Algoritmos de Agrupamento - Aprendizado Não Supervisionado

Fabrício Jailson Barth

Abril de 2013

Sumário

- Introdução e Definições
- Aplicações
- Algoritmos de Agrupamento
  - ⋆ Agrupamento Plano
  - ⋆ Agrupamento Hierárquico
- Considerações Finais

Introdução

Introdução e Definições

- Os algoritmos de agrupamento particionam um conjunto de objetos em agrupamentos [Manning and Schütze, 2003].
- Normalmente, objetos são descritos e agrupados usando um conjunto de atributos e valores.
- Não existe nenhuma informação sobre a classe ou categoria dos objetos.
- **O objetivo dos algoritmos de agrupamento é colocar os objetos similares em um mesmo grupo e objetos não similares em grupos diferentes**.

Aplicações

- Agrupamento de objetos similares, onde **"objetos"** podem ser:
  - ⋆ agrupamento de documentos (textos) similares
  - ⋆ identificação de grupos em redes sociais
  - ⋆ segmentação de clientes
  - ⋆ identificação de plantas com características comuns

Algoritmos

Algoritmos de Agrupamento

Existem dois tipos de estruturas produzidas por algoritmos de agrupamento:

- não hierárquicos ou **planos**
- agrupamentos **hierárquicos**

Agrupamento Plano

- Agrupamentos planos simplesmente contêm um certo número de agrupamentos e a **relação** entre os agrupamentos e geralmente **não-determinada**.
- A maioria dos algoritmos que produzem agrupamentos planos são **iterativos**.
- Eles iniciam com um conjunto inicial de agrupamentos e realocam os objetos em cada agrupamento de maneira iterativa.
- Até uma determinada **condição de parada**.

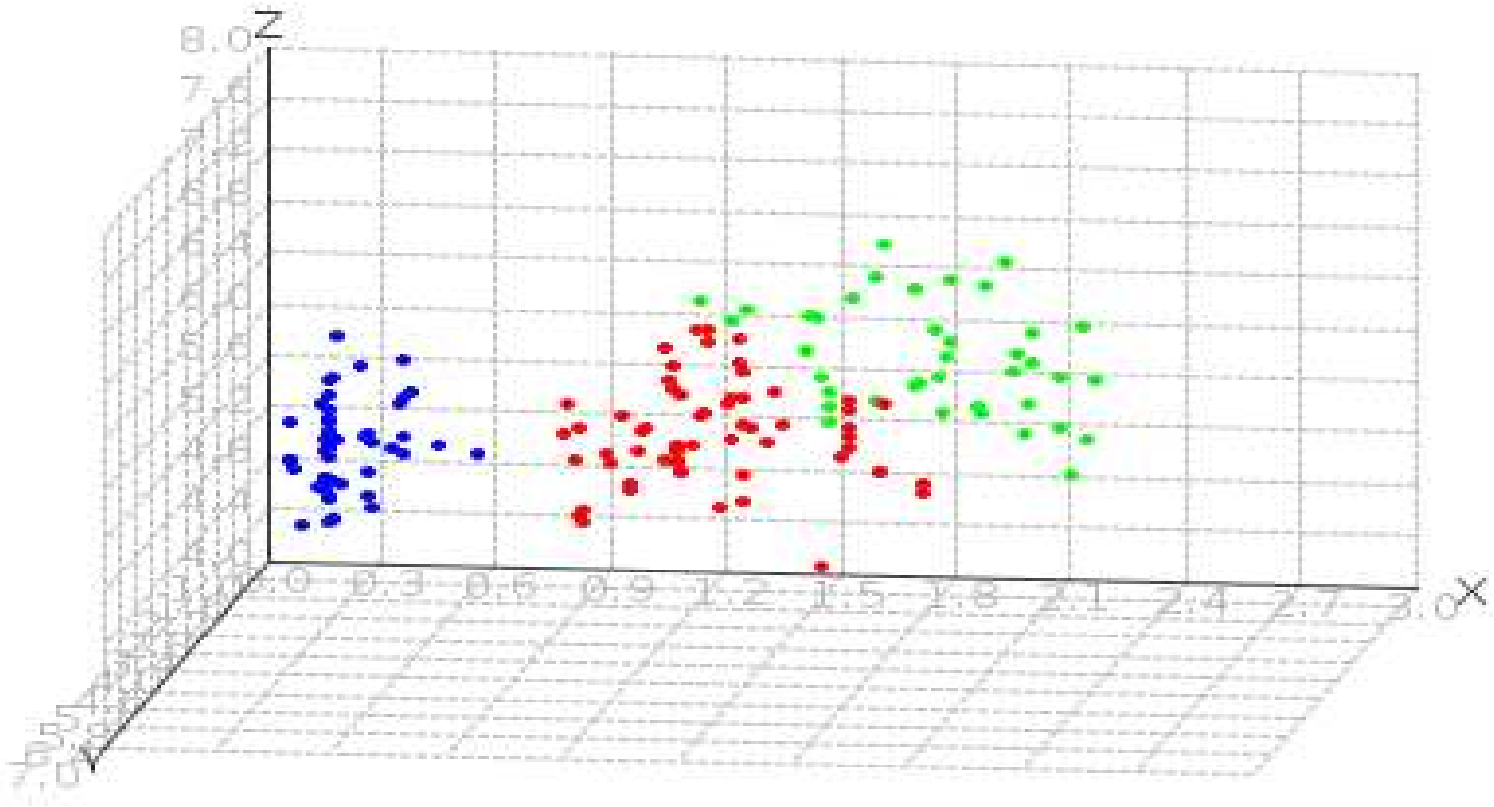
Agrupamentos **soft** e **hard**

Além da divisão entre os algoritmos hierárquicos e planos, tem-se a divisão entre os algoritmos **soft** e **hard**.

- Na abordagem **hard** cada objeto é inserido em um e somente um agrupamento.
- Na abordagem **soft** um objeto pode ser inserido em vários agrupamentos com diferentes níveis de pertinência.

Em agrupamentos hierárquicos, geralmente a abordagem é **hard**. Em agrupamentos planos, ambos os tipos de abordagens são comuns.

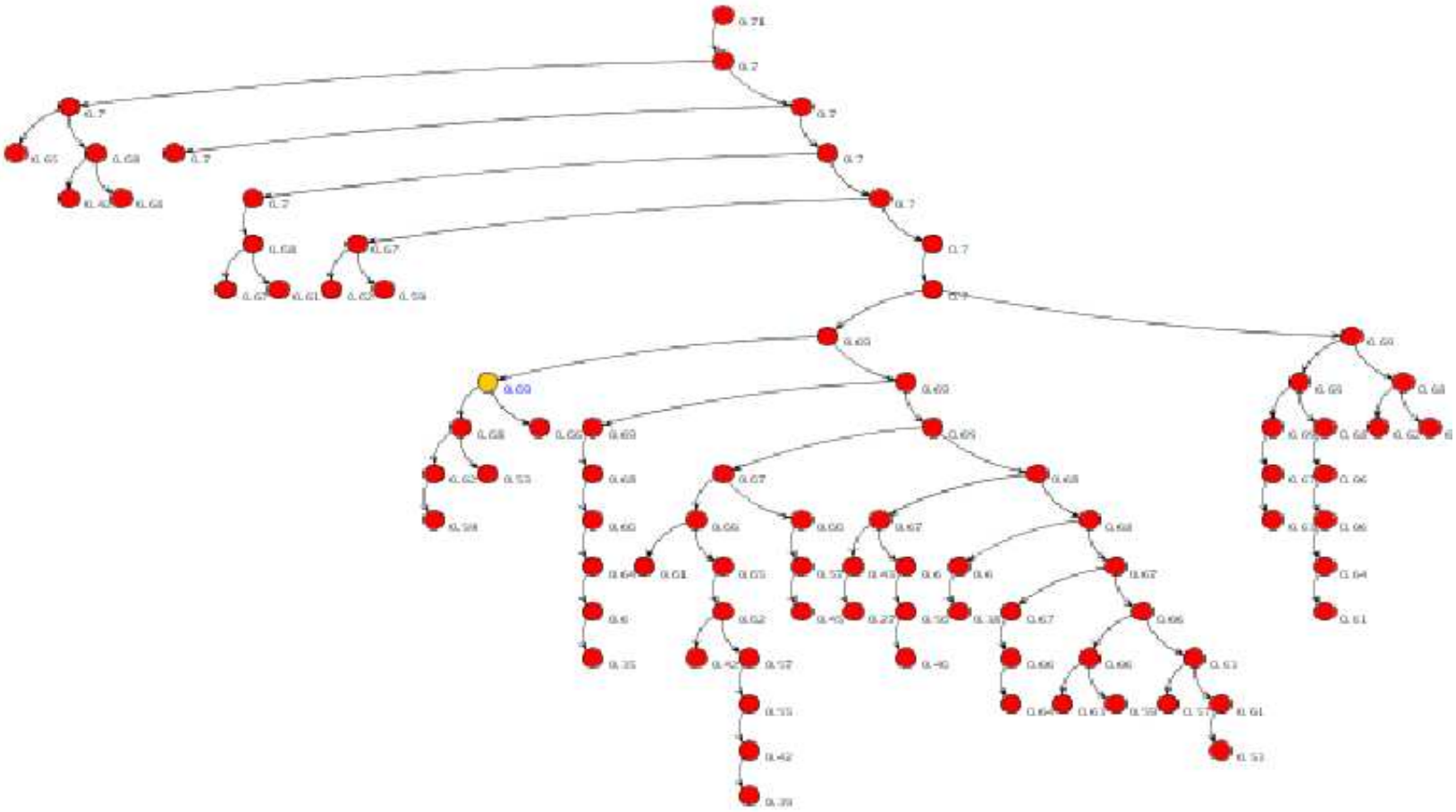
Agrupamento Plano *hard* (Exemplo)



Agrupamento Hierárquico

- Um agrupamento hierárquico é representado por uma árvore.
- Os nós folhas são os objetos.
- Cada nó intermediário representa o agrupamento que contêm todos os objetos de seus descendentes.

Agrupamento Hierárquico (Exemplo)



Algoritmos para Agrupamento Plano

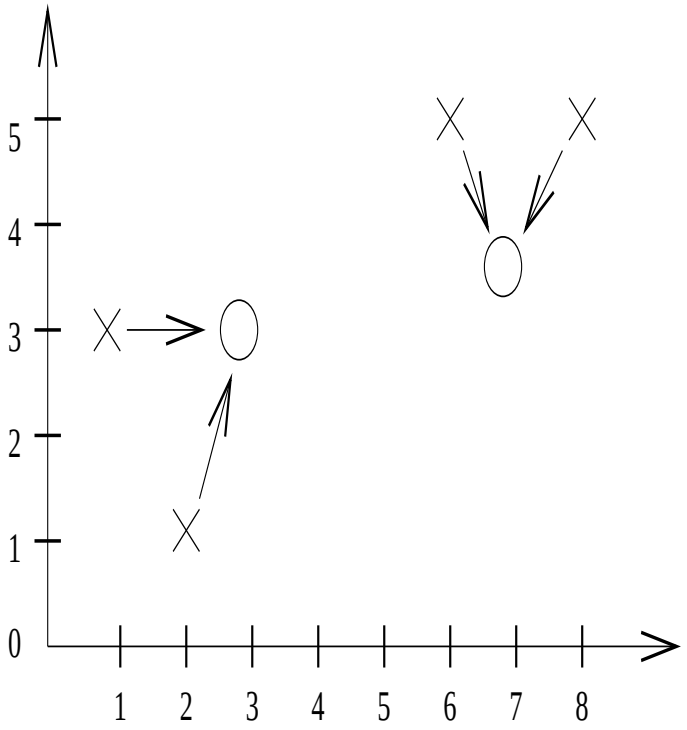
Algoritmos para agrupamento plano

- Utiliza diversas **iterações** para realocar os objetos nos melhores agrupamentos.
- **Critério de parada** é baseado na qualidade dos agrupamentos (similaridade média e cálculo para informação comum entre agrupamentos).
- É necessário determinar o **número de agrupamentos**:
  - ⋆ usando conhecimento à priori
  - ⋆ k, $k-1$ aumento significativo da qualidade, $k+1$ aumento reduzido da qualidade. Procurar por um k com este comportamento.

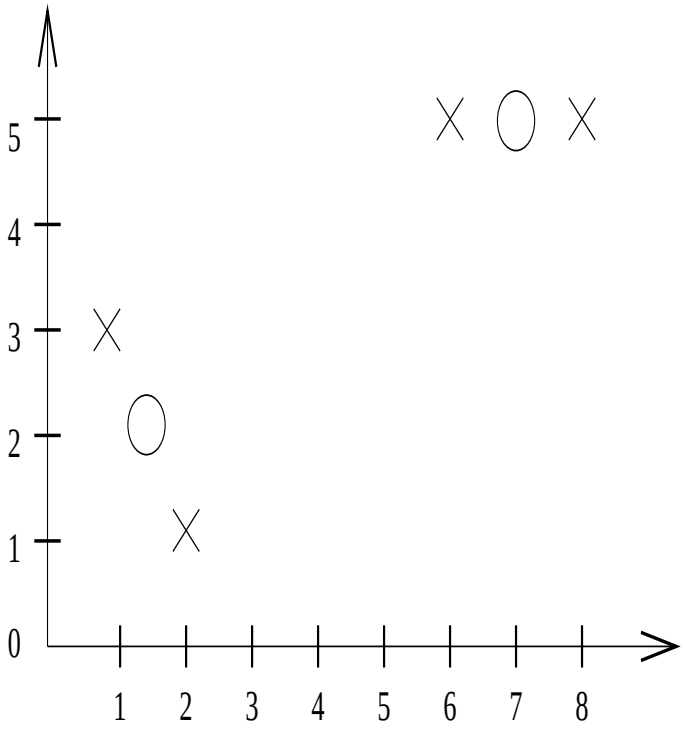
K-means

- Algoritmo de agrupamento **hard**
- Define o agrupamento pelo centro de massa dos seus membros.
- É necessário um conjunto inicial de agrupamentos.
- Seqüência de ações iterativas.
- Usualmente, diversas iterações são necessárias para o algoritmo convergir.

Iteração do algoritmo **K-means**



(1) Atribuição dos objetos aos agrupamentos



(2) Definição do centro do agrupamento

Algoritmo **K-means**

entrada: um conjunto $X = \{\overrightarrow{x_1}, \cdots, \overrightarrow{x_n}\} \subset \Re^m$

{conjunto inicial de agrupamentos}

uma medida de distância: d: $\Re^m \times \Re^m \rightarrow \Re$

uma função para computar o ponto central:

μ: $P(\Re) \rightarrow \Re^m$

selecionar k centros iniciais $\overrightarrow{f_1}, \cdots, \overrightarrow{f_k}$

while o critério de parada não for verdadeiro **do**

for todos os agrupamentos c_j **do**

$c_j = \{\overrightarrow{x_i} \mid \forall \overrightarrow{f_l}\, d(\overrightarrow{x_i}, \overrightarrow{f_j}) \leq d(\overrightarrow{x_i}, \overrightarrow{f_l})\}$ {os agrupamentos c_j recebem objetos levando-se em consideração o seu centro f_j}

end for

for todos os pontos centrais $\overrightarrow{f_j}$ **do**

$\overrightarrow{f_j} = \mu(c_j)$

end for

end while

Algoritmo **K-means**

- A medida de distância pode ser a distância Euclidiana:

$$| \overrightarrow{x} - \overrightarrow{y} | = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (1)$$

- a função para computar o ponto central pode ser:

$$\overrightarrow{\mu} = \frac{1}{M} \sum_{\overrightarrow{x} \in C} \overrightarrow{x} \qquad (2)$$

onde M é igual ao número de pontos no agrupamento C.

Problema...

Blue Flag Iris

- Considere uma base de dados sobre um determinado tipo de flor.
- Esta base de dados possui informações sobre o **comprimento** e **largura** do **caule** e das **pétalas** de várias flores parecidas (todas azuis).

Blue Flag Iris - Dados

```
1   @ATTRIBUTE sepallength  REAL
2   @ATTRIBUTE sepalwidth   REAL
3   @ATTRIBUTE petallength  REAL
4   @ATTRIBUTE petalwidth   REAL
5   @DATA
6   5.1,3.5,1.4,0.2
7   4.9,3.0,1.4,0.2
8   4.7,3.2,1.3,0.2
9   4.6,3.1,1.5,0.2
10  5.0,3.6,1.4,0.2
11  6.6,2.9,4.6,1.3
12  5.2,2.7,3.9,1.4
```

Todas as medidas são em cm.

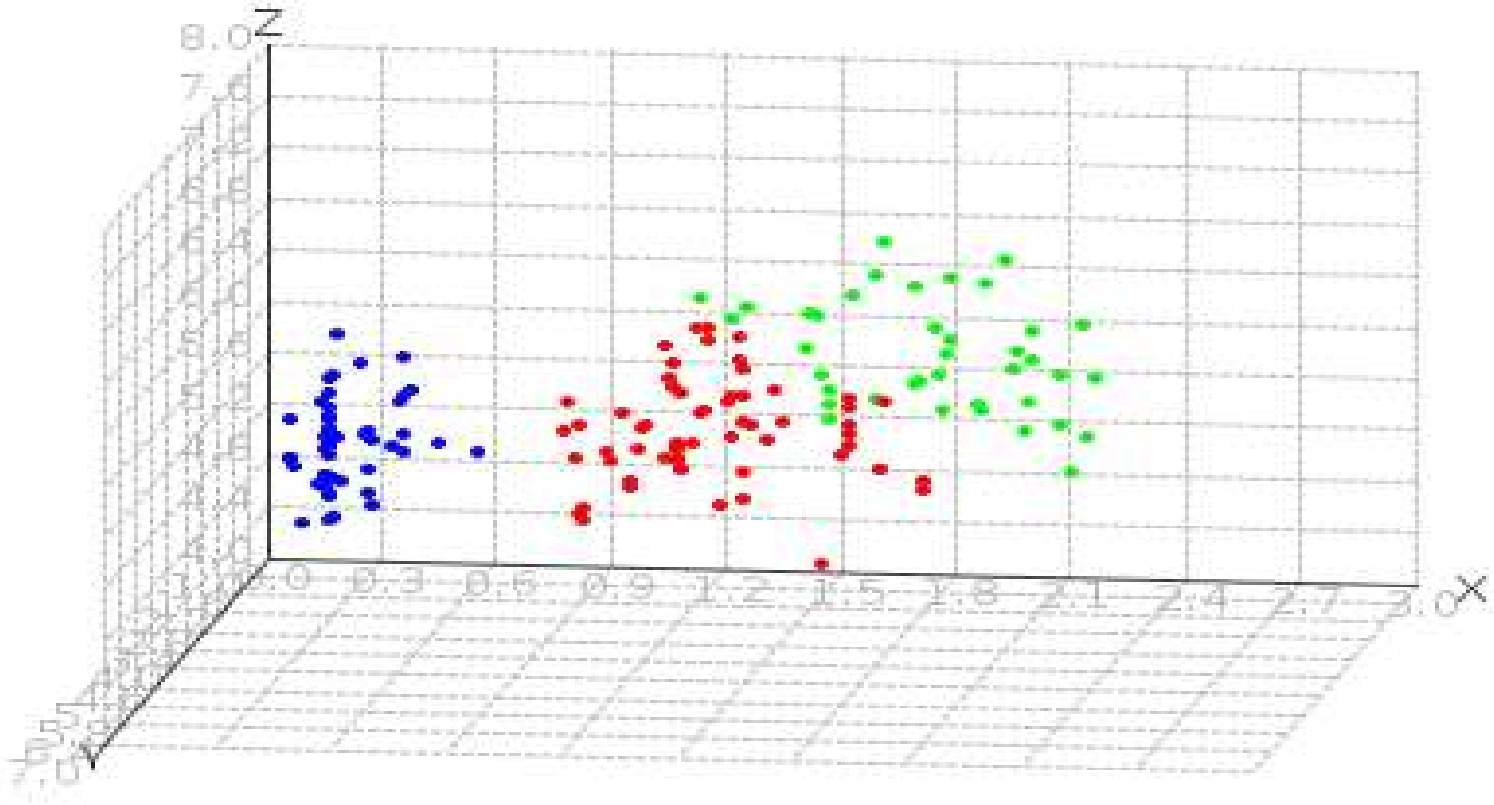
Aplicando o algoritmo K-means

Cluster Model

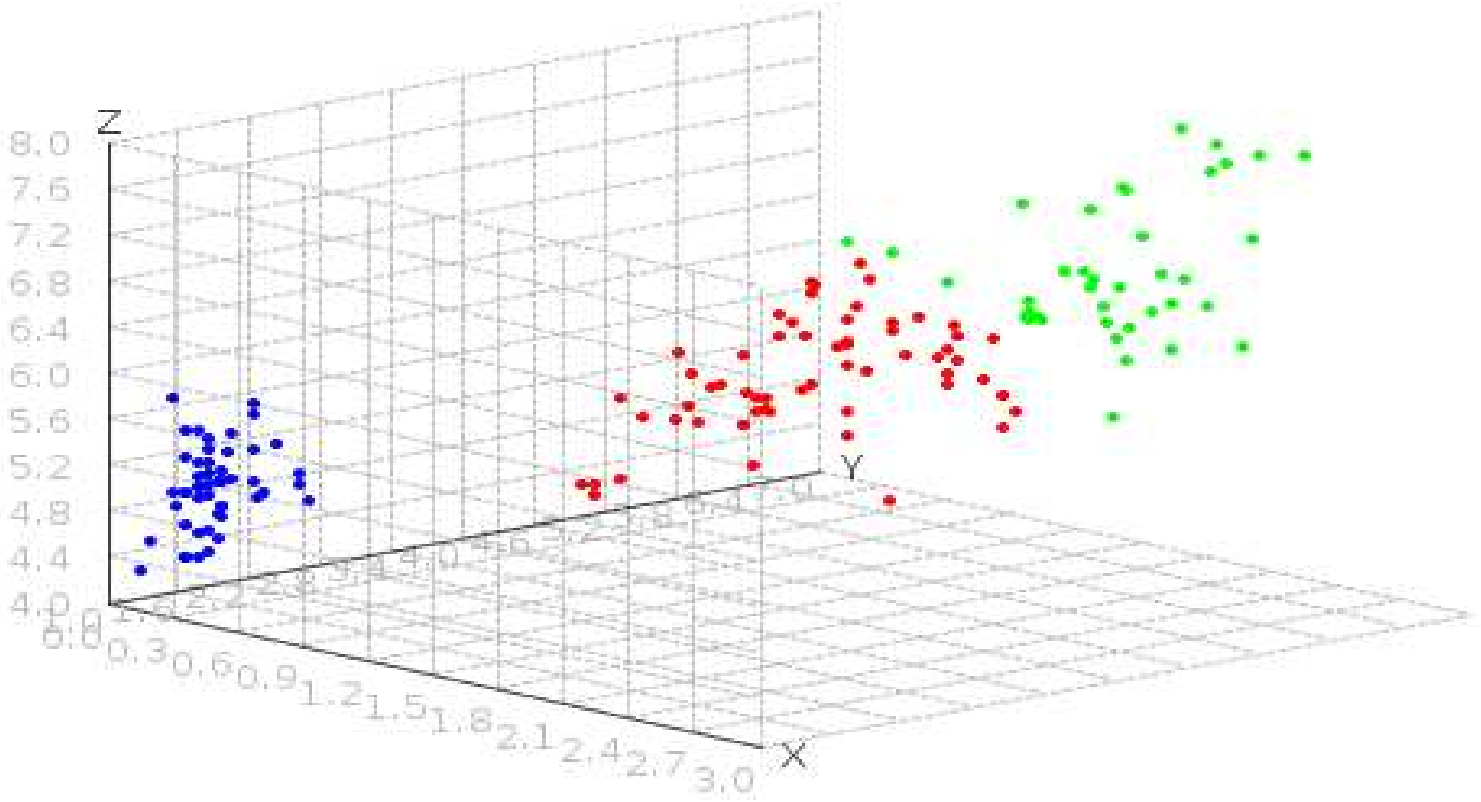
A cluster model with the following properties:

Cluster 0 (no description available) : 62 items
Cluster 1 (no description available) : 38 items
Cluster 2 (no description available) : 50 items
Total number of items: 150

Cluster centroids:
Cluster 0: sepallength = 5,902 sepalwidth = 2,748 petallength = 4,394 petalwidth = 1,434
Cluster 1: sepallength = 6,850 sepalwidth = 3,074 petallength = 5,742 petalwidth = 2,071
Cluster 2: sepallength = 5,006 sepalwidth = 3,418 petallength = 1,464 petalwidth = 0,244

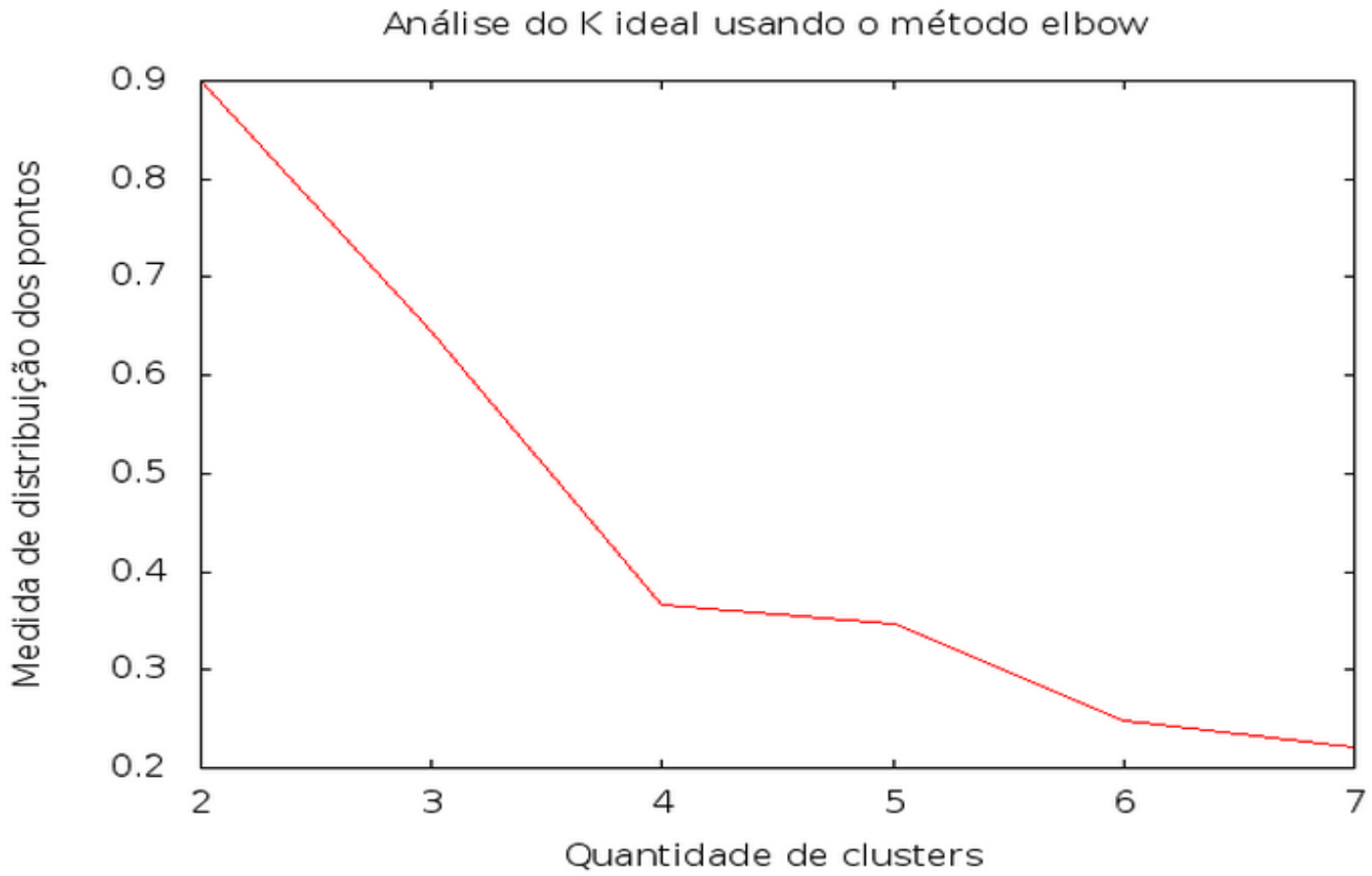


x = Petal Width, y = Petal Length, z = Sepal Length.



x = Petal Width, y = Petal Length, z = Sepal Length.

Como determinar o melhor k?



A medida de distribuição dos pontos normalmente empregada é *sum of squared errors*.

Algoritmos para Agrupamento Hierárquico

Algoritmos para agrupamento hierárquico

Os algoritmos que utilizam a abordagem de agrupamento hierárquico podem implementar abordagens:

- **bottom-up (agglomerative clustering)**
- **top-down (divisive clustering)**

Agrupamento hierárquico **bottom-up**

Entrada: um conjunto $x = \{x_1, \cdots, x_n\}$ de objetos

e uma função sim: $P(X) \times P(X) \rightarrow \Re$

for i:=1 até n **do**

$c_i := \{x_i\}$ {Inicia com um agrupamento para cada objeto}

end for

j := n + 1

while $|C| > 1$ **do**

$(c_{n1}, c_{n2}) := \arg\max_{c_u, c_v \in C \times C} sim(c_u, c_v)$ {Em cada passo, os dois agrupamentos mais similares são determinados}

$c_j := c_{n1} \cup c_{n2}$ {Unidos em um novo agrupamento}

$C := C \setminus \{c_{n1}, c_{n2}\} \cup \{c_j\}$ {Elimina-se os dois agrupamentos mais similares e adiciona-se o novo agrupamento ao conjunto de agrupamentos}

$j := j + 1$

end while

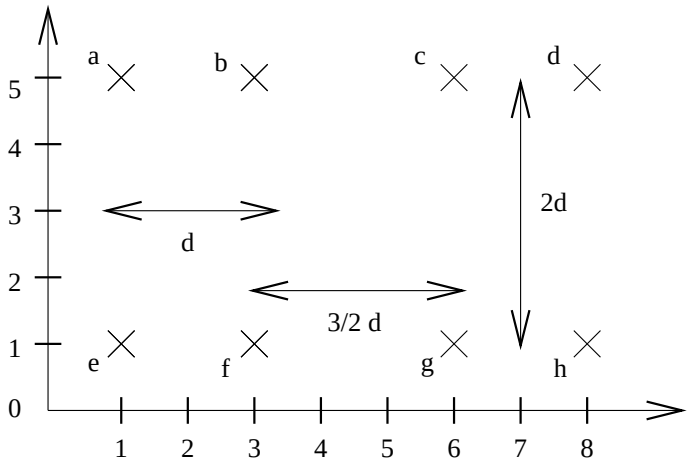
Agrupamento hierárquico **bottom-up** - Função de similaridade

- A função de similaridade pode ser a distância Euclidiana:

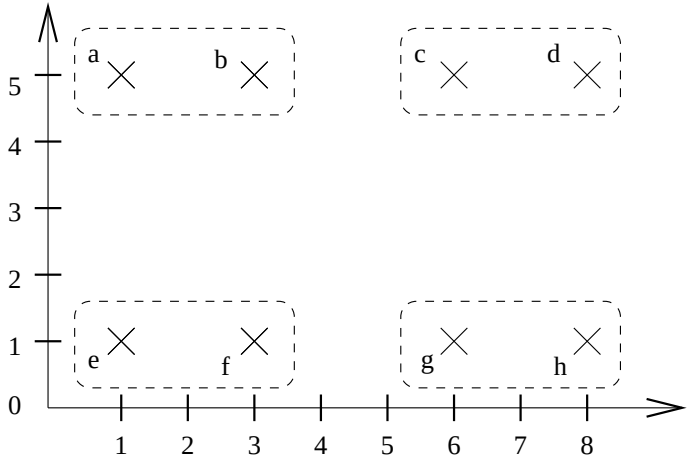
$$| \overrightarrow{x} - \overrightarrow{y} | = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (3)$$

Tipos de funções de similaridade

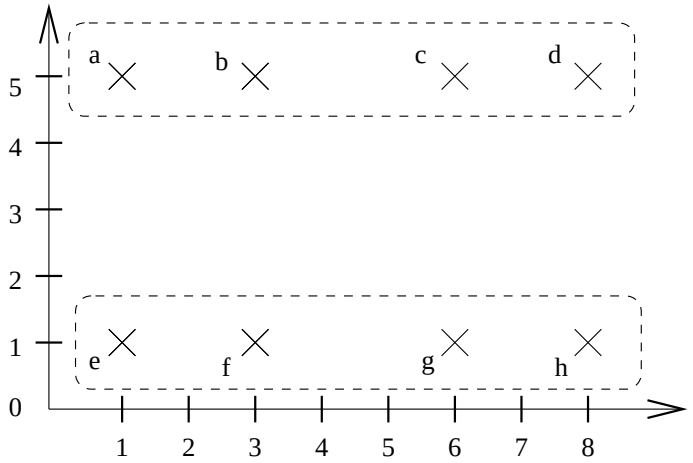
- ligação simples (**single link**): a similaridade entre dois agrupamentos é o melhor resultado retornado da similaridade entre os seus membros **mais** similares.
- ligação completa (**complete link**): a similaridade entre dois agrupamentos é o melhor resultado retornado da similaridade entre os seus membros **menos** similares.
- média do grupo (**group-average**): a similaridade entre dois agrupamentos é a **média** da similaridade entre os membros dos agrupamentos.



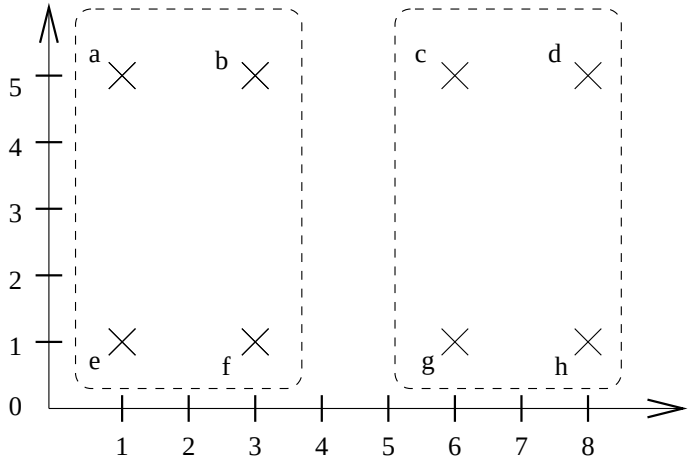
(1) Pontos em um plano



(2) Agrupamentos intermediários dos pontos da figura 1



(3) Agrupamentos de ligação simples para os pontos da figura 1



(4) Agrupamentos de ligação completa para os pontos da figura 1

Agrupamento hierárquico **top-down**

Entrada: um conjunto $x = \{x_1, \cdots, x_n\}$ de objetos,
uma funcao de coesao coh: $P(X) \rightarrow \Re$
e uma funcao de divisao $split$: $P(X) \rightarrow P(X) \times P(X)$
$C := \{X\} (= \{c_1\})$ {Inicia com um agrupamento com todos os objetos}
$j := 1$
while $\{\exists c_i \in C \mid |c_i| > 1\}$ **do**
  $c_u := \arg\min_{c_v \in C} coh(c_v)$ {Determina que agrupamento eh o menos coerente}
  $(c_{j+1}, c_{j+2}) := split(c_u)$ {Divide o agrupamento}
  $C := C \setminus \{c_u\} \cup \{c_{j+1}, c_{j+2}\}$
  $j := j + 2$
end while

Restrição sobre os agrupamentos hierárquicos

Agrupamento hierárquico só faz sentido se a função de similaridade é monotônica decrescente das folhas para a raiz da árvore:

$$\forall c, c', c'' \subseteq S : \min(sim(c, c'), sim(c, c'')) \geq sim(c, c' \cup c'') \tag{4}$$

Algumas considerações sobre agrupamentos

- Um agrupamento é um grupo de objetos centrados em torno de um ponto central.
- Os agrupamentos mais compactos são os preferidos.

Considerações Finais

Sumário dos atributos dos algoritmos

Agrupamento hierárquico:

- É a melhor abordagem para análise exploratória de dados.
- Fornece mais informação que agrupamento plano.
- Menos eficiente que agrupamento plano (tempo e memória gastos).

Sumário dos atributos dos algoritmos

Agrupamento plano:

- É preferível se a eficiência é um atributo importante e se o conjunto de dados é muito grande.
- O algoritmo **K-means** é o método mais simples e deve ser usado sobre novos conjuntos de dados porque os resultados são geralmente suficientes.

References

[Manning and Schütze, 2003] Manning, C. D. and Schütze, H. (2003). *Foundations of Statistical Natural Language Processing*. MIT Press.

[Mitchell, 1997] Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.