

Faculdade de Computação e Informática  Mackenzie

Curso: Sistemas de Informação

CIÊNCIA de DADOS


Aprendizado de Máquina
(Machine Learning)

Análise de AGRUPAMENTO
(Clusterização)



Prof. Dr. Arnaldo R. A. Vallim Fº
aavallim@mackenzie.br

1

Aprendizado de Máquina  Mackenzie

SUMÁRIO

- **Aprendizado de Máquina (ML – Machine Learning)**
 - Aprendizado Supervisionado vs. Aprendizado Não Supervisionado
- **Agrupamento de Dados (Clusterização)**
 - “Clusterização”: Visão Geral
 - Agrupamento por Partição: **K-Médias (K-Means)**
 - Agrupamento **Hierárquico**

2

SUMÁRIO

■ **Aprendizado de Máquina (ML – Machine Learning)**

- Aprendizado Supervisionado vs. Aprendizado Não Supervisionado

■ **Agrupamento de Dados (Clusterização)**

- “Clusterização”: Visão Geral
- Agrupamento por Partição: **K-Médias (K-Means)**
- Agrupamento **Hierárquico**

3

BASES da ML

APRENDIZADO DE MÁQUINA (ML – MACHINE LEARNING)

O aprendizado de máquina é uma forma de IA que permite que um sistema aprenda com os dados, e não através de programação explícita

DOIS TIPOS PRINCIPAIS DE APRENDIZADO

- . **Aprendizado Supervisionado**
- . **Aprendizado não Supervisionado**

4

Tipos de Aprendizado

Aprendizado Supervisionado (AS)

O objetivo do Aprendizado Supervisionado é encontrar padrões nos dados, que posteriormente, podem ser aplicados a outros dados.

O AS parte de um conjunto de dados em que já exista uma classificação das Observações (*registros da Base de Dados devem estar classificados*).

Para isto, o Conjunto de Dados deve ter uma coluna (um Atributo) com um “rótulo” que defina o significado de cada Observação (classifica cada registro da base de dados).

Exemplo 1: Dataset de Peças

Com várias características (atributos) das peças, como: comprimento, peso, largura, material, etc.

Em uma das colunas há a informação (rótulo): Peça com Defeito ou sem Defeito

Baseado nesse *dataset*, no futuro, o algoritmo poderá classificar uma peça a partir dos outros Atributos

Exemplo 2: Dataset de Alunos

Com a classificação (rótulo): aluno **ruim, médio, bom ou ótimo**

O algoritmo poderá relacionar esses rótulos com outras características (Atributos) dos alunos, como: nota, índice de faltas e tarefas entregues.

Assim, no futuro, o algoritmo poderá classificar um aluno a partir desses outros Atributos.

5

Tipos de Aprendizado

Aprendizado NÃO Supervisionado (AnS)

O objetivo do Aprendizado Não Supervisionado é encontrar relações de similaridade entre os dados.

Ao contrário do anterior, **no AnS os dados não possuem rótulo**. Não há qualquer tipo de classificação ou organização dos dados.

O algoritmo de AnS deve explorar a base de dados buscando encontrar algum tipo de estrutura, como, por exemplo, grupos de Exemplares com “similaridades”, montando assim, Grupos de Observações (grupos de exemplares).

Exemplo 1: Dataset de Clientes de um Supermercado

Sem nenhuma classificação (rótulo) dos Clientes.

A Base de Dados deve ter várias características (atributos) dos clientes, como: produtos comprados, frequência de compra, valor das compras, etc.

Baseado nesse *dataset*, o algoritmo poderá agrupar os clientes segundo esses Atributos.

Exemplo 2: Dataset de Alunos

Sem nenhuma classificação (rótulo) dos alunos.

A Base de Dados deve ter várias características (atributos) dos alunos, como: nota, índice de faltas, tarefas entregues, etc.

O algoritmo poderá agrupar os alunos com base nesses Atributos.

Assim, no futuro, o algoritmo poderá alocar um aluno a um Grupo, a partir de seus Atributos.

6

SUMÁRIO

- Aprendizado de Máquina (*ML – Machine Learning*)
 - Aprendizado Supervisionado vs. Aprendizado Não Supervisionado
- **Agrupamento de Dados (Clusterização)**
 - “Clusterização”: Visão Geral
 - Agrupamento por Partição: **K-Médias** (*K-Means*)
 - Agrupamento Hierárquico

7

O PROBLEMA de CLUSTERIZAÇÃO **DEFINIÇÃO de CLUSTERS**

- ☛ Passo fundamental em muitos casos práticos de Planejamento Operacional, Tático e Estratégico

Pode influenciar de forma ampla uma operação

☛ **QUESTÃO PRÁTICA PRINCIPAL**

*Problemas de Grande Porte
Bases de Dados de Grande Porte!*

*Como compreender os dados?
Como tomar decisão a partir dos Dados?*

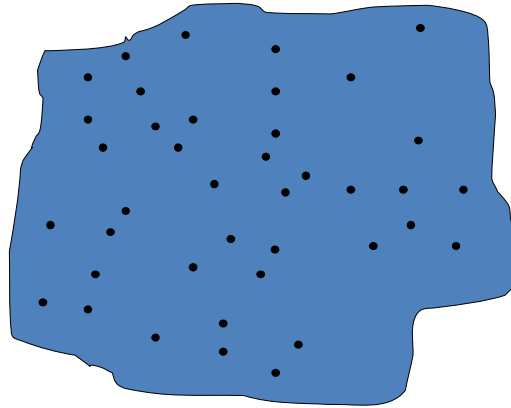
8

○ Problema de Clusterização



*Seja um conjunto de pontos,
que representa Exemplares de um Dataset...*

Situação Inicial



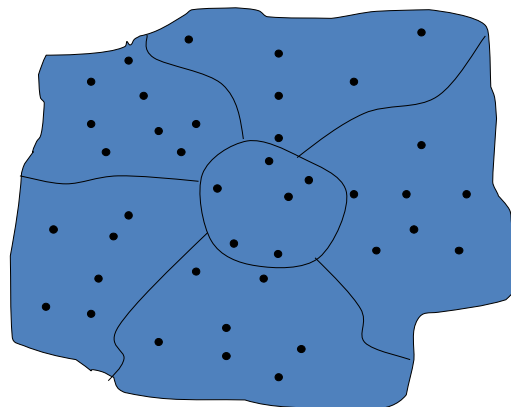
9

○ Problema de Clusterização



OBJETIVO do PROBLEMA
Agrupar os pontos em Clusters

Situação Final



10

Conceitos

“Clusterização” é o processo de organizar elementos em grupos cujos membros são similares de alguma forma (Matteucci, 2006)

Estatística é a área que primeiro estudou o problema

Em Estatística, a Análise de Clusters (**Cluster Analysis**) se enquadra nas chamadas técnicas de Análise Multivariada

A Análise de Clusters identifica tipos de elementos ou populações (no sentido estatístico) que se abrigariam sob uma massa de dados, que se constituiriam em amostras dessas populações

Matteucci, M. (2006). Tutorial in Clustering. Politecnico di Milano. Department of Electronics and Information. Milano, Italy.
Disponível em: https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/

11

Conceitos

“Clusterização” é o processo de organizar elementos em grupos cujos membros são similares de alguma forma (Matteucci, 2006)

Mais recentemente, com o crescimento da Ciência de Dados, o assunto ganhou novo impulso.

Em Mineração de Dados tem-se o seguinte conceito básico:

“os elementos de um cluster devem possuir um grau de similaridade maior com os demais elementos de seu próprio cluster, do que com indivíduos de outros clusters”

E esta é uma questão chave: a definição da medida de similaridade.

Pode-se ter dois tipos de problema:

- . K-Clusterização, quando K, o número de clusters, é pré-definido
- . Problema de Clusterização Automática (PCA), quando K não é previamente conhecido

Matteucci, M. (2006). Tutorial in Clustering. Politecnico di Milano. Department of Electronics and Information. Milano, Italy.
Disponível em: https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/

12

■ Clusterização é um Problema Combinatório

Combinatório = Multiplicidade de Soluções Viáveis

☞ *Como resolver este Problema?*
complexidade cresce com o número de objetos!



Tradicionalmente... por Modelos Matemáticos
Modelos Exatos de Otimização

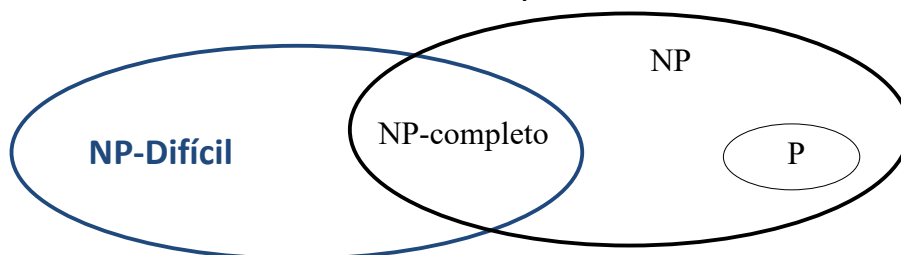
Programação Linear Inteira Mista (PLIM)
(Mixed Integer Linear Programming – MILP)

Problema de Clusterização é complexo quando resolvido por MILP
Complexidade = No. de operações computacionais elementares
(somar, subtrair, comparar, etc.)

13

Problemas

☞ Classes de Complexidade



Solução por MILP é NP-Difícil
a mais alta classe de complexidade!

Problema NP-Difícil → Tempo Computacional tende a infinito quando problema cresce

14

■ Problema Combinatório

complexidade cresce com o número de objetos !

Como superar esta dificuldade ?



Heurísticas e/ou Metaheurísticas



IA / Machine Learning

15

CLUSTERIZAÇÃO

TIPOS de MÉTODOS

Principais métodos são de dois tipos:

- . HIERÁRQUICO ou
- . PARTICIONAMENTO.

HIERÁRQUICO

O tipo Hierárquico pode ser de Aglomeração ou de Divisão:

- Na estratégia hierárquica de Aglomeração (*bottom-up*) pode-se iniciar com cada elemento se constituindo em um cluster. A seguir, os clusters vão sendo unidos. Assim, cada cluster que tiver um tamanho maior que 1, é porque foi composto por dois ou mais clusters. A união de clusters sempre ocorre com base em alguma medida de similaridade
- Na estratégia hierárquica de Divisão (*top-down*), se inicia com um só cluster e procede-se a uma sequência de divisões, com base em algum critério

16

TIPOS de MÉTODOS

Principais métodos são de dois tipos:

- . HIERÁRQUICO ou
- . PARTICIONAMENTO.

PARTICIONAMENTO

No método do tipo Particionamento, parte-se de um conjunto “K” de clusters e procede-se a um intercâmbio de elementos entre esses clusters

A cada rearranjo dos clusters tem-se uma nova configuração

Para se aceitar uma nova configuração tem-se uma Função Objetivo

Esta função é fator chave!

Deve ser efetivamente representativa do objetivo que se busca alcançar

17

SUMÁRIO

- Aprendizado de Máquina (*ML – Machine Learning*)
 - Aprendizado Supervisionado vs. Aprendizado Não Supervisionado
- **Agrupamento de Dados (Clusterização)**
 - “Clusterização”: Visão Geral
 - **Agrupamento por Partição: K-Médias (K-Means)**
 - Agrupamento Hierárquico

18

CLUSTERIZAÇÃO



K-Médias ou K-Means

É um Método de Clusterização por PARTICIONAMENTO

PARTICIONAMENTO: No Particionamento, parte-se de um conjunto “K” de clusters e procede-se a um intercâmbio de elementos entre os clusters

K-MÉDIAS (K-Means)

Este método segue o princípio acima, através dos seguintes os Passos:

- Passo 1: Definir aleatoriamente as posições de um número K de Centróides de clusters
- Passo 2: Calcular a distância de cada elemento (cada Exemplar) a todos os K centróides
- Passo 3: Formar grupos, unindo os Exemplares ao Centróide mais próximo
- Passo 4: Se for a 1ª Iteração, seguir para o Passo 5.
A partir da 2ª Iteração:
 - . Se não há alteração nos Grupos, encerrar o processo.
 - . Em caso contrário, seguir para o Passo 5
- Passo 5: Com os grupos formados, recalculer os Centróides de cada grupo
Retornar ao Passo 2

19

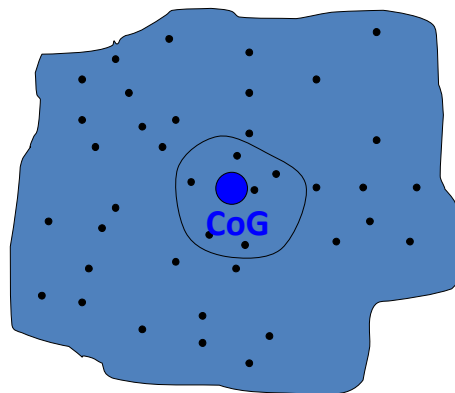
CLUSTERIZAÇÃO



Pontos e Centróides

1 - Centróide

(Center of Gravity – CoG)



CENTROIDE: Corresponde aproximadamente ao Centro de Gravidade dos pontos

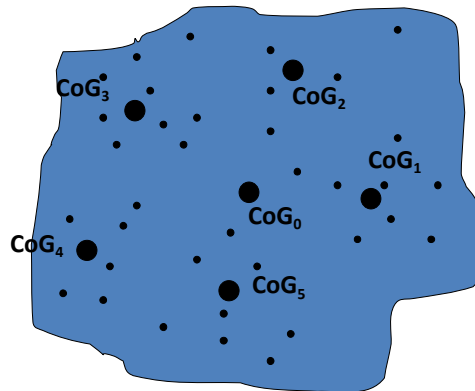
20

CLUSTERIZAÇÃO



K-Médias ou *K-Means*

K – Centróides Aleatórios
(Centers of Gravity – CoG)



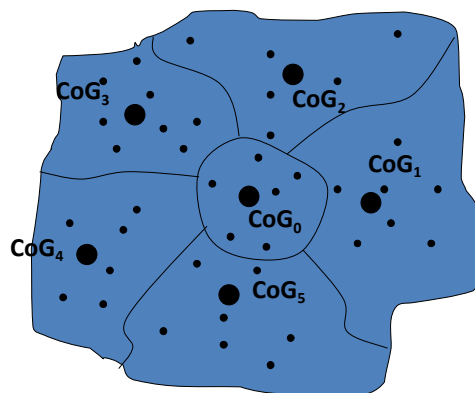
21

CLUSTERIZAÇÃO



K-Médias ou *K-Means*

K – Centróides e Clusters
(Centers of Gravity – CoG)



22

K-Médias ou *K-Means*

K-MÉDIAS (*K-Means*)

PONTOS CHAVE do ALGORITMO

- . Centroides de Clusters
- . Posições dos Pontos e dos Centróides
- . Distância dos elementos (Exemplares) aos Centroides

23

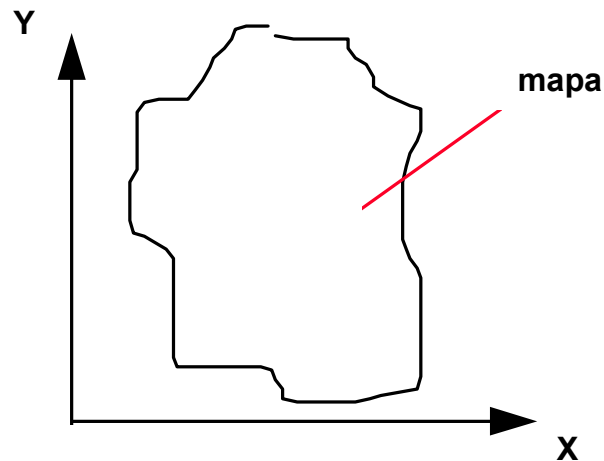
K-Médias ou *K-Means*

K-MÉDIAS (*K-Means*): PONTOS CHAVE do ALGORITMO

- . **Centroide de Cluster**
Corresponde aproximadamente ao Centro de Gravidade dos pontos (dos elementos ou exemplares) de cada Cluster
- . **Posições dos Elementos do Cluster**
As posições são definidas pelos valores dos Atributos de cada elemento
São, na verdade, Coordenadas (Atributos = Coordenadas = Dimensões)
No caso de se ter apenas 2 Atributos, pode-se representar todos os Exemplares e Centróides, como Pontos em um gráfico XY (2 Dimensões)
Cada Ponto teria 2 Coordenadas: Atributo 1 (X) e Atributo 2 (Y)
O Atributo 1 ficaria no eixo X e o Atributo 2 no eixo Y
- . **Posições dos Centróides**
A posição de cada Centróide é definida pelas Médias de cada Atributo dos pontos do Cluster. No caso de 2 Atributos, ter-se-ia 2 Médias
- . **Distância dos elementos (Exemplares) aos Centroides**
É calculada segundo o padrão de Distância Euclidiana (distância em linha reta)

24

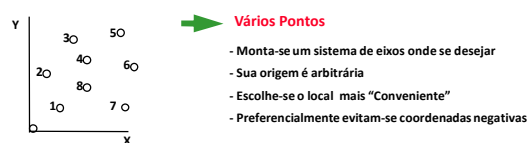
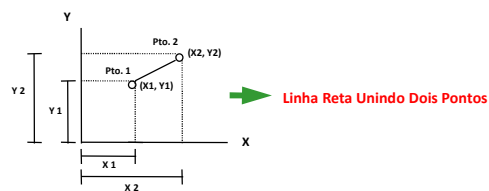
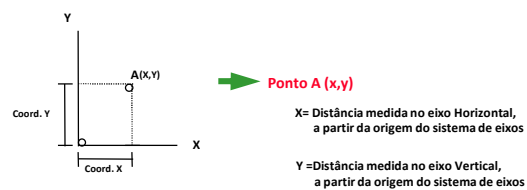
Coordenadas Cartesianas e Distâncias



25

DISTÂNCIAS

Coordenadas Cartesianas



26

DISTÂNCIA EUCLIDIANA

Medida mais comumente utilizada

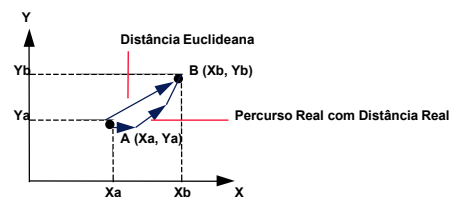
Representa a Distância Física entre 2 pontos no espaço

Distância em Linha Reta entre dois pontos A e B
($Dist_{A-B}$)

27

👉 **Cálculo de Distâncias**

DISTÂNCIA EUCLIDEANA

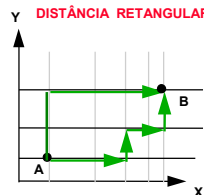


■ Distância Euclidean é a distância em linha reta entre dois pontos

- Distância Real (Física) \geq Distância Euclidean

Outros Tipos de DISTÂNCIA

DISTÂNCIA RETANGULAR



- Estrutura da rede é retangular
- Então, escolhe-se sistema de eixos paralelo às direções da rede.

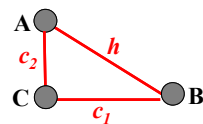
28

Duas Dimensões (2 Atributos)

Distância Euclidiana

Distância em Linha Reta entre dois pontos A e B ($Dist_{A-B}$)

Triângulo Retângulo



Teorema de Pitágoras

$$h^2 = c_1^2 + c_2^2$$

h = hipotenusa

c_1 = cateto 1

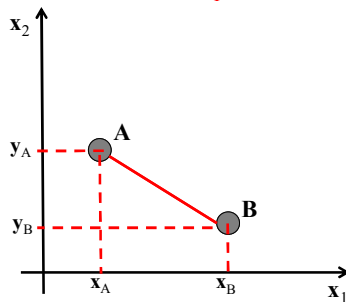
c_2 = cateto 2

h = $Dist_{A-B}$

$c_1 = x_B - x_A$

$c_2 = y_A - y_B$

$$d_{AB} = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$$



29

PROCESSO de CÁLCULO

Distância em Linha Reta entre dois pontos A e B ($Dist_{A-B}$)

2 Dimensões (2Atributos) $d_{AB} = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$

3 Dimensões (3Atributos) $d_{AB} = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2}$

4 Dimensões (4Atributos) $d_{AB} = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2 + (w_A - w_B)^2}$

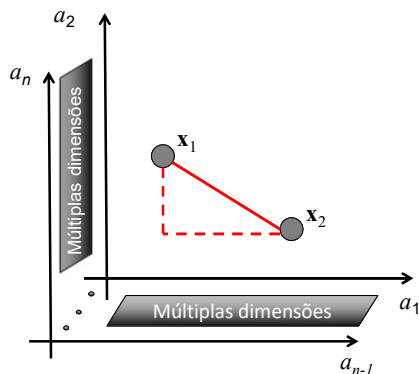
⋮

n Dimensões (n Atributos) $d_{AB} = \sqrt{\underbrace{(x_{A1} - x_{B1})^2}_{\text{Dimensão 1}} + \underbrace{(x_{A2} - x_{B2})^2}_{\text{Dimensão 2}} + \dots + \underbrace{(x_{An} - x_{Bn})^2}_{\text{Dimensão n}}}$

30

PROCESSO de CÁLCULO

Cálculo de Distâncias para “n” Dimensões (n Atributos)



d_{ij} = Distância do ponto i ao ponto j
 k = dimensão ; $k = 1, 2, \dots, n$

$$d_{ij} = \left(\sum_{k=1}^n (x_{ik} - x_{jk})^2 \right)^{1/2}$$

Fonte: Adaptado de Apresentação de “K-NN”
 Prof. Dr. Leandro Augusto da Silva – FCI/Mackenzie

31

CLUSTERIZAÇÃO

SUMÁRIO

- Aprendizado de Máquina (ML – Machine Learning)
 - Aprendizado Supervisionado vs. Aprendizado Não Supervisionado
- Agrupamento de Dados (Clusterização)
 - “Clusterização”: Visão Geral
 - Agrupamento por Partição: **K-Médias (K-Means)**
 - Agrupamento **Hierárquico**

32

Agrupamento Hierárquico

HIERÁRQUICO

Dois processos:

- **Agrupamento por Aglomeração** (*bottom-up*) inicialmente, cada objeto é um cluster. A seguir, os clusters vão sendo unidos, com base em uma medida de similaridade
- **Agrupamento por Divisão** (*top-down*), se inicia com um só cluster e procede-se a uma sequência de divisões, com base em algum critério

*Métodos de divisão geralmente não estão disponíveis e raramente foram aplicados
Neste texto, vamos apresentar apenas o método por Aglomeração (bottom-up)*

33

Clusterização Hierárquica por Aglomeração PROCESSO

O processo básico de agrupamento hierárquico por Aglomeração, definido por S.C. Johnson em seu famoso *paper*, de 1967, tem os seguintes passos:

1. Inicia-se definindo-se cada objeto (cada exemplar) como um cluster. Para N objetos, tem-se N clusters;
2. Monta-se a Matriz de Distâncias NxN (medida de similaridade) entre os N clusters;
3. Encontra-se o par de clusters mais próximo (mais similar), e estes são unidos em um único cluster. O número de clusters é reduzido em um;
4. Monta-se nova Matriz de Distâncias (similaridades), considerando-se agora, o novo cluster;
5. Repete-se os passos 3 e 4 até que todos os objetos estejam agrupados em um único cluster de tamanho N (vide OBS.)

OBS: É claro, que não há muito sentido em se ter todos os objetos agrupados em um único cluster. Mas a questão é que à medida que os clusters vão sendo unidos, vai-se montando uma Árvore Hierárquica das Partições do conjunto de objetos em números cada vez menores de clusters, até que se chegue a um único cluster. Esta Árvore Hierárquica é chamada de DENDOGRAMA. Tem-se ao final, o DENDOGRAMA Completo, com todas as partições dos dados em diferentes números de clusters, com k variando de N a 1 (k=No. de clusters; N=No. de Objetos ou Exemplares). Assim, desejando-se k clusters, basta se efetuar um corte no Dendograma no Nível k. Poda-se a árvore naquele patamar de k clusters (naquela cota), e os níveis seguintes (N-k níveis) são descartados.

REFERÊNCIAS:

Johnson, S. C. (1967). Hierarchical Clustering Schemes. *Psychometrika*, Vol. 32 No. 3: pp. 241-253
Matteucci, M. (2006). Tutorial in Clustering. Politecnico di Milano. Department of Electronics and Information. Milano, Italy.
Disponível em: <http://www.elet.polimi.it/upload/matteucci/Clustering/tutorial.html/>

34

DENDOGRAMA - *Árvore Hierárquica de Partições*

Ao final do processo de agrupamento, tem-se uma sequência ou hierarquia de partições do conjunto S , de dados, denotadas de P_0, P_1, \dots, P_{n-1}

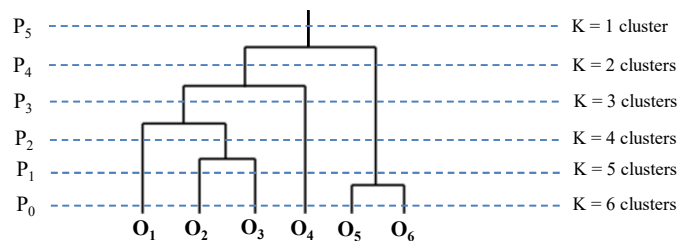
Em particular, P_0 , chamada de partição disjunta, apresenta os N objetos em N clusters separados. Este é o chamado cluster "fraco", de Johnson (1967)

P_{n-1} , a chamada partição conjunta, consiste de um único cluster com todos objetos incluídos. Este é o chamado agrupamento "forte", de Johnson (1967)

Árvore Hierárquica de Partições

DENDOGRAMA

Agrupamentos sucessivos de 6 Objetos ($O_i ; i=1, 2, \dots, 6$)

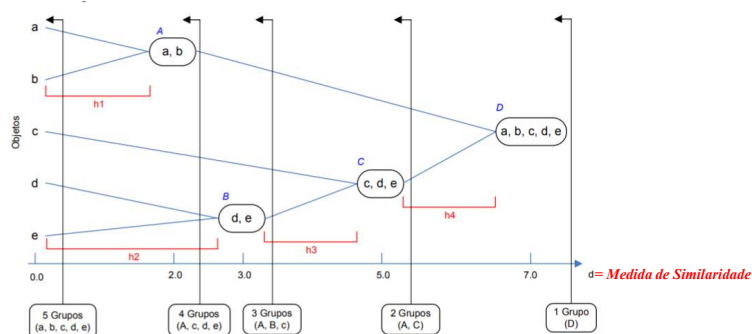


35

DENDOGRAMA - Exemplo

DENDOGRAMA

Agrupamentos sucessivos de 5 Objetos (a, b, c, d, e)



h_1 = medida de similaridade (distância) entre a e b...analogamente para as demais, h_2, h_3 e h_4
No eixo horizontal tem-se uma escala da Medida de Similaridade (distância) adotada

Fonte: Vale, Marcos Neves doo (2005). Agrupamentos de Dados: Avaliação de Métodos e Desenvolvimento de Aplicativo para Análise de Grupos. Dissertação de Mestrado. PUC-RJ

36

Algoritmos de Clusterização Hierárquica Critério de Definição dos Grupos mais Próximos

A etapa 4 pode ser realizada de diferentes critérios, gerando tipos diferentes de agrupamentos (clusters):

- . agrupamento de *ligação-simples* (*single-linkage clustering*)
- . agrupamento *deligação-completa* (*complete-linkage clustering*) e
- . agrupamento *deligação-média* (*average-linkage clustering*)

. *ligação-simples* (*single-linkage clustering*) a distância entre clusters é a **menor distância entre membros** de um cluster e do outro (também chamado de conexão ou método do mínimo).

Se os dados consistirem em similaridades, consideraremos que a similaridade entre um cluster e outro é igual à maior similaridade entre qualquer membro de um cluster e outro.

. agrupamento *deligação-completa* (*complete-linkage clustering*) a distância entre clusters é a **maior distância** entre qualquer membro de um cluster e do outro (também chamado de diâmetro ou método do máximo),

. agrupamento de *ligação-média* (*average-linkage clustering*) a distância entre clusters é a **distância média** entre os membros de um cluster e do outro.

Uma variação no agrupamento de ligação-média é o método UCLUS de D'Andrade (1978), que usa a mediana das distâncias, que é muito mais à prova de outliers do que a média.

REFERÊNCIAS:

D'Andrade, R. (1978). U-Statistic Hierarchical Clustering. *Psychometrika*, 4:58-67

Matteucci, M. (2006). Tutorial in Clustering. Politecnico di Milano. Department of Electronics and Information. Milano,

Disponível em 15/05/2006 no endereço: <http://www.elet.polimi.it/upload/matteucci/Clustering/tutorial.html/>

Italy.

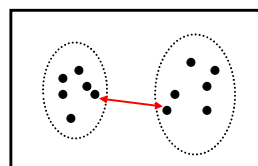
37

Definição dos Grupos mais Próximos

LIGAÇÃO-SIMPLES

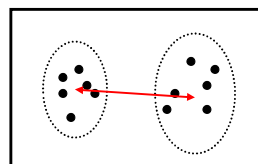
Single Link: Distância entre dois grupos é a distância entre os OBJETOS MAIS PRÓXIMOS.

Também chamado “agrupamento de vizinhos”.



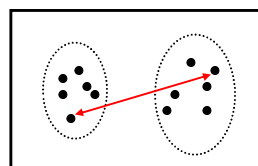
LIGAÇÃO-MÉDIA

Average Link: Distância entre grupos é a distância entre os CENTRÓIDES.



LIGAÇÃO-COMPLETA

Complete Link: Distância entre grupos é a distância entre os OBJETOS MAIS DISTANTES.

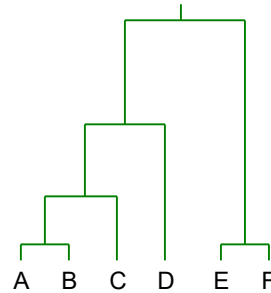
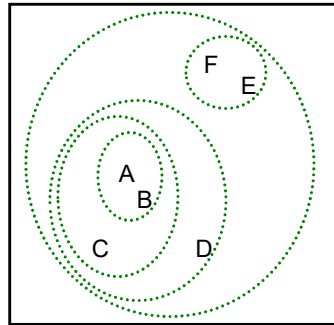


Fonte: Adaptado de Apresentação de “Agrupamento Aglomerativo Hierárquico de Dados”. Prof. Dr. Leandro Augusto da Silva – FCI/Mackenzie

38

Ilustração do Algoritmo

Ilustração do **single-link clustering**
com Distâncias Euclidianas
6 pontos (A, B, C, D, E, F)

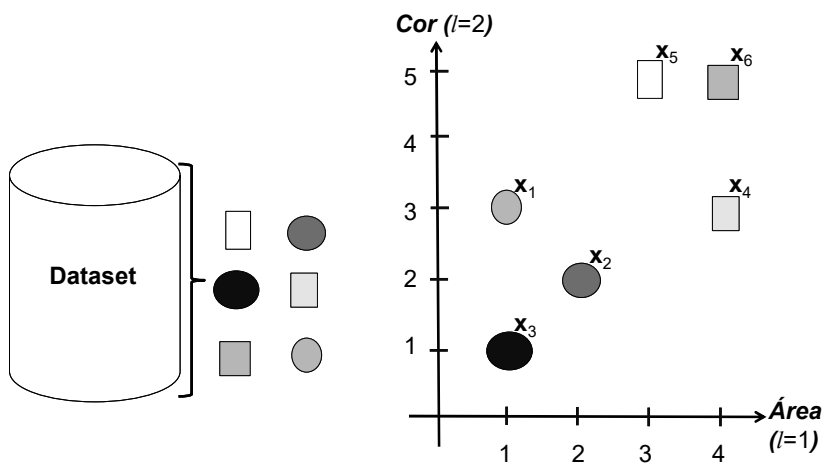


Fonte: Adaptado de Apresentação de "Agrupamento Aglomerativo Hierárquico de Dados". Prof. Dr. Leandro Augusto da Silva – FCI/Mackenzie

39

EXEMPLO 1

6 Objetos (x_i ; $i = 1, 2, \dots, 6$)
com 2 Atributos: Cor e Área do objeto



Fonte: Adaptado de Apresentação de "Agrupamento Aglomerativo Hierárquico de Dados". Prof. Dr. Leandro Augusto da Silva – FCI/Mackenzie

40

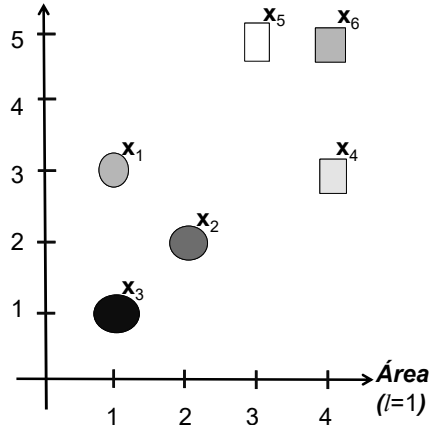
CLUSTERIZAÇÃO



EXEMPLO 1

Calculo da Matriz de Similaridade

Cor ($l=2$)



$$d_{ij} = \left(\sum_{l=1}^L (x_{il} - x_{jl})^2 \right)^{1/2}$$

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0					
x_2	1,4	0				
x_3	2,0	1,4	0			
x_4	3,0	2,8	3,6	0		
x_5	2,8	3,2	4,5	2,2	0	
x_6	3,6	3,6	5,0	2,0	1,0	0

Fonte: Adaptado de Apresentação de "Agrupamento Aglomerativo Hierárquico de Dados". Prof. Dr. Leandro Augusto da Silva – FCI/Mackenzie

41

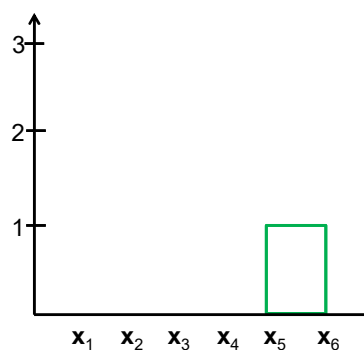
CLUSTERIZAÇÃO



EXEMPLO 1

(Agrupar o par de Exemplos mais próximo – *Single*)

distância



	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0					
x_2	1,4	0				
x_3	2,0	1,4	0			
x_4	3,0	2,8	3,6	0		
x_5	2,8	3,2	4,5	2,2	0	
x_6	3,6	3,6	5,0	2,0	1,0	0

Fonte: Adaptado de Apresentação de "Agrupamento Aglomerativo Hierárquico de Dados". Prof. Dr. Leandro Augusto da Silva – FCI/Mackenzie

42

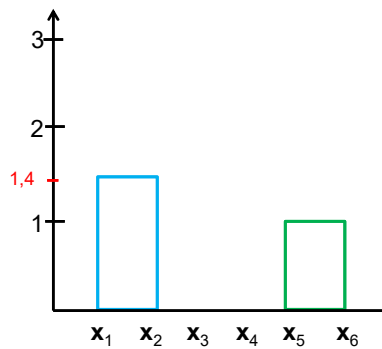
CLUSTERIZAÇÃO



EXEMPLO 1

(Agrupar o par de Exemplos mais próximo – *Single*)

distância



	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0					
x_2	1,4	0				
x_3	2,0	1,4	0			
x_4	3,0	2,8	3,6	0		
x_5	2,8	3,2	4,5	2,2	0	
x_6	3,6	3,6	5,0	2,0	1,0	0

SIMÉTRICOS

Fonte: Adaptado de Apresentação de "Agrupamento Aglomerativo Hierárquico de Dados". Prof. Dr. Leandro Augusto da Silva – FCI/Mackenzie

43

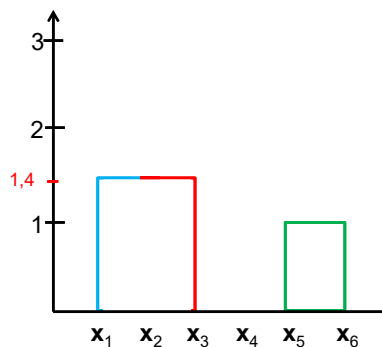
CLUSTERIZAÇÃO



EXEMPLO 1

(Agrupar o par de Exemplos mais próximo – *Single*)

distância



	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0					
x_2	1,4	0				
x_3	2,0	1,4	0			
x_4	3,0	2,8	3,6	0		
x_5	2,8	3,2	4,5	2,2	0	
x_6	3,6	3,6	5,0	2,0	1,0	0

SIMÉTRICOS

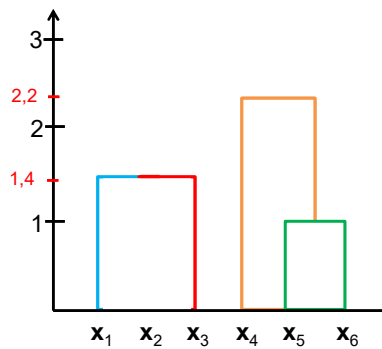
Fonte: Adaptado de Apresentação de "Agrupamento Aglomerativo Hierárquico de Dados". Prof. Dr. Leandro Augusto da Silva – FCI/Mackenzie

44

EXEMPLO 1

(Agrupar o par de Exemplos mais próximo – *Single*)

distância



	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0					
x_2	1,4	0				
x_3		1,4	0			
x_4	3,0	2,8	3,6	0		
x_5	2,8	3,2	4,5	2,2	0	
x_6	3,6	3,6	5,0	2,0	1,0	0

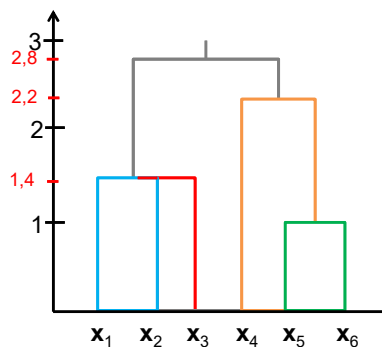
Fonte: Adaptado de Apresentação de "Agrupamento Aglomerativo Hierárquico de Dados". Prof. Dr. Leandro Augusto da Silva – FCI/Mackenzie

45

EXEMPLO 1

(Agrupar o par de Exemplos mais próximo – *Single*)

distância



	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0					
x_2	1,4	0				
x_3		1,4	0			
x_4	3,0	2,8	3,6	0		
x_5	2,8	3,2	4,5	2,2	0	
x_6	3,6	3,6	5,0		1,0	0

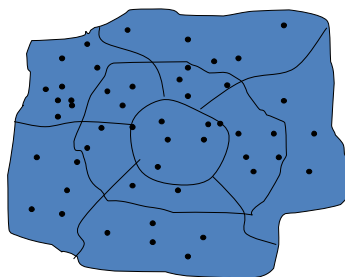
Fonte: Adaptado de Apresentação de "Agrupamento Aglomerativo Hierárquico de Dados". Prof. Dr. Leandro Augusto da Silva – FCI/Mackenzie

46

EXEMPLO 2

Aplicação em Logística – Montagem de “Distritos” Pontos de Demanda em uma Região Geográfica

. Deseja-se definir **setores** (*districts*) de atendimento dos pontos de demanda (entrega ou coleta de produtos)



- . Clusters baseados na proximidade (distância física) entre pontos
- . Busca-se ponto mais próximo de cada cluster e agrega-se ao cluster
- . Critério de **Ligação-Média** (*Average-Link*)
- . Processo é interrompido quando número desejado de clusters é atingido

47

Heurística proposta por Ballou (1994)

*Preliminarmente todos os pontos devem ter suas **coordenadas** definidas*

Passos da Heurística

- Passo 1:** Inicie com todos os pontos e defina o número desejado de clusters. Inicialmente cada ponto se constitui em um cluster
- Passo 2:** Compute as distâncias entre todos os pares de clusters.
- Passo 3:** Identifique o par de clusters com a menor distância e combine este par em um novo cluster. O número de clusters é reduzido em **1**. As coordenadas deste novo cluster corresponderão ao Centróide (Centro de Gravidade) dos pontos pertencentes ao cluster.
- Passo 4:** Repita este procedimento de agregação (passos **2** e **3**) até que o número desejado de clusters seja atingido.

Fonte: Ballou, R. H. (1994) Measuring Transport Costing Error in Customer Aggregation for Facility Location. Transportation Journal, Vol. 33, No. 3, SPRING 1994: pp. 49-59 (<https://www.jstor.org/stable/20713205>)

48

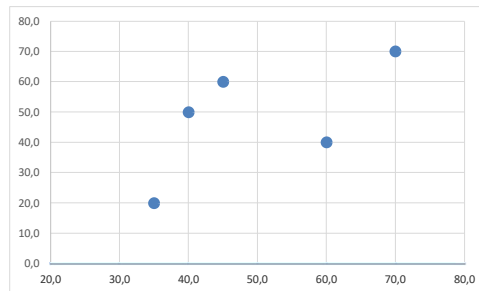
EXEMPLO 2

Pontos de Demanda em uma Região Geográfica

Preliminarmente todos os pontos devem ter suas **coordenadas** definidas
Cada ponto tem uma **Demanda** (peso a entregar)

Dados:

PONTO	X (km)	Y (km)	Peso (t)
1	45,0	60,0	20,0
2	40,0	50,0	40,0
3	35,0	20,0	60,0
4	60,0	40,0	50,0
5	70,0	70,0	30,0
			200,0



Matriz de Distâncias

Cluster	1	2	3	4	5
1	0,0	11,2	41,2	25,0	26,9
2	11,2	0,0	30,4	22,4	36,1
3	41,2	30,4	0,0	32,0	61,0
4	25,0	22,4	32,0	0,0	31,6
5	26,9	36,1	61,0	31,6	0,0

49

Passos da Heurística

ITERAÇÃO 1

Matriz de Distâncias

Cluster	1	2	3	4	5
1	0,0	11,2	41,2	25,0	26,9
2	11,2	0,0	30,4	22,4	36,1
3	41,2	30,4	0,0	32,0	61,0
4	25,0	22,4	32,0	0,0	31,6
5	26,9	36,1	61,0	31,6	0,0

O Centróide dos Clusters (*Centro de Gravidade*) é calculado pela Média Ponderada das Coordenadas dos pontos. Ponderada pela Demanda (*peso a entregar*) de cada ponto

CG do Cluster 1					
Cluster	X (km)	Y (km)	Peso (t)	X.t	Y.t
1	45,0	60,0	20,0	900	1.200
2	40,0	50,0	40,0	1.600	2.000
			Total:	60,0	3.200
			CG	Xcg	Ycg
				41,7	53,3

ITERAÇÃO 2

Matriz de Distâncias

Cluster	1	2	3	4	5
1	0,0		34,0	22,7	32,9
2					
3	34,0		0,0	32,0	61,0
4	22,7		32,0	0,0	31,6
5	32,9		61,0	31,6	0,0

CG do Cluster 1					
Cluster	X (km)	Y (km)	Peso (t)	X.t	Y.t
1	45,0	60,0	20,0	900	1.200
2	40,0	50,0	40,0	1.600	2.000
4	60,0	40,0	50,0	3.000	2.000
			Total:	110,0	5.200
			CG	Xcg	Ycg
				50,0	47,3

50

Passos da Heurística

ITERAÇÃO 3

Matriz de Distâncias

Cluster	1	2	3	4	5
1	0,0		31,1		30,3
2					
3	31,1		0,0		61,0
4					
5	30,3		61,0		0,0

CG do Cluster 1

Cluster	X (km)	Y (km)	Peso (t)	X.t	Y.t
1	45,0	60,0	20,0	900	1.200
2	40,0	50,0	40,0	1.600	2.000
4	60,0	40,0	50,0	3.000	2.000
5	70,0	70,0	30,0	2.100	2.100
Total:			140,0	7.600	7.300
CG				Xcg	Ycg
				54,3	52,1

ITERAÇÃO 4

Matriz de Distâncias

Cluster	1	2	3	4	5
1	0,0		37,5		
2					
3	37,5		0,0		
4					
5					

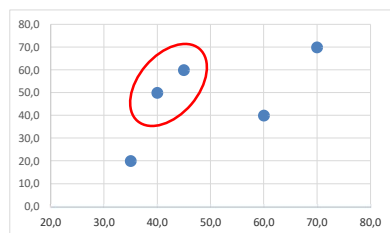
CG do Cluster 1

Cluster	X (km)	Y (km)	Peso (t)	X.t	Y.t
1	45,0	60,0	20,0	900,0	1.200,0
2	40,0	50,0	40,0	1.600,0	2.000,0
3	35,0	20,0	60,0	2.100,0	1.200,0
4	60,0	40,0	50,0	3.000,0	2.000,0
5	70,0	70,0	30,0	2.100,0	2.100,0
Total:			170,0	7.600	6.400
CG				Xcg	Ycg
				44,7	37,6

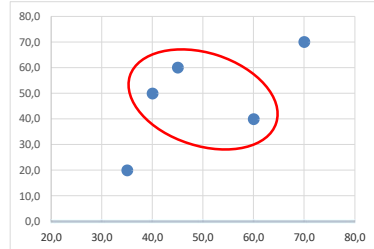
51

Passos da Heurística

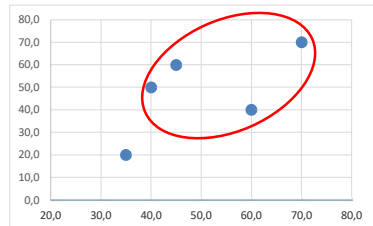
ITERAÇÃO 1



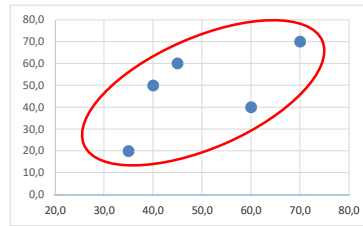
ITERAÇÃO 2



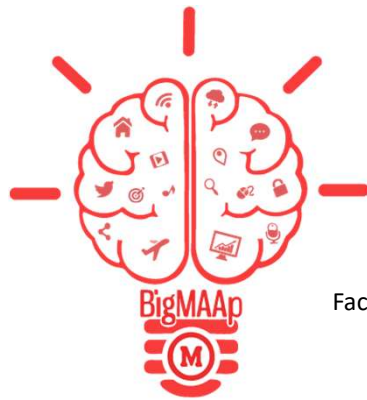
ITERAÇÃO 3



ITERAÇÃO 4



52



BigMAAp
Laboratório de
BIG e Métodos
DATA Analíticos
Aplicados

Prof. Dr. Arnaldo R. A. Vallim Fº
aavallim@mackenzie.br
Faculdade de Computação e Informática