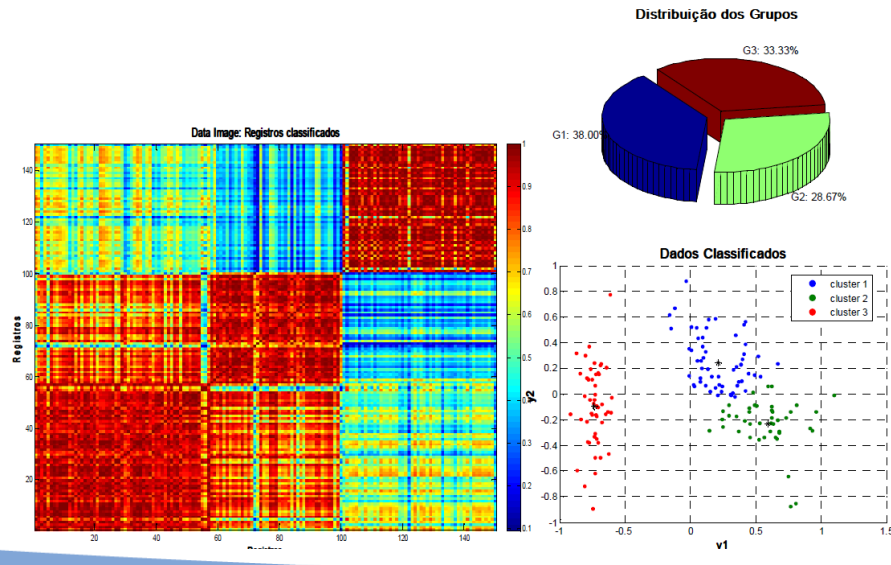
- É um algoritmo baseado em princípios fundamentais da teoria de conjuntos.
- O algoritmo k-médias busca encontrar partições no conjunto de dados (os chamados agrupamentos).

K-médias

- Em cada iteração, o algoritmo k-médias:
  1. Calcula a distância de cada registro aos centros de grupo;
  2. Aloca cada registro ao grupo cujo centro é mais próximo;
  3. Atualiza as coordenadas do centro de cada grupo pela média dos registros alocados ao grupo.



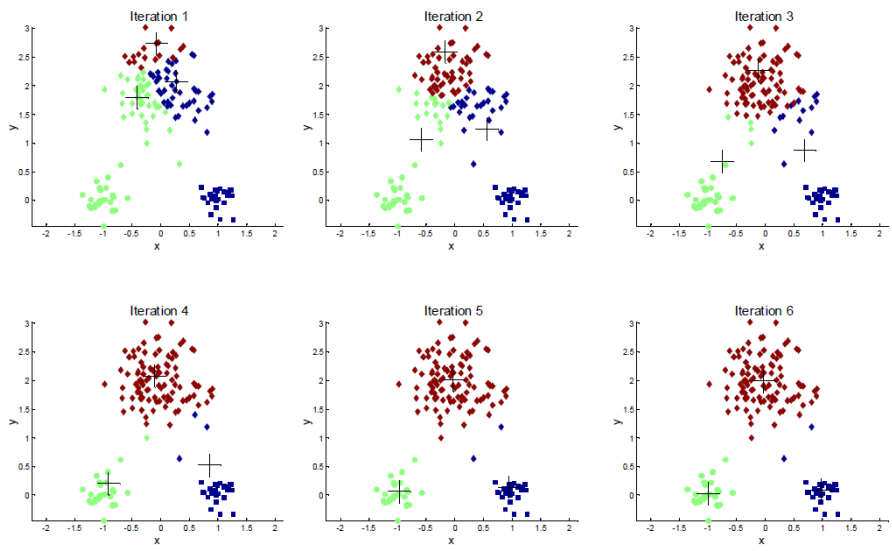
Exemplo: Flores Íris



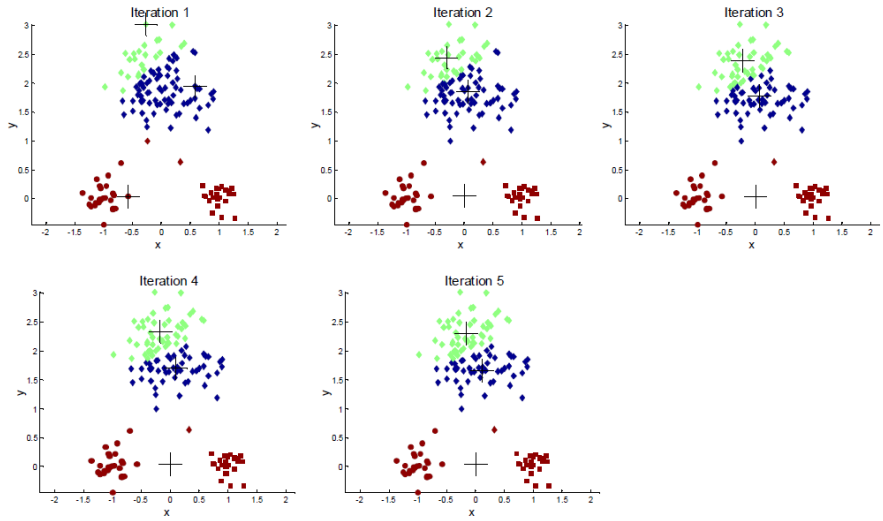
Observações

- As alternativas de inicialização:
  - Amostra do conjunto de dados escolhida aleatoriamente
  - Pontos sorteados aleatoriamente
  - Gerar um conjunto inicial a partir de um outro algoritmo determinístico.
- O algoritmo pode gerar um grupo vazio durante o processo. Neste caso, as alternativas mais comuns são:
  - Continuar o processo com K-1 grupos
  - Sortear um novo centro
  - Criar um novo grupo com o centro mais distante da iteração anterior.

Exemplo: inicialização funciona



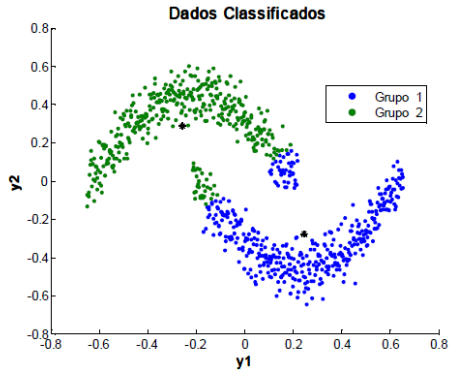
Exemplo: inicialização não funciona

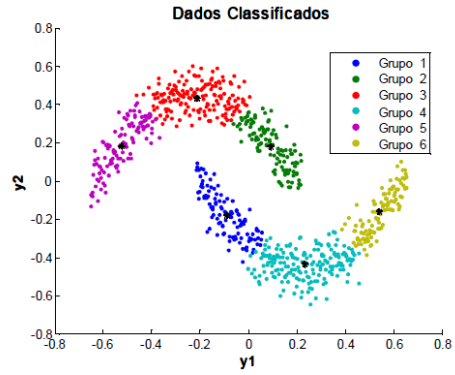


Limitações: formatos dos grupos

- Uma das maiores dificuldades do algoritmo é que a quantidade de grupos deve ser definida à priori.
- É possível usar o algoritmo para encontrar um número maior de grupos e posteriormente diminuir esta quantidade até um ponto que seja interpretável.
- O exemplo a seguir mostra que não foi possível identificar os dois grupos com K=2, porém ao usarmos K=6, isso torna-se possível pela agregação de grupos.

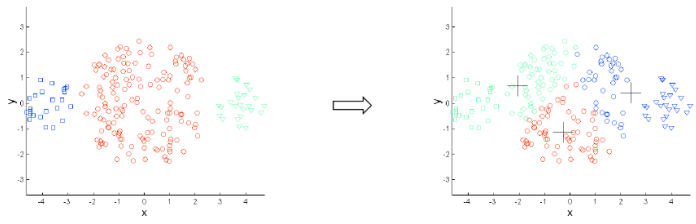
Limitações: formatos dos grupos



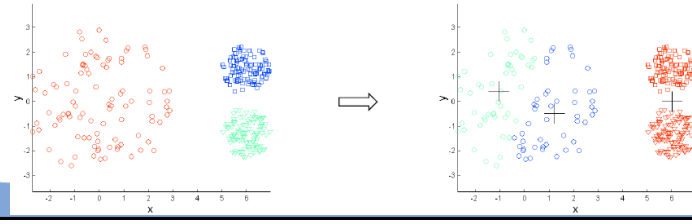


Limitações

- Tamanhos diferentes



- Densidades diferentes



Agrupamento Hierárquico (Hierarchical Clustering)

Métodos hierárquicos

- No algoritmo “K-médias” precisamos especificar “*a priori*” o número de agrupamentos desejados e encontrar o número ótimo de agrupamentos pode ser uma tarefa das mais difíceis.
- O algoritmo chamado de agrupamento hierárquico pode ser uma boa opção neste caso.
- Ele cria uma hierarquia por meio da técnica “*bottom-up*” e não necessita de uma definição prévia do número de agrupamentos.

O algoritmo

- Posicionar cada observação dentro de seu próprio agrupamento.
- Identificar os agrupamentos mais próximos e agrega-los em um agrupamento único.
- Repetir os passos acima até que todas as observações estejam em um agrupamento único.
- Ao terminar, apresentar o resultado no formato de dendrograma (do grego "*dendro*" (árvore) e "*gramma*" (desenho)).

Critérios de proximidade

- Existem diferentes formas de definir o quão próximos estão dois agrupamentos:
  - Pela maior distância possível entre pontos que pertençam a dois diferentes agrupamentos (*complete linkage clustering*);
  - Pela menor distância possível entre pontos que pertençam a dois diferentes agrupamentos (*single linkage clustering*);
  - Pela média de todas as distâncias entre pares de pontos pertencentes a dois diferentes agrupamentos (*mean linkage clustering*);
  - Pelo cálculo da distância entre centroides de dois agrupamentos (*centroid linkage clustering*).
- Os métodos *complete linkage* e *mean linkage* são os mais utilizados.