Análise de Agrupamento no R

Parte 2

Débora Silva
Lucas Bianchi

Breve recapitulação do conceito

Um dos principais desafios ao analisar um conjunto de dados é identificar padrões.

Uma abordagem simples para isso é encontrar características que permita agrupar um grande conjunto de dados em grupos menores.

Assim, um dos métodos utilizados é a análise de agrupamentos.

A idéia básica é: indivíduos pertencentes ao mesmo grupo são mais semelhantes que indivíduos pertencentes a grupos diferentes

Breve recapitulação do conceito

Na aula anterior foi visto que através de algumas **medidas de distância** e um **método estatístico** é possível definirmos grupos.

Algumas distâncias mais conhecidas são:

- Euclidiana, Mahalanobis, City-block (Manhattan) e Chebyschev

E os métodos estatísticos mais conhecidos são:

- Hierárquico: vizinhos mais próximos, mais distantes e ligação média

Breve recapitulação do conceito

Análise de agrupamento hierárquico

- O número de grupos não é definido antes da análise
- Pode ser aglomerativo ou divisivo

Análise de agrupamento não-hierárquico

- O número de grupos é definido antes da análise. Essa definição pode ser via a um conhecimento a priori ou usando alguma técnica (semente).

Passo-a-passo

Análise de agrupamento hierárquico

1. Definir a medida de dissimilaridade
2. Formação de agrupamentos
3. Definição de número de grupos

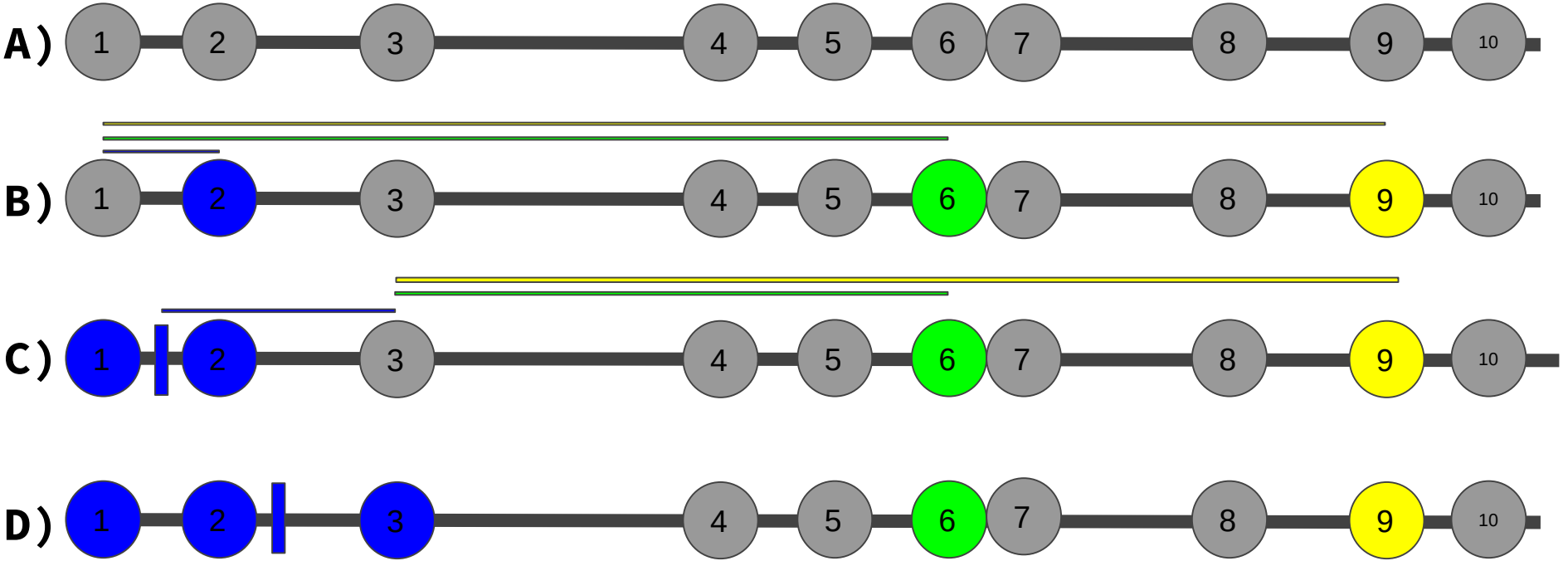
Análise de agrupamento não-hierárquico

1. Definir o número de grupos k
2. Selecione k diferentes pontos aleatoriamente
3. Calcule as distâncias de cada ponto até os pontos selecionados aleatoriamente
4. Calcule a média de cada grupo
5. Repita o passo 3 calculando a distância até a media do grupo obtido no passo 4.
6. Repita o processo até todos os pontos terem sido agrupados.

K-means

- É uma técnica de clusterização que visa separar o conjunto de dados em k grupos.
- O método visa encontrar centróide mais próximo e atribuir o ponto encontrado a esse cluster.
- Após este passo, os centróides são atualizados sempre tomando o valor médio de todos os pontos naquele cluster e aprimora de forma iterativa seus resultados até alcançar um resultado final.

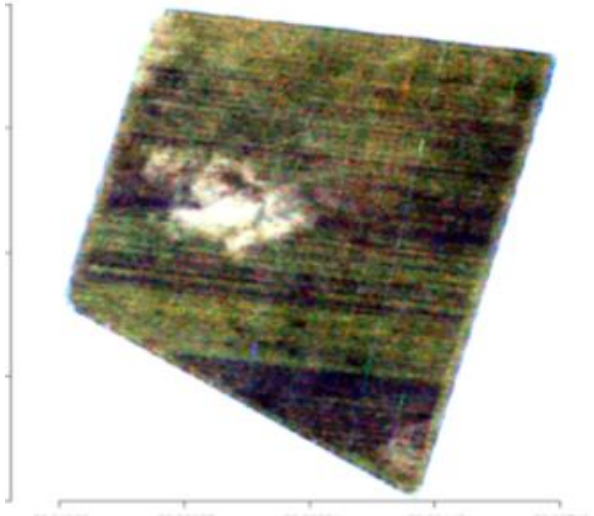
Exemplo 1: K-means

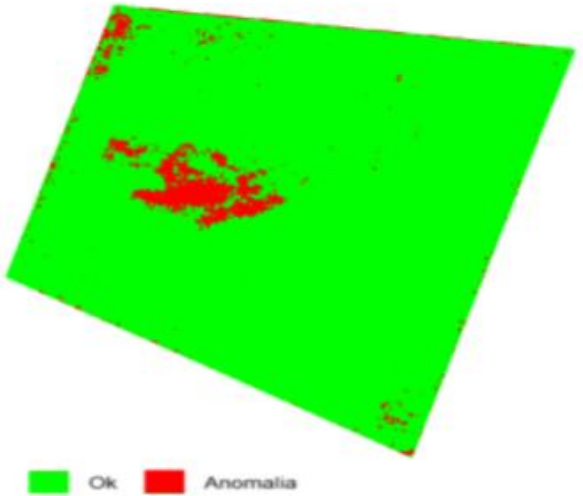


Exemplo 2: K-means (aplicação real)

Aplicação: Detecção de nuvens em imagens de satélite

Foi utilizado o algoritmo k-means com k = 2, visando agrupar as informações das bandas RGB e NIR para identificar nuvens.

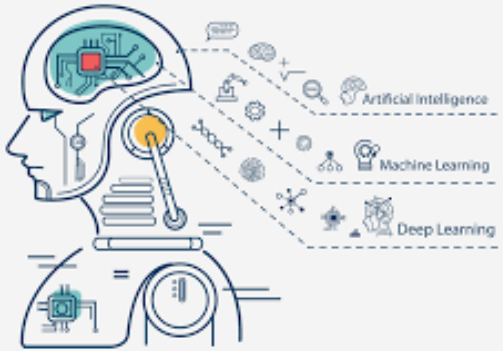




Análise de agrupamentos não-hierárquicos em *Machine Learning*

Além do k-means, outro método bem conhecido é o random forest!

Esse método consiste em construir árvores de decisões aleatórias, ou seja, com base nas características das variáveis informadas, o algoritmo busca identificar os pontos de corte que melhor segregam os dados.



O aprendizado pode ser supervisionado ou não-supervisionado!

Principais pacotes no R

- randomForest
- caret
- cluster

Mãos a obra! Trabalhando com dados reais!

Exercício

1. Baixe o conjunto de dados de cancer de mama (Breast Cancer Wisconsin (Diagnostic) Data Set | Kaggle);
2. Selecione 5 variáveis (menos id e diagnosis) utilizando o critério de sua preferência;
3. Aplique um dos métodos de agrupamentos apresentados
4. Verifique se os grupos formados são semelhantes ao “diagnosis”. Um meio de fazer isso é usando a função table().

Ex:

```
tabela <- table(dados$diagnosis, dados$clusters).
    acuracia <- sum(diag(tabela))/sum(tabela)
```

Quanto mais próximo de 1 for a acurácia, melhor foi o arranjo do agrupamento.