Universidade Federal de Pernambuco – UFPE
Centro de Informática – Cin
Pós-graduação em Ciência da Computação

Clustering: K-means and Aglomerative

Equipe: Hugo, Jeandro, Rhudney e Tiago
Professores: Ricardo Massa e Paulo Maciel

Agenda

- Contexto
- Motivação
- Introdução à Mineração de Dados
- Clusterização
- Algorítmos Particionados
  - K-Means
- Algorítmos Hierarquicos
  - Aglomeração
- Exercícios

Centro de Informática U·F·P·E

Introdução a Mineração de Dados

- Processo de descoberta de novas informações e conhecimento, no formato de regras e padrões, a partir de grandes bases de dados.

Introdução a Mineração de Dados

- Mineração preditiva
  - Deseja-se prever o valor desconhecido de um determinado atributo, a partir da análise histórica dos dados armazenados na base.

- Mineração descritiva
  - Padrões e regras descrevem características importantes dos dados com os quais se está trabalhando.

Introdução a Mineração de Dados

- KDD (Knowledge Discovery in Databases)
  - Processo de extração de informação de uma base de dados
- O processo de KDD é composto por seis fases (Navathe):
  - Seleção dos dados
  - Limpeza dos dados
  - Enriquecimento dos dados
  - Transformação dos dados
  - Mineração dos dados
  - Apresentação e análise dos resultados

Introdução a Mineração de Dados

- Tarefas mais comuns em Mineração de Dados
  - Associação
  - Classificação
  - Clusterização

Introdução a Mineração de Dados

- Tipos de aprendizado
  - Supervisionado
    - Dados rotulados/ base de treinamento já classificada
  - Não-supervisionado
    - Dados não rotulados/ base de treinamento não classificada
      - Coletar e rotular um grande conjunto de exemplos pode custar muito tempo, esforço e dinheiro.

Introdução a Mineração de Dados

- Associação
  - As regras de Associação têm como premissa básica encontrar elementos que implicam na presença de outros elementos em uma mesma transação.

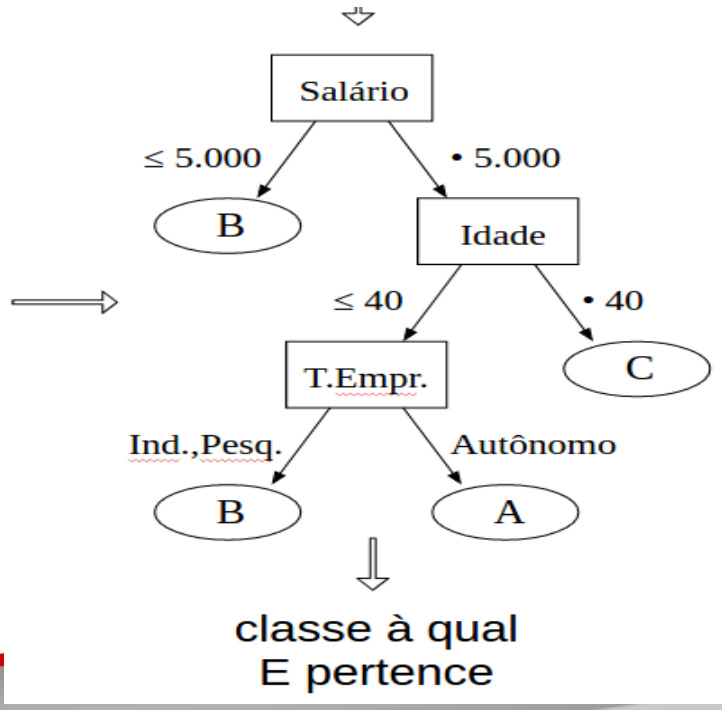
Id-Transação (TID)	Itens Comprados
1	leite, pão, refrigerante
2	cerveja, carne
3	cerveja, fralda, leite, refrigerante
4	cerveja, fralda, leite, pão
5	fralda, leite, refrigerante

{fralda} ⇒ {cerveja} confiança de 66% (suporte médio)
{fralda} ⇒ {leite} confiança de 100% (suporte alto)
{leite} ⇒ {fralda} confiança de 75% (suporte alto)
{carne} ⇒ {cerveja} confiança de 100% (suporte baixo)

Introdução a Mineração de Dados

- Classificação
  - Um classificador identifica, entre um conjunto pré-definido de classes, aquela a qual pertence um elemento, a partir de seus atributos.

ID	Salário	Idade	Tipo Emprego	Classe
1	3.000	30	Autônomo	B
2	4.000	35	Indústria	B
3	7.000	50	Pesquisa	C
4	6.000	45	Autônomo	C
5	7.000	30	Pesquisa	B
6	6.000	35	Indústria	B
7	6.000	35	Autônomo	A
8	7.000	30	Autônomo	A
9	4.000	45	Indústria	B



Introdução a Mineração de Dados

- Clusterização
  - É a tarefa de identificar um conjunto finito de categorias (ou grupos - clusters) que contêm objetos similares.
  - A similaridade é difícil de ser definida

Introdução a Mineração de Dados

- Clusterização
  - É a tarefa de identificar um conjunto finito de categorias (ou grupos - clusters) que contêm objetos similares.

Exemplo: Deseja-se separar os clientes em grupos de forma que aqueles que apresentam o mesmo comportamento de consumo fiquem no mesmo grupo.

Consumidor	Qtd.Tot.	Preço.Méd.
1	2	1.700
2	10	1.800
3	2	100
4	3	2.000
5	12	2.100
6	3	200
7	4	2.300
8	11	2.040
9	3	150

Grupo	Consumidor	Qtd.Tot.	Preço.Méd.
1	1	2	1.700
	4	3	2.000
	7	4	2.300
2	2	10	1.800
	5	12	2.100
	8	11	2.040
3	3	2	100
	6	3	200
	9	3	150

Análise de clusters

- O que é um cluster?
  - Um cluster é uma coleção de objetos de dados que são:
    - Similares aos objetos que estão no mesmo grupo
    - Dissimilar aos objetos que estão em outros grupos
- Análise de clusters
  - Divisão de um conjunto de objetos de dados em clusters
- Tipo de aprendizado não-supervisionado

Cluster Analysis: A Multi-Dimensional Categorization

- Technique-Centered
  - Distance-based methods
  - Density-based and grid-based methods
  - Probabilistic and generative models
  - Leveraging dimensionality reduction methods
  - High-dimensional clustering
  - Scalable techniques for cluster analysis
  - Data Type-Centered
    - Clustering numerical data, categorical data, text data, multimedia data, time-series data, sequences, stream data, networked data, uncertain data
    - Additional Insight-Centered
    - Visual insights, semi-supervised, ensemble-based, validation-based

Tipos de clusterização

- Quanto ao cluster
  - Particionado
    - Objetos são divididos em grupos no mesmo nível, ou seja, sem sobreposição de clusters ou não-aninhados.
    - Ex.: K-means, K-medoids e DBSCAN
  - Hierárquico
    - Os grupos de objetos estão aninhados, ou seja, estão organizados como em uma árvore.
    - Um cluster pode ser formado por sub-clusters.
    - Ex.: Aglomerativos e Divisivos

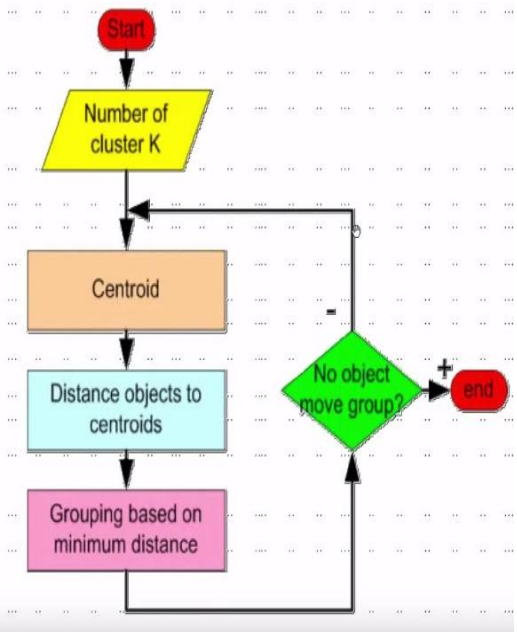
Tipos de clusterização

- Quanto aos dados
  - Exclusivo
    - Quando cada objeto de uma massa de dados está atribuído apenas a um cluster.
  - Sobreposto ou não-exclusivo
    - Quando um objeto pode coexistir em mais de um cluster
  - Fuzzy
    - Um tipo de agrupamento sobreposto em que cada objeto pertença a cada cluster com um grau de pertinência (0% a 100%)

Tipos de clusterização

- Quanto ao particionamento
  - Completo
    - Todos os objetos são atribuídos necessariamente a um cluster
  - Parcial
    - Objetos podem não ser atribuídos a clusters

K-means



- Entrada: n objetos e k clusters
- Saída: n objetos organizados em k clusters

Algoritmo

1. Escolher aleatoriamente k objetos como os centros (centróides) iniciais dos k clusters;
2. Repetir
   a. (re)associar cada objeto ao cluster cujo centróide esteja mais perto;
   b. (re)calcular o centróide de cada cluster como sendo o valor médio dos objetos de cada cluster;

- Até que os centróides permaneçam estáveis

K-means - Entrada de Dados

Matriz de Dados: contém os valores dos p atributos que caracterizam cada um dos n objetos.

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

K-means - Entrada de Dados

Matriz de Dissimilaridade (distâncias): contém as distâncias entre cada par de objetos.

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \ldots & \ldots & 0 \end{bmatrix}$$

K-means - Função Objetivo

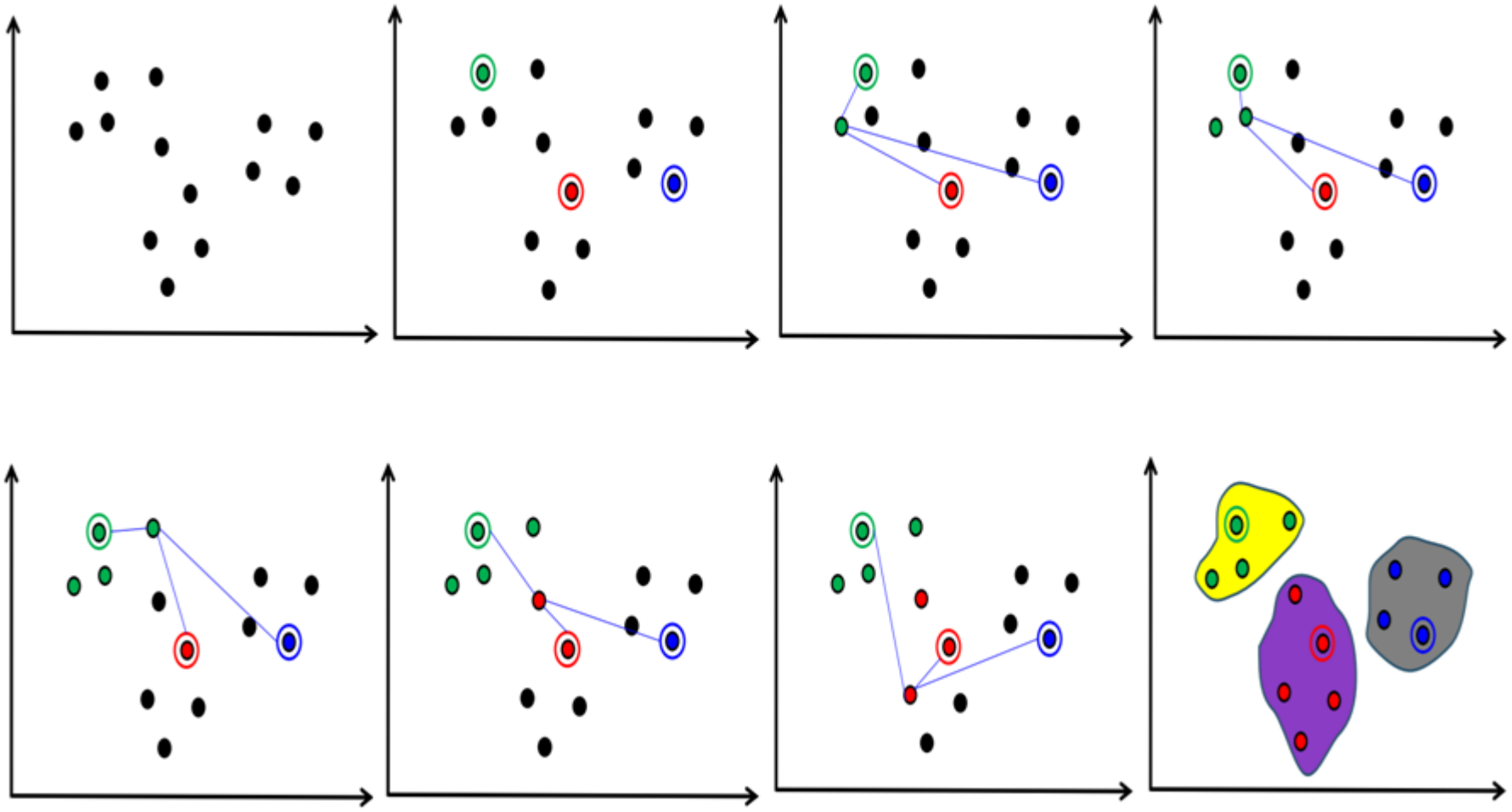
- Para análise da qualidade do cluster, é utilizado a Soma dos Erros Quadrados (sum of the squared error)

$$\mathrm{SSE} = \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} dist(\mathbf{c}_i, \mathbf{x})^2$$

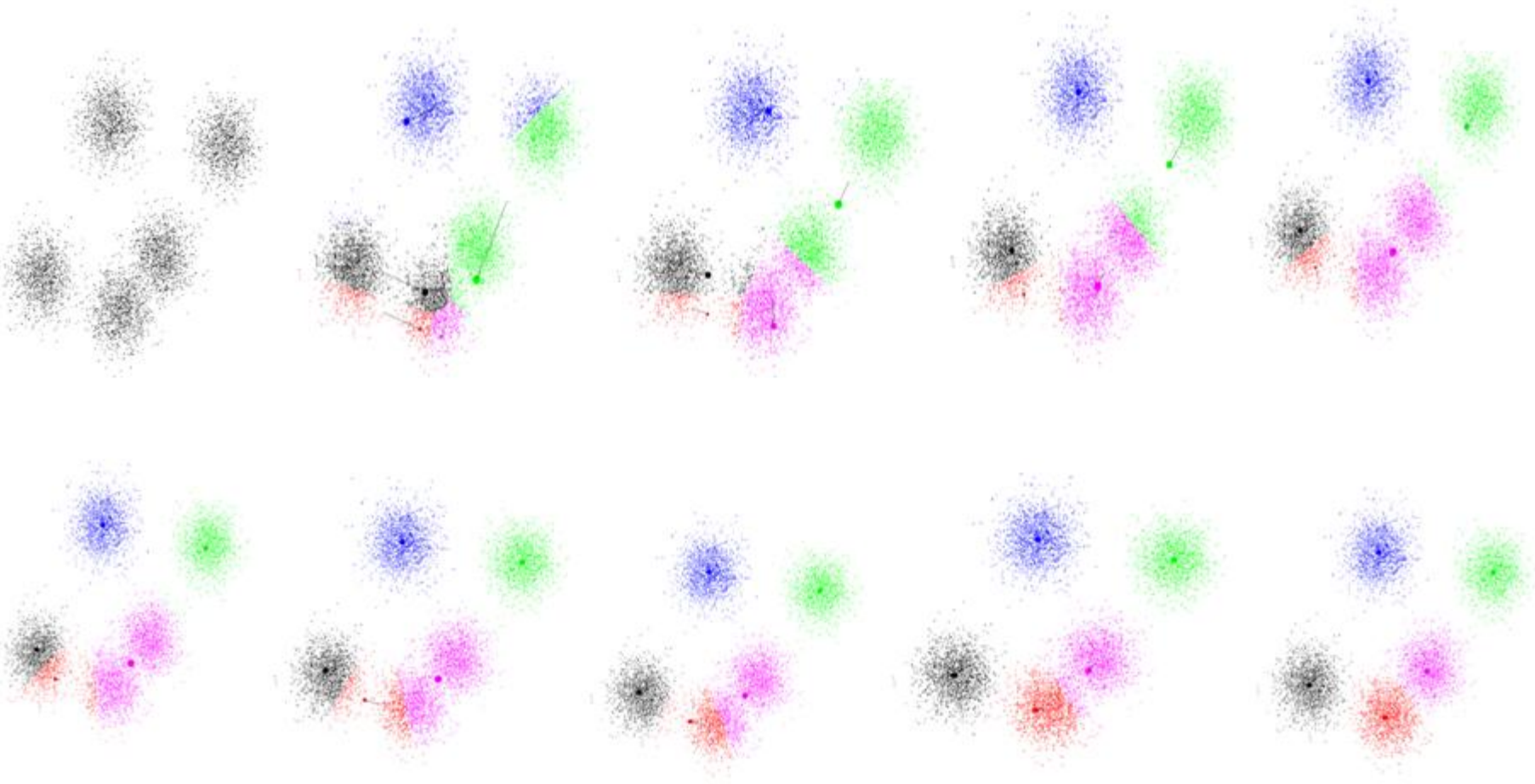
- Calcular o centróide de cada cluster como sendo o valor médio dos objetos de cada cluster.

$$\mathbf{c}_i = \frac{1}{m_i} \sum_{\mathbf{x} \in C_i} \mathbf{x}$$

K-means - Exemplo



K-means - Exemplo

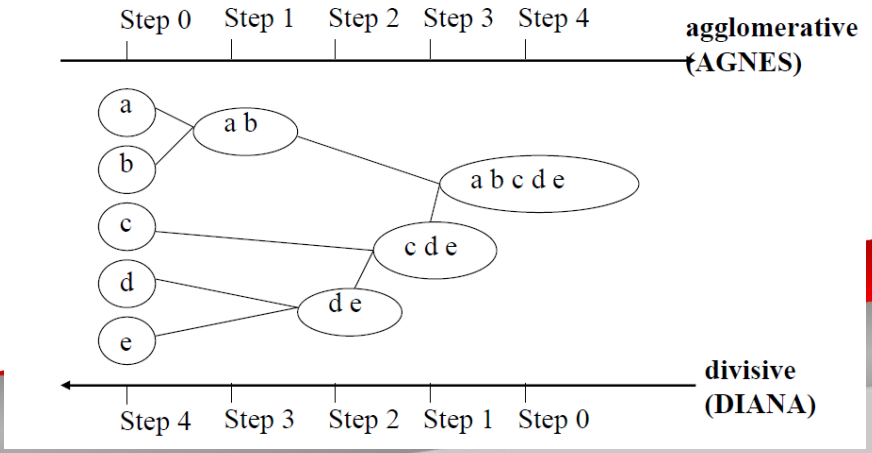


K-means - Desvantagens

- Definição inicial dos centróids
- Sensível ruídos
- Dados convexos

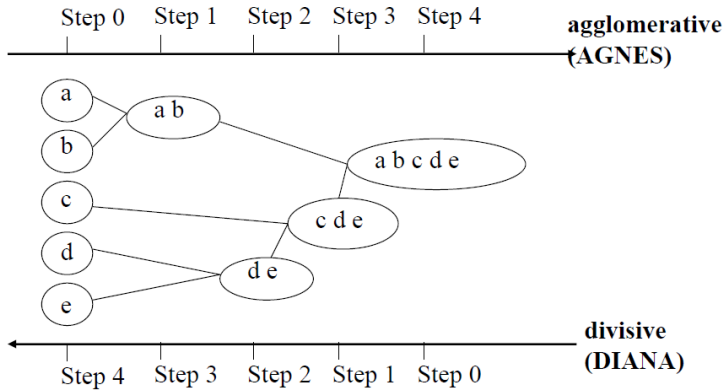
Métodos de Clusterização Hierárquicos

- Cluster hieráquico
  - Gera um agrupamento hieráquico (plotado como um dendrograma)
  - Não necessita especificar o k (número de clusters)
  - Mais determinístico
  - Refinamento não iterativo



Métodos de Clusterização Hierárquicos

- Duas categorias de algoritmos:
  - Aglomerativos: inicia com clusters *singleton,* continuamente unem dois clusters para formar uma hierarquia ***bottom-up***.
  - Divisivos: inicia com um “macro-cluster”, divide continuamente em dois grupos, gerando uma hierarquia ***top-down.***



Questão

Em qual dos seguintes cenários é mais favorável aplicar o particionamento hierárquico ao invés do de nível simples ?

a) Clusterizar documentos por tópicos, onde seja possível organizar a hierarquia por tópicos (e.g Ciência da Computação → Mineração de Dados → Clusterização)

b) Organismos clusterizados baseados em suas características(e.g aparência, comportamento, constituição genética)

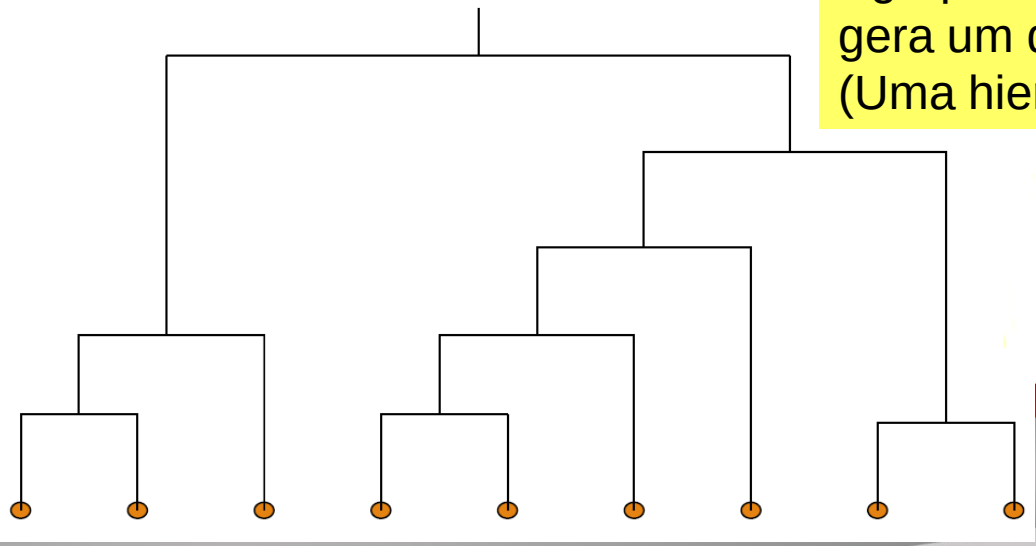
c) Clusterização de pessoas por geolocalização, onde deve examinar-se a população em diferentes níveis de resolução(e.g, país→ estado → município)

Dendrograma

- Dendograma: decompõe um conjunto de objetos em uma árvore de clusters em mutiníves com particionamentos aninhados

- Um agrupamento de objetos é obtido pelo corte no dendrograma no nível desejado, assim cada componente conectado forma um cluster.

Agrupamento hierárquico gera um dendrograma. (Uma hierarquia de grupos)



Matriz de Distância

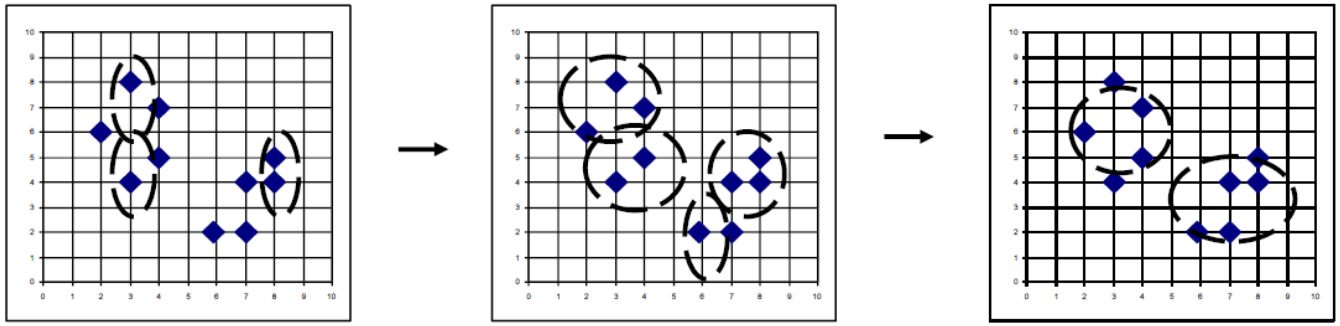
- Matriz que representa a distância, par-a-par, entre os elementos de dois conjuntos.
  - Os valores da diagonal principal são sempre zero
  - Os valores não diagonais são sempre positivos
  - A matriz é simétrica

	a	b	c
a	0	1	3
b	1	0	4
c	3	4	0

Algoritmos de Clusterização Aglomerativos

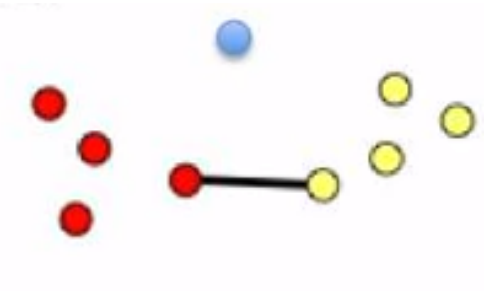
Algoritmos de Clusterização Aglomerativos

- Clusterização aglomerativa varia em diferentes medidas de similaridade entre clusters
  - Single Link (vizinho mais próximo)
  - Complete Link (diâmetro)
  - Average Link (média do grupo)
  - Centroid Link (Similaridade do Centróide)



Algoritmos de Clusterização Aglomerativos

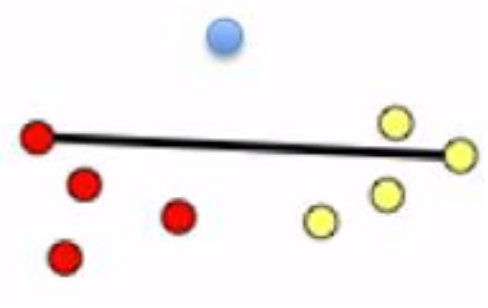
- Single Link
  - A similaridade entre dois clusters é a similaridade entre seus membros mais similares (vizinho mais próximo)
  - Baseado em similaridade local: ênfase mais em regiões fechadas, ignorando a estrutura completa do cluster
  - Sensível a ruídos e outliers



$$D(c_1, c_2) = \min_{x_1 \in c_1, x_2 \in c_2} D(x_1, x_2)$$

Algoritmos de Clusterização Aglomerativos

- Complete Link
  - A similaridade entre dois clusters é a similaridade entre seus membros mais dissimilares
  - Unir dois clusters para formar um com menor diâmetro
  - Não localidade no comportamento, obtendo clusters compactos
  - Sensível a outliers



$$D(c_1, c_2) = \max_{x_1 \in c_1, x_2 \in c_2} D(x_1, x_2)$$

Algoritmos de Clusterização Aglomerativos : R

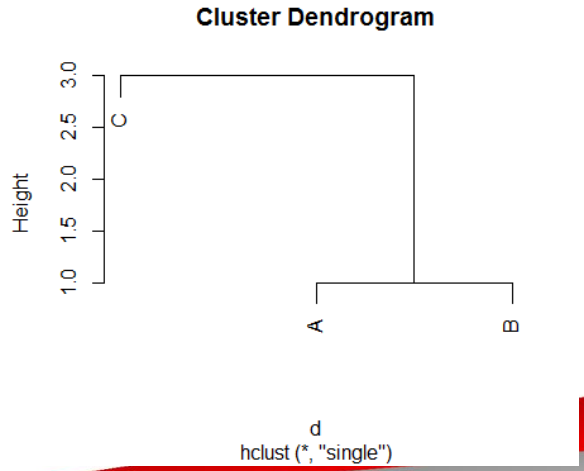
- AGNES (AGglomerative NESting)
  - Utiliza o método single-link e matriz de dissimilaridade
  - Continuamente une os nós que possuem a menor dissimilaridade
  - Eventualmente todos os nós pertencem ao mesmo cluster

Algoritmos de Clusterização Aglomerativos : R

- Exemplo

	a	b	c
a	0	1	3
b	1	0	4
c	3	4	0

d	k	K
0	3	{a},{b},{c}
1	2	{a,b},{c}
2	2	{a,b}, {c}
3	1	{a,b,c}



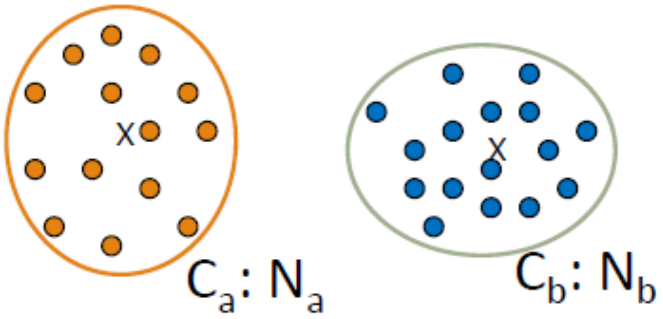
Algoritmos de Clusterização Aglomerativos : R

- Pacote hclust

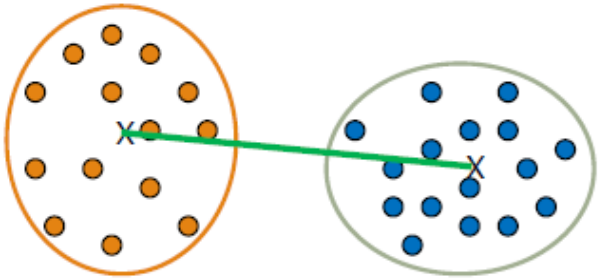
```
> tb = read.csv(“file.csv”)
> tb2 = matrix(
        c(0,1,3,1,0,4,3,4,0),
        nrow = 3,
        )

> d = as.dist(tb)
> hclust(d, method=“single”)

> d2 = as.dist(tb2)
> plot(hclust(tb2))
```

Algoritmos de Clusterização Aglomerativos

- Link médio:
  - Link médio: a distância média entre um elemento em um cluster e um elemento em outro
  - Alto custo computacional



- Link Centróide:
  - Link Centróid: distância entre os centróides de dois clusters



Algoritmos de Clusterização Aglomerativos

- Group Averaged Agglomerative Clustering (GAAC)
  - Permite dois clusters C_a e C_b se unir resultando em $C_{aUb.}$ O novo centróide é:

$$c_{a\cup b} = \frac{N_a c_a + N_b c_b}{N_a + N_b}$$

  - N_a é a cardinalidade do cluster $C_{a,}$ e c_a é o centróide de C_a
  - A similaridade medida para o GAAC é a média de suas distâncias

Algoritmos de Clusterização Aglomerativos

- Clusterização aglomerativa com critério de Ward
  - Critério de Ward: o aumento no critério do SSE para a clusterização obtida pela junção de

$C_a \cup C_b$: $W(C_{a\cup b}, c_{a\cup b}) - W(C, c) = \frac{N_a N_b}{N_a + N_b} d(c_a, c_b)$

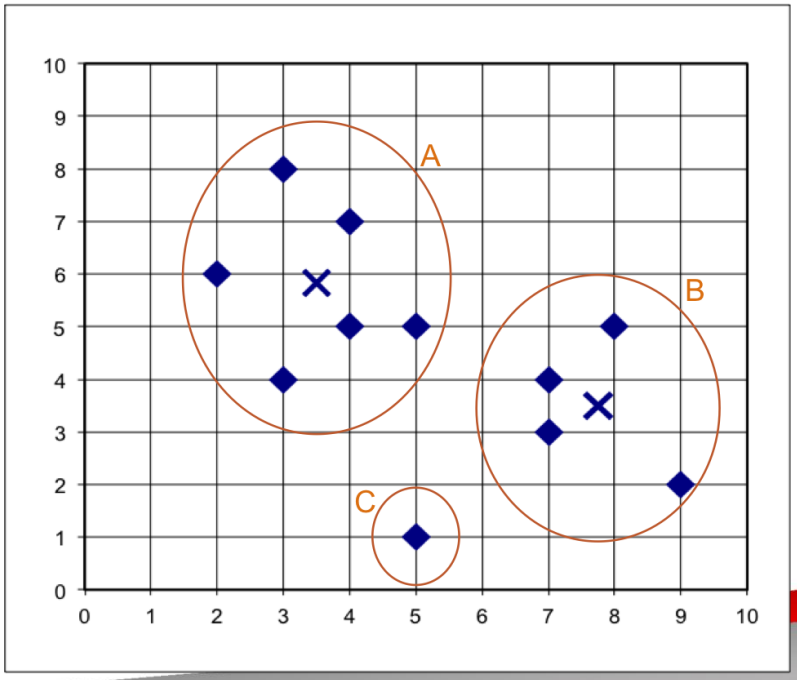
Pacotes/Funções no R para Clusterização Hierárquica

- Hclust(stat)
  - https://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html
- Tsclust (Clusterização em Séries Temporais)
  - Inclui modelos para as medidas de dissimilaridade
  - https://cran.r-project.org/web/packages/TSclust/index.html

Exercício 1

Quais clusters seriam aglomerados primeiro usando o algoritmo aglomerativo de Complete e Single link?

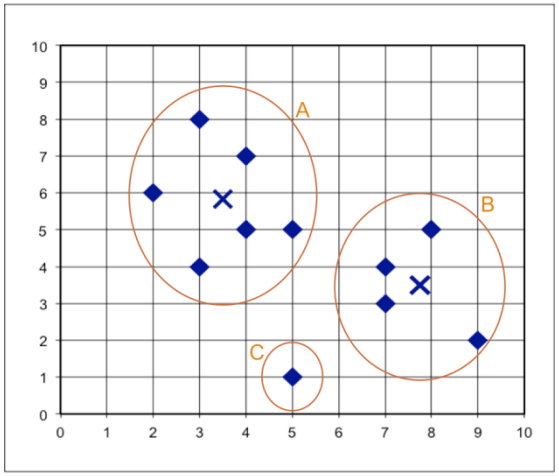
Obs: Use distância Euclidiana como medida de similaridade



Resposta 1

Single Link:

$$d(A,B) = dE((5,5),(7,4)) = \sqrt{(5-7)^2 + (5-4)^2} = \sqrt{5}$$
$$d(A,C) = dE((5,5),(5,1)) = \sqrt{(5-5)^2 + (5-1)^2} = \sqrt{16}$$
$$d(B,C) = dE((7,3),(5,1)) = \sqrt{(7-5)^2 + (3-1)^2} = \sqrt{16}$$



Complete Link:

$$d(A,B) = dE((3,8),(9,2)) = \sqrt{(3-9)^2 + (8-2)^2} = \sqrt{72}$$
$$d(A,C) = dE((3,8),(5,1)) = \sqrt{(3-5)^2 + (8-1)^2} = \sqrt{53}$$
$$d(B,C) = dE((8,5),(5,1)) = \sqrt{(8-5)^2 + (5-1)^2} = \sqrt{25}$$

Exercício 2

Dado a matriz de distância abaixo, preencha o restante da tabela para Single e Complete Link e plote os dendogramas no R.

	A	B	C	D
A	0	1	4	5
B	1	0	2	6
C	4	2	0	3
D	5	3	6	0

d	k	K
0	4	{A},{B},{C},{D}
1	3	{A,B},{C},{D}
2	...	...

Resposta 2

Single Link

d	k	K
0	4	{A},{B},{C},{D}
1	3	{A,B},{C},{D}
2	2	{A,B,C}, {D}
3	1	{A,B,C,D}

Complete Link

d	k	K
0	4	{A},{B},{C},{D}
1	3	{A,B},{C},{D}
2	3	{A,B},{C},{D}
3	2	{A,B},{C,D}
4	2	{A,B},{C,D}
5	2	{A,B},{C,D}
6	1	{A,B,C,D}