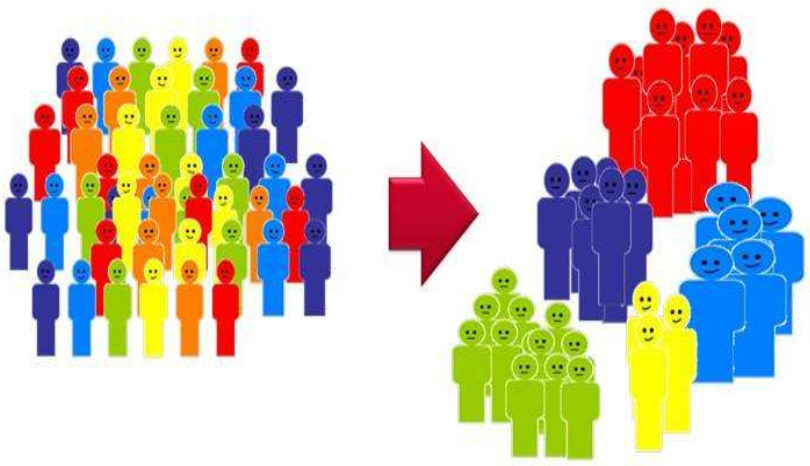


Análise de Agrupamento Cluster

Aprendizado Não Supervisionado

Técnica Exploratória

24

Análise de Agrupamentos (Cluster)

- Agrupamento: uma coleção de objetos de dados
  - Similares aos objetos do mesmo grupo
  - Diferentes dos objetos dos demais grupos
- Análise de Agrupamento
  - Agrupar de um conjunto de objetos de dados em clusters (grupos)
- Agrupamento corresponde à uma classificação não supervisionada, sem classes pré-definidas.
- Aplicações típicas
  - Como uma análise isolada, obter conhecimento sobre a distribuição de objetos de dados.
  - Como um passo anterior para outros algoritmos para preparação para outras análises, reduzindo a complexidade de análise ao reduzir uma população em subgrupos memores que podem ser analisados separadamente.

25

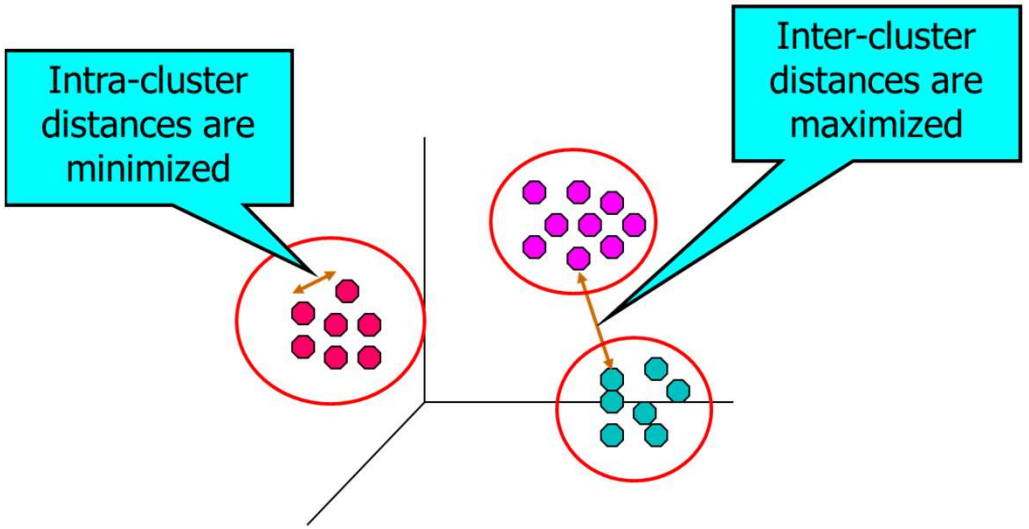
Aplicações da Análise de Cluster

- **Marketing:** descobrir grupos distintos de clientes e depois utilizar este conhecimento para aplicar diferentes estratégias de relacionamento.
- **Uso do Solo**: identificar áreas com uso semelhante do solo a partir de uma base de dados geo referenciada.
- **Seguros:** identificar grupos distintos de risco para determinação de níveis de custos em apólices.
- **Planejamento Urbano**: identificar grupos de edificações de acordo com seu tipo, valor e localização geográfica.
- **Programas Sociais**: identificar grupos distintos de cidadãos de acordo com suas características e necessidades para aplicação de programas sociais.
- E muitas outras áreas como Saúde, Educação e Política.

26

O que é um bom Agrupamento

- Um bom método de agrupamento produzirá grupos com:
  - Alta similaridade intragrupo;
  - Baixa similaridade intergrupo.
- Uma definição precisa para a qualidade do agrupamento é muito difícil:
  - Depende da aplicação (problema)
  - É subjetiva



27

Análise de Agrupamento

- É uma técnica exploratória.
- O resultado depende da escola das variáveis, da definição do número de clusters e/ou das “sementes”.
- Não apresenta uma solução ótima global ou única.
- Sempre gerará uma solução.
- A interpretação dos resultados tem forte componente subjetivo (conhecimento do negócio)
- É uma técnica de Aprendizado de Máquina Não Supervisionado.

28

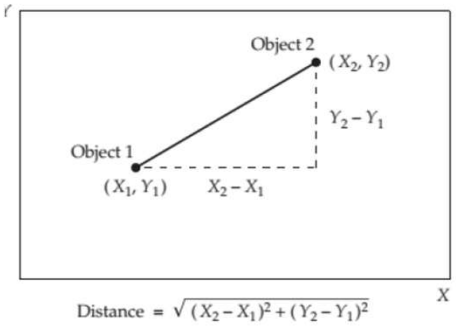
Como medir a similaridade?

Lealdade à Loja
Lealdade à Marca

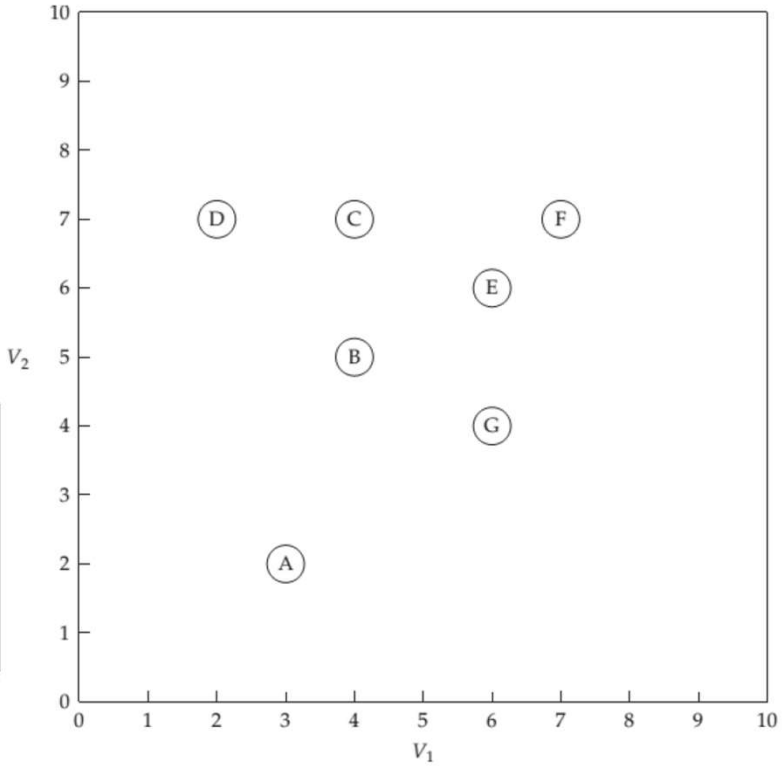
Clustering Variable	Respondents						
	A	B	C	D	E	F	G
V_1	3	4	4	2	6	7	6
V_2	2	5	7	7	6	7	4

Como medir a Similaridade?

Distância Euclidiana



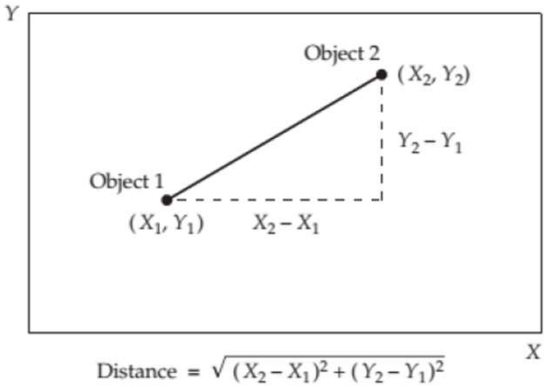
Distance = $\sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$



Fonte: HAIR JR., J. F. et al. Análise multivariada de dados. 6. ed. Porto Alegre: Bookman, 2009. 688 p.

29

Distância Euclidiana



	Observation						
Observation	**A**	**B**	**C**	**D**	**E**	**F**	**G**
A	—						
B	3.162	—					
C	5.099	2.000	—				
D	5.099	2.828	2.000	—			
E	5.000	2.236	2.236	4.123	—		
F	6.403	3.606	3.000	5.000	1.414	—	
G	3.606	2.236	3.606	5.000	2.000	3.162	—

30

Similaridade e Diferença entre Objetos

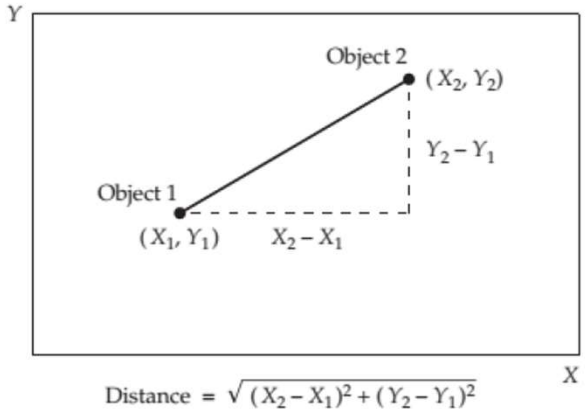
- Distância euclidiana (p = 2):

$$d(i,j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + ... + |x_{ip} - x_{jp}|^2)}$$

- Propriedades da métrica *d(i, j)*:
  - $d(i, j) \geq 0$
  - $d(i, i) = 0$
  - $d(i, j) = d(j, i)$
  - $d(i, j) \leq d(i, k) + d(k, j)$

31

Distância Euclidiana



Distancia Euclidiana no Espaço Multidimensional

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}.$$

dist(exemplo4[2:6], method="euclidean")

Similarity Measure: Euclidean Distance

	Case						
Case	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1	*nc*						
2	3.32	*nc*					
3	6.86	6.63	*nc*				
4	10.25	10.20	6.00	*nc*			
5	15.78	16.19	10.10	7.07	*nc*		
6	13.11	13.00	7.28	3.87	3.87	*nc*	
7	11.27	12.16	6.32	5.10	4.90	4.36	*nc*

nc = distances not calculated.

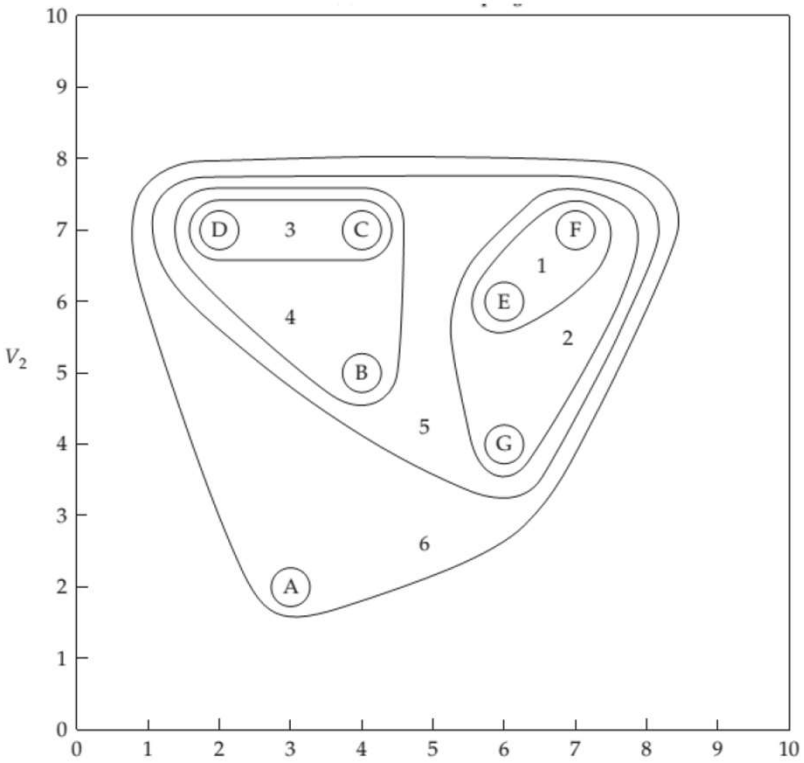
32

Como formar Clusters?

Clustering Variable	Respondents						
	A	B	C	D	E	F	G
Lealdade à Loja V_1	3	4	4	2	6	7	6
Lealdade à Marca V_2	2	5	7	7	6	7	4

Como formar Clusters?

Método Hierárquico Aglomerativo com Nearest Neighbour Approach ("Single Linkage")



Fonte: Hair et al. (2009)

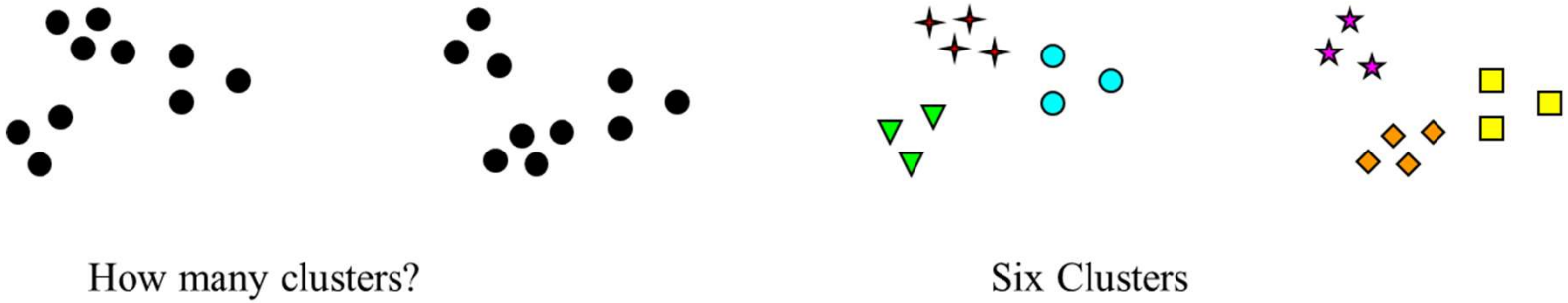
33

Requisitos para Agrupamento

- Escalabilidade
- Habilidade para lidar com diferentes tipos de atributos
- Descobrir grupos com formatos arbitrários
- Mínimo conhecimento do domínio é requerido para determinar os parâmetros de entrada
- Habilidade para lidar com ruídos e anomalias (*outliers*)
- Insensibilidade à ordem dos dados
- Robustez vs. alta dimensionalidade
- Incorporação de restrições especificadas pelo analista
- Interpretabilidade e usabilidade

34

O Agrupamento pode ser Ambíguo



How many clusters?

Six Clusters

Two Clusters

Four Clusters

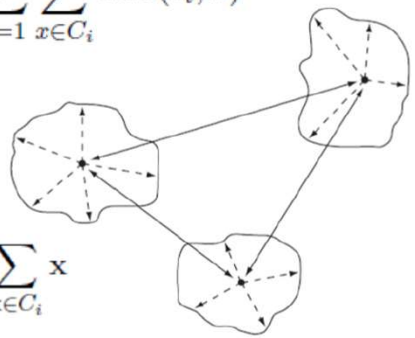
35

Quantos Grupos Formar?

Critério da Soma de Distancias ao Quadrado (wss)

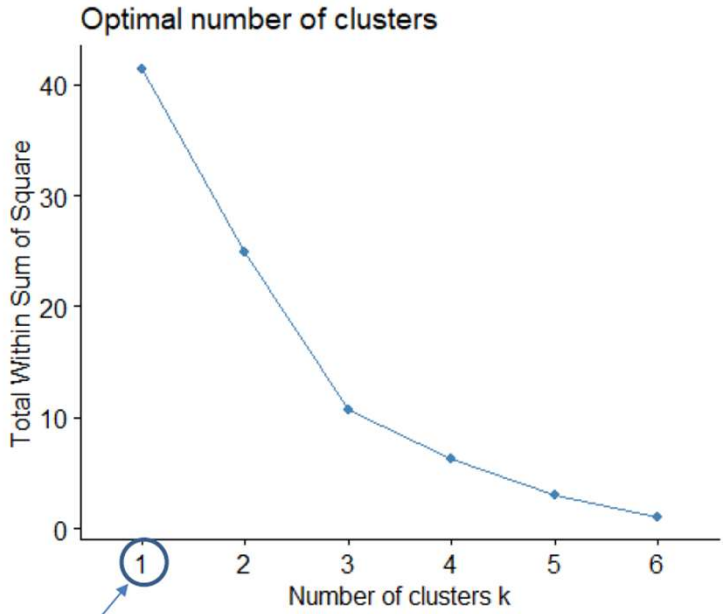
WSS

$$\text{SSE} = \sum_{i=1}^{K} \sum_{x \in C_i} dist(c_i, x)^2$$



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

Fonte: TAN, P.; STEINBACH, M.; KUMAR, V. (2016)



Um grupo

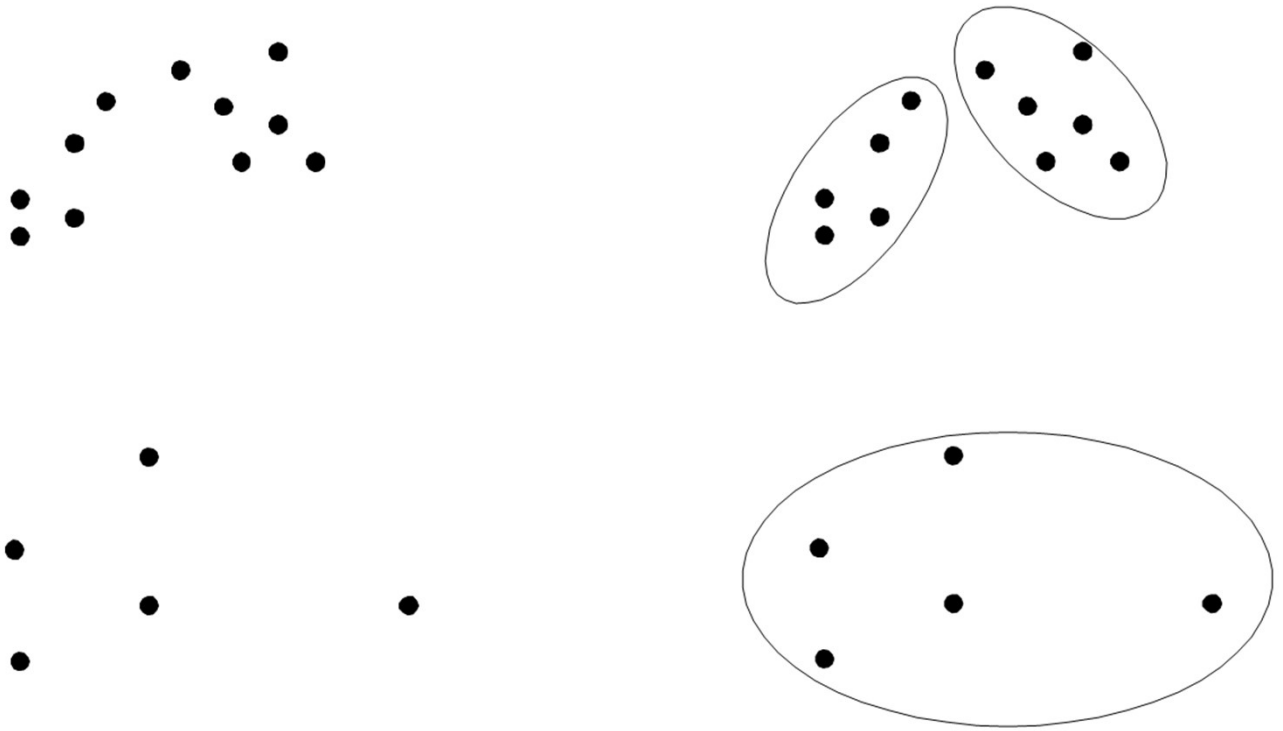
36

Principais Técnicas de Agrupamento

- **Particionamento**: construir várias partições e depois avalia-las com algum critério (proximidade às sementes ou protótipos);
- **Hierarquia**: criar uma decomposição hierárquica do conjunto de objetos usando algum critério (formam árvores ou hierarquias);
- **Modelo**: criar modelos para cada cluster e achar os melhores modelos estatísticos para os objetos.
- **Densidade**: guiadas por funções de conectividade e densidade (proximidade usando algum critério).

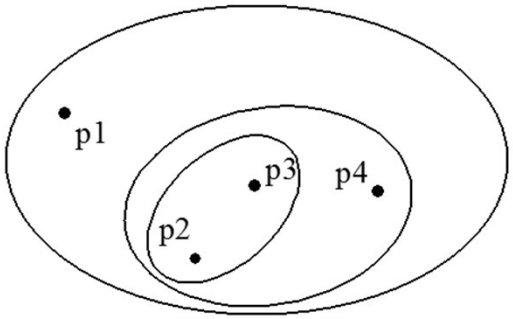
37

Particionamento



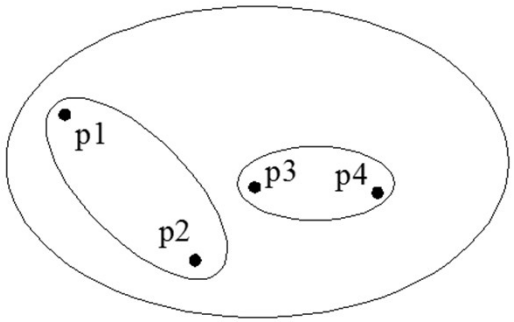
38

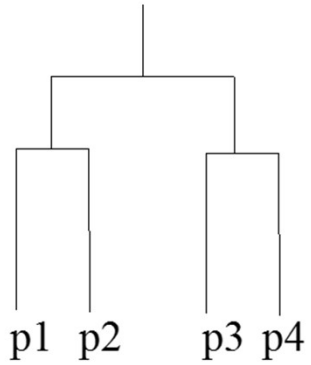
Hierárquico



Traditional Hierarchical Clustering

Traditional Dendrogram





Non-traditional Hierarchical Clustering

Non-traditional Dendrogram

39

Algoritmos de Particionamento

- Método de Particionamento: construir uma partição de uma base de dados **D** com **n** objetos dentro de um conjunto de **k** clusters.
- Dado **k** (número de clusters), achar a partição de k clusters que otimiza o critério de particionamento:
  - Otimização global: exaustivamente enumera todas as partições;
  - Métodos heurísticos: algoritmos K-means e K-medoids;
  - **K-means** (MacQueen, 1967): cada cluster é representado pelo centro do cluster (centróide);
  - **K-medoids** (Kaufman & Rousseew, 1987) cada cluster é representado por dois objetos do cluster.

40

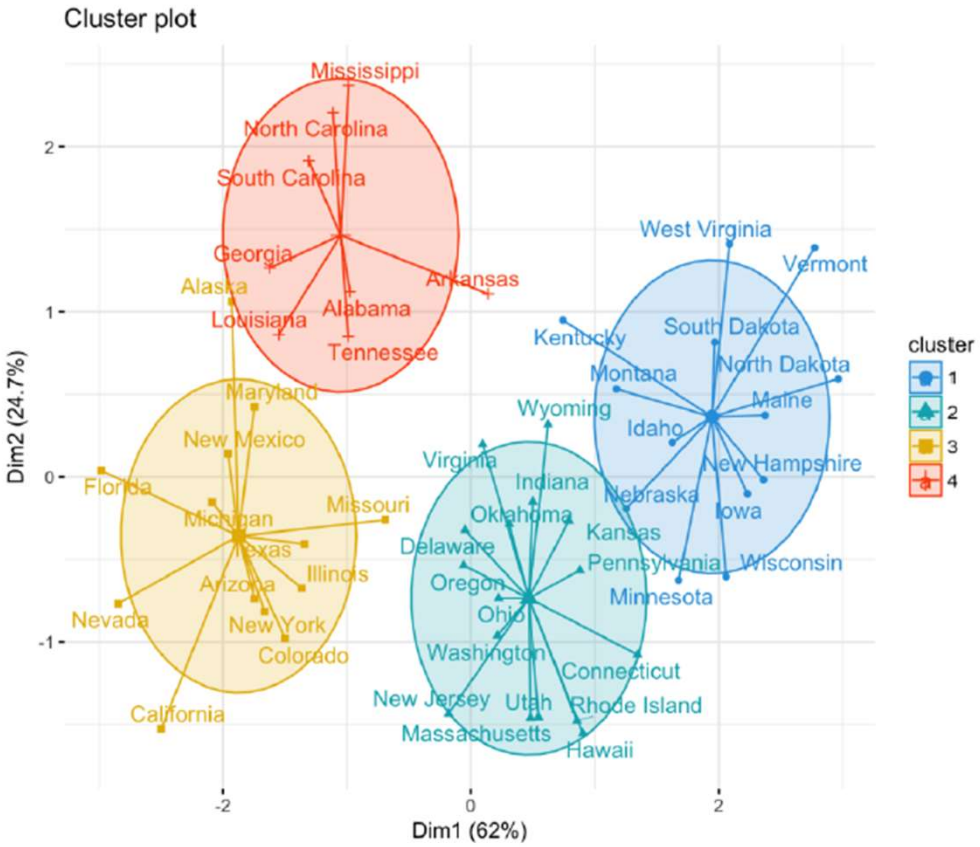
K-Means Clustering

- Dado **k** (número de clusters), o algoritmo consiste em quatro etapas:
  1. Selecionar os centroides iniciais aleatoriamente;
  2. Assinalar cada objeto ao cluster com o centroide mais próximo.
  3. Recalcular cada centroide como uma média de objetos assinalados a ele e reposicionar o centroide.
  4. Repetir as duas etapas anteriores até que não haja alterações (o centroide não se mova).

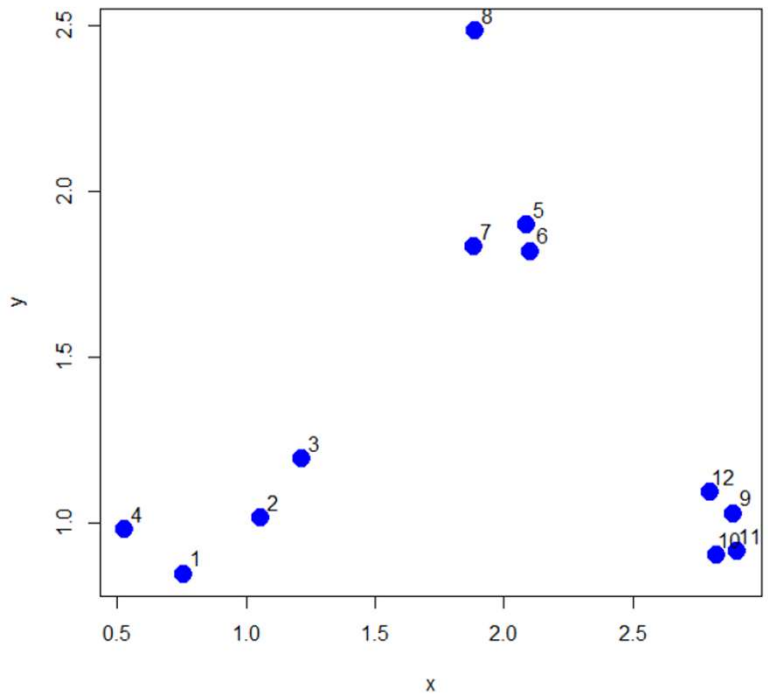
41

K-Means Clustering

- Exemplo:



Etapa 1



- exemplo2.csv

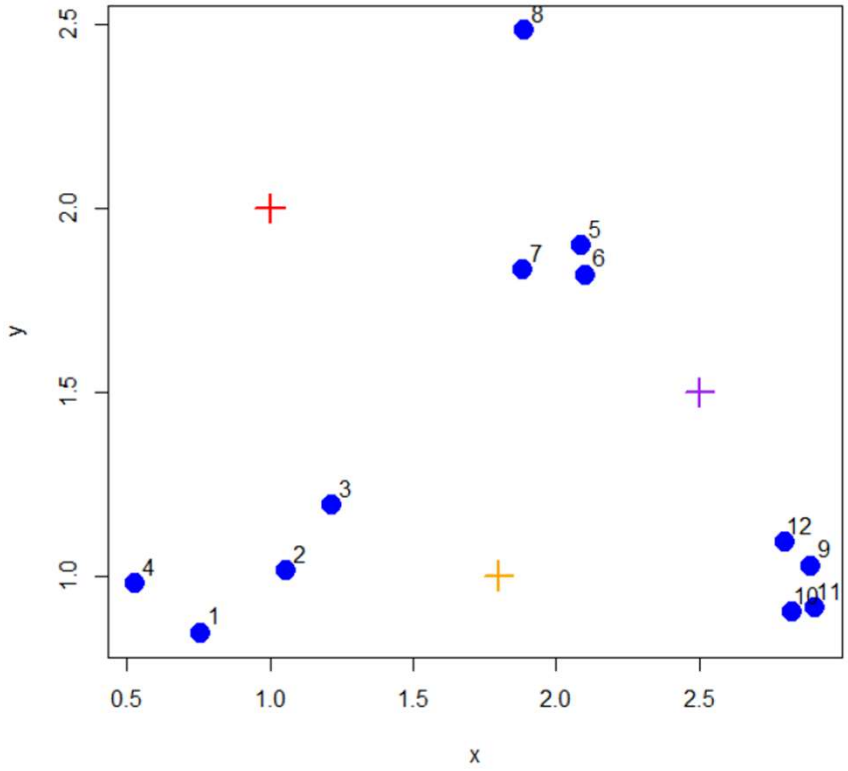
Etapa 1 – Definição do k

Serão considerados 3 clusters na análise

```
exemplo2 <- read.csv2("exemplo2.csv")
scatterplot(y~x, regLine=FALSE, smooth=FALSE, id=list(method='mahal', n = 12),
boxplots=FALSE, data=exemplo2)
```

43

Etapa 2

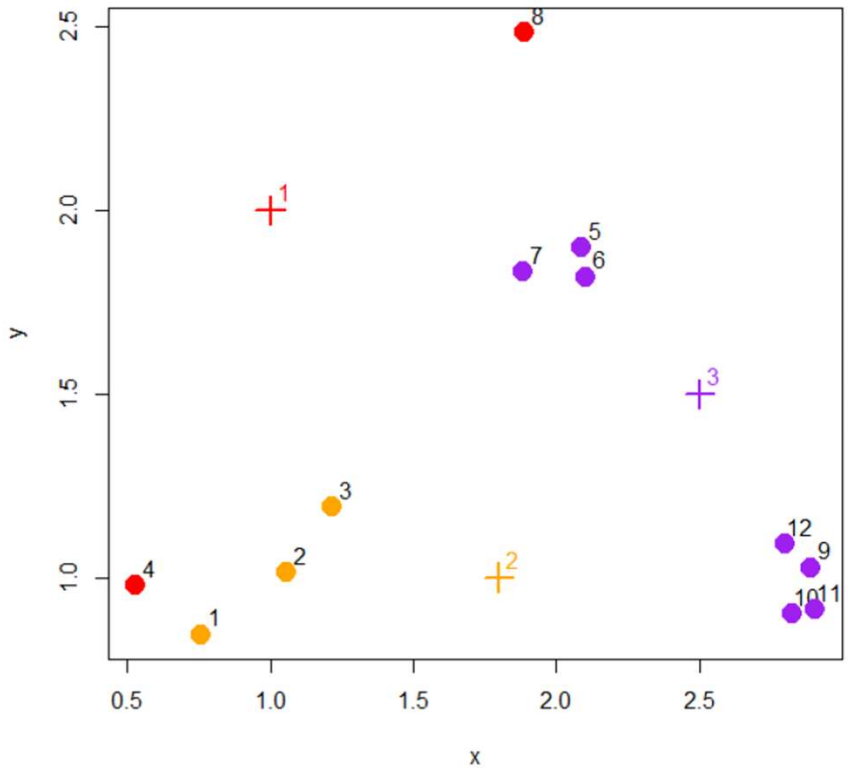


Etapa 2 – Centroides Iniciais

```
cx <- c(1,1.8,2.5)
cy <- c(2,1,1.5)
cols1 <- c("red","orange","purple")
points(cx,cy,col=cols1,pch=3,cex=2,lwd=2)
text(cx + 0.05, cy + 0.05, labels =
as.character(1:3), col=cols1)
```

44

Etapa 3

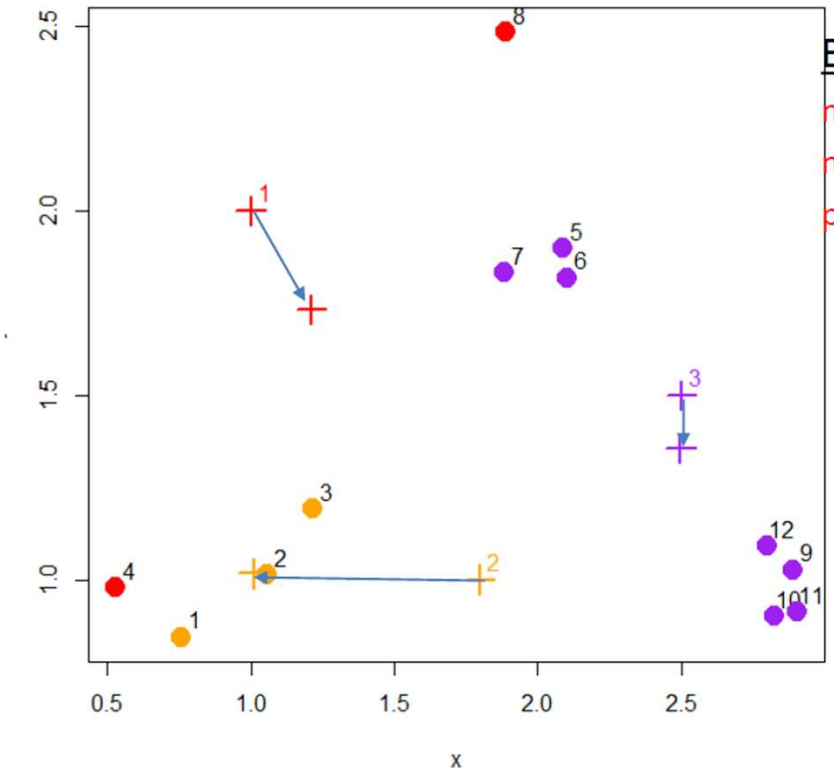


Etapa 3 – Designa os pontos aos Centroides mais Próximos

```
distTmp <- matrix(NA,nrow=3,ncol=12)
distTmp[1,] <- (exemplo2$x-cx[1])^2 + (exemplo2$y-cy[1])^2
distTmp[2,] <- (exemplo2$x-cx[2])^2 + (exemplo2$y-cy[2])^2
distTmp[3,] <- (exemplo2$x-cx[3])^2 + (exemplo2$y-cy[3])^2
distTmp
newClust <- apply(distTmp,2,which.min)
points(exemplo2$x,exemplo2$y,pch=19,cex=2,col=cols1[newClust])
```

45

Etapa 4



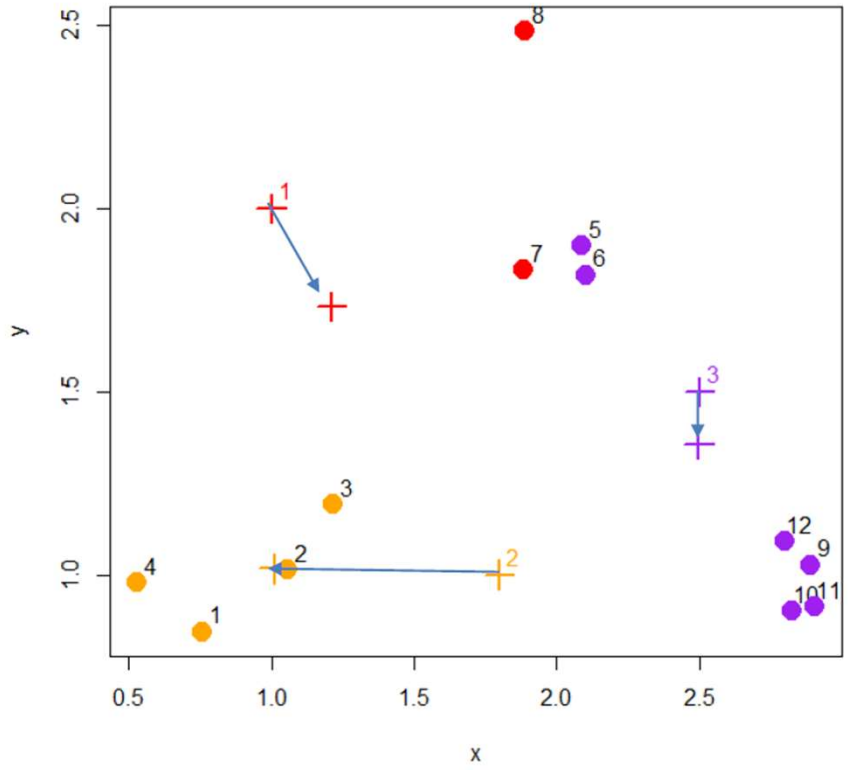
Etapa 4 – Recalcula os Centróides

```
newCx <- tapply(exemplo2$x,newClust,mean)
newCy <- tapply(exemplo2$y,newClust,mean)
points(newCx,newCy,col=cols1,pch=3,cex=2,lwd=2)
```

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

46

Etapa 3/2

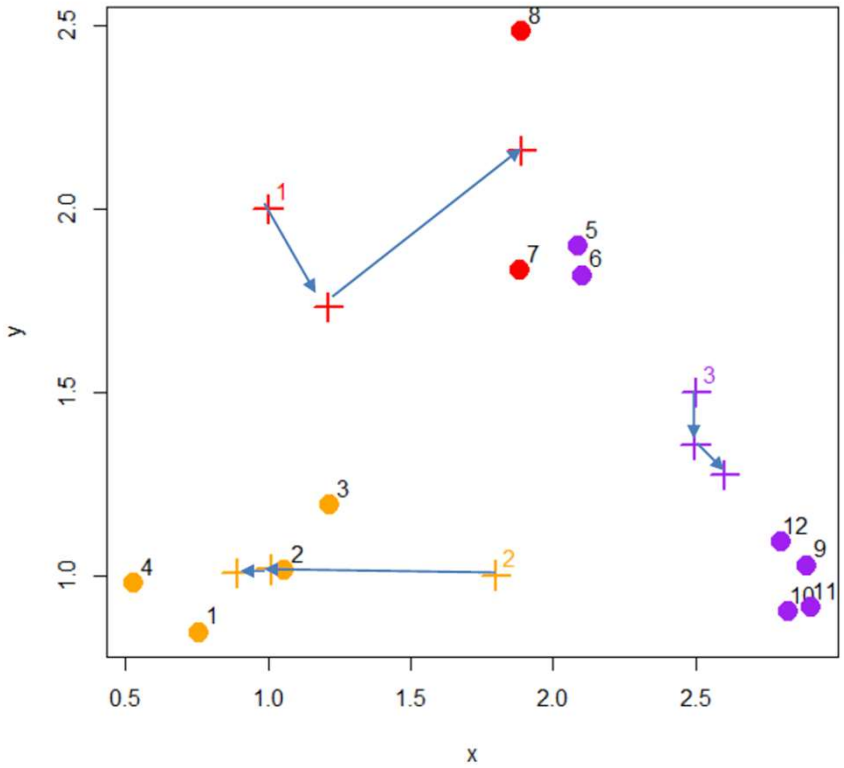


Etapa 3(2) – Designa os pontos aos Centroides mais Próximos

```
distTmp <- matrix(NA,nrow=3,ncol=12)
distTmp[1,] <- (exemplo2$x-newCx[1])^2 +
(exemplo2$y-newCy[1])^2
distTmp[2,] <- (exemplo2$x-newCx[2])^2 +
(exemplo2$y-newCy[2])^2
distTmp[3,] <- (exemplo2$x-newCx[3])^2 +
(exemplo2$y-newCy[3])^2
newClust2 <- apply(distTmp,2,which.min)
points(exemplo2$x,exemplo2$y,pch=19,cex=2,c
ol=cols1[newClust2])
```

47

Etapa 4/2

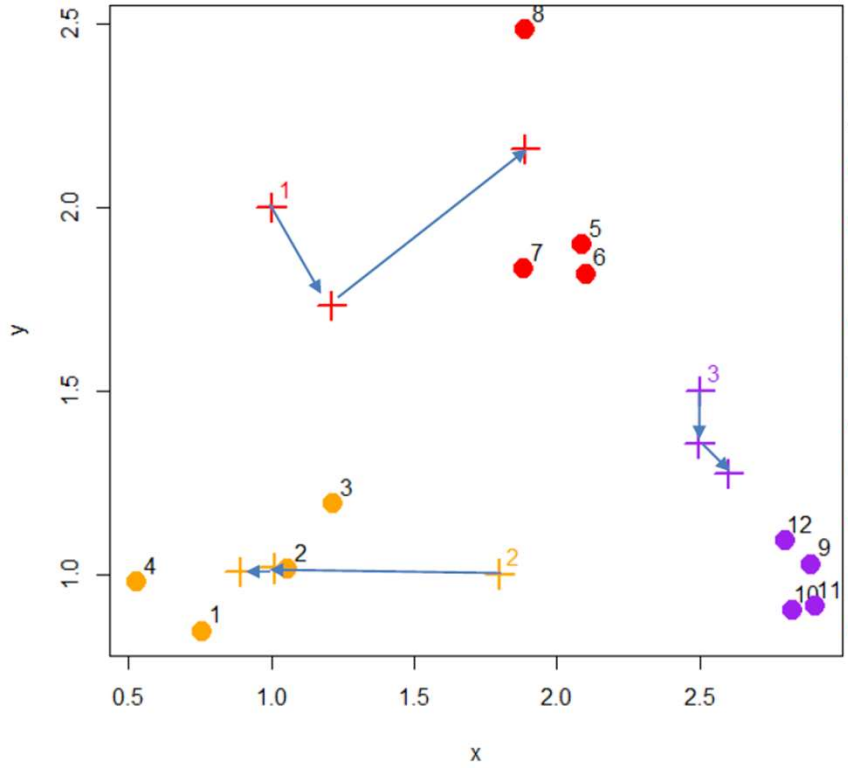


Etapa 4(2) – Recalcula os Centróides

```
newCx2 <- tapply(exemplo2$x,newClust2,mean)
newCy2 <- tapply(exemplo2$y,newClust2,mean)
points(newCx2,newCy2,col=cols1,pch=3,cex=2,lwd
=2)
```

48

Etapa 3/3

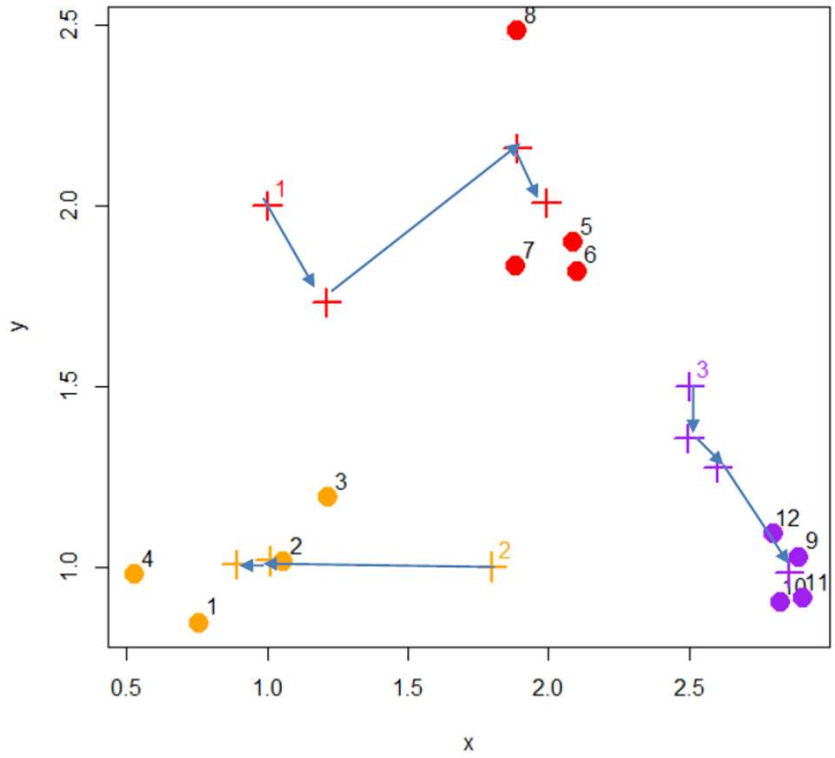


Etapa 3(3) – Designa os pontos aos Centroides mais Próximos

```
distTmp <- matrix(NA,nrow=3,ncol=12)
distTmp[1,] <- (exemplo2$x-newCx2[1])^2 +
(exemplo2$y-newCy2[1])^2
distTmp[2,] <- (exemplo2$x-newCx2[2])^2 +
(exemplo2$y-newCy2[2])^2
distTmp[3,] <- (exemplo2$x-newCx2[3])^2 +
(exemplo2$y-newCy2[3])^2
finalClust <- apply(distTmp,2,which.min)
points(exemplo2$x,exemplo2$y,pch=19,cex=2,col
cols1[finalClust])
```

49

Etapa 4/3



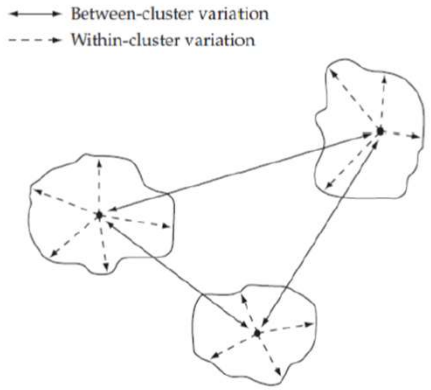
Etapa 4(3) – Recalcula os Centróides

```
finalCx <- tapply(exemplo2$x,finalClust,mean)
finalCy <- tapply(exemplo2$y,finalClust,mean)
points(finalCx,finalCy,col=cols1,pch=3,cex=2,lwd=2)
```

50

K-Means Cluster

- A ideia basica do algoritmo k-Means é definir os clusters de maneira a minimizar a variação dentro do cluster e ao mesmo tempo maximizar a variação entre os clusters



$$\text{Cluster SSE} = \sum_{\mathbf{x} \in C_i} dist(\mathbf{c}_i, \mathbf{x})^2$$

$$\text{SSE} = \sum_{i=1}^{K} \sum_{x \in C_i} dist(c_i, x)^2$$

$$\text{Total SSB} = \sum_{i=1}^{K} m_i \, dist(\mathbf{c}_i, \mathbf{c})^2$$

$$\mathbf{c}_i = \frac{1}{m_i} \sum_{\mathbf{x} \in C_i} \mathbf{x} \qquad \text{TSS} = \sum_{i=1}^{K} \sum_{x \in C_i} (x - c)^2 = \text{SSE} + \text{SSB}$$

Fonte: Tan et al., 2006

51

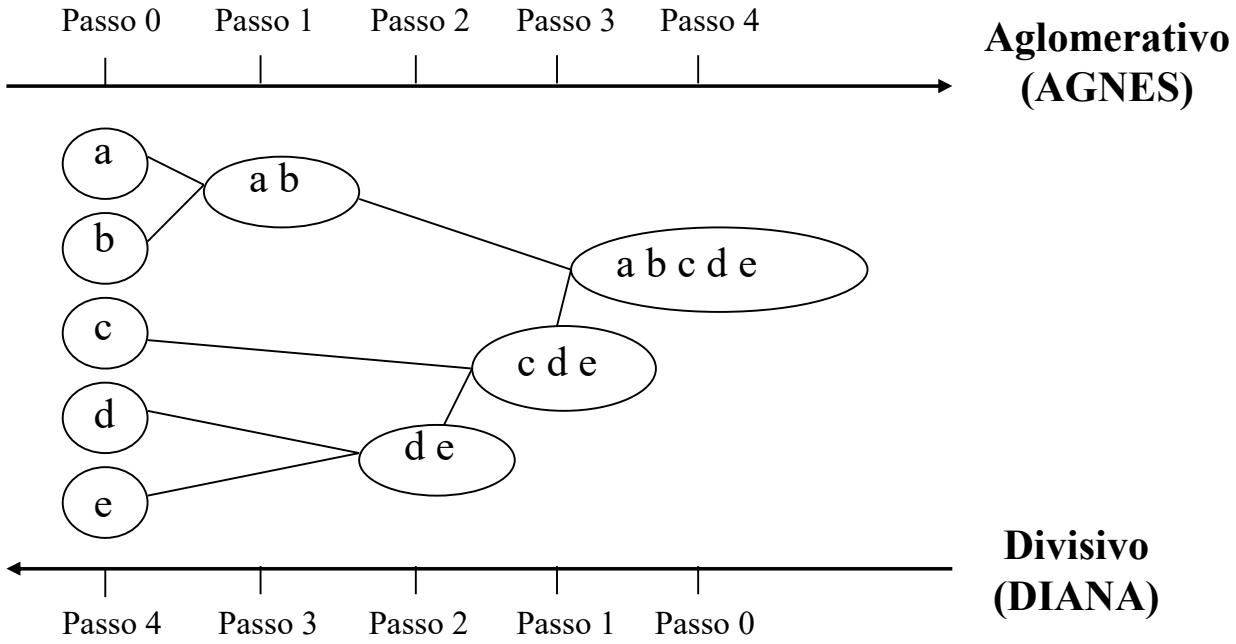
Comentários sobre K-Means

- Pontos Fortes
  - Relativamente eficiente: O (tkn), onde n é o número de objetos, k o número de clusters, e t o número de iterações. Normalmente, k e t são muito menores que n.
  - Normalmente finaliza otimizado: a otimização pode ser conseguida usando técnicas como simulação e algoritmo genético.
- Pontos Fracos
  - Aplicável apenas quando uma média pode ser definida; não aplicável a dados categóricos.
  - É necessário especificar **k**, o número de clusters, com antecedência.
  - Dificuldade com dados com ruídos e *outliers*.
  - Não é adequado para descobrir clusters em formas não convexas.

52

Cluster Hierárquico

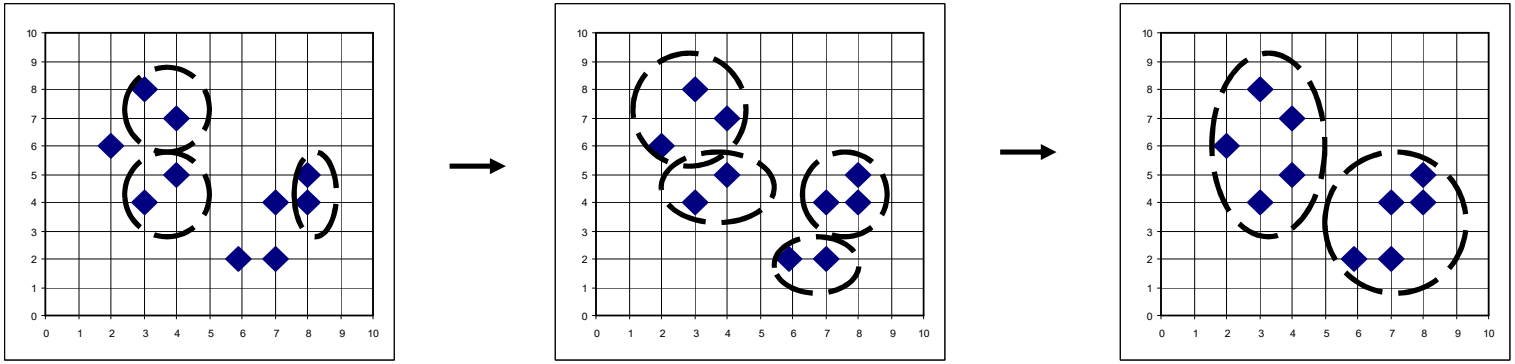
- Usa uma matriz de distância como critério para o agrupamento.
- Não requer o número de clusters k como uma entrada, mas precisa de uma condição para encerramento.



53

AGNES (Aninhamento Aglomerativo)

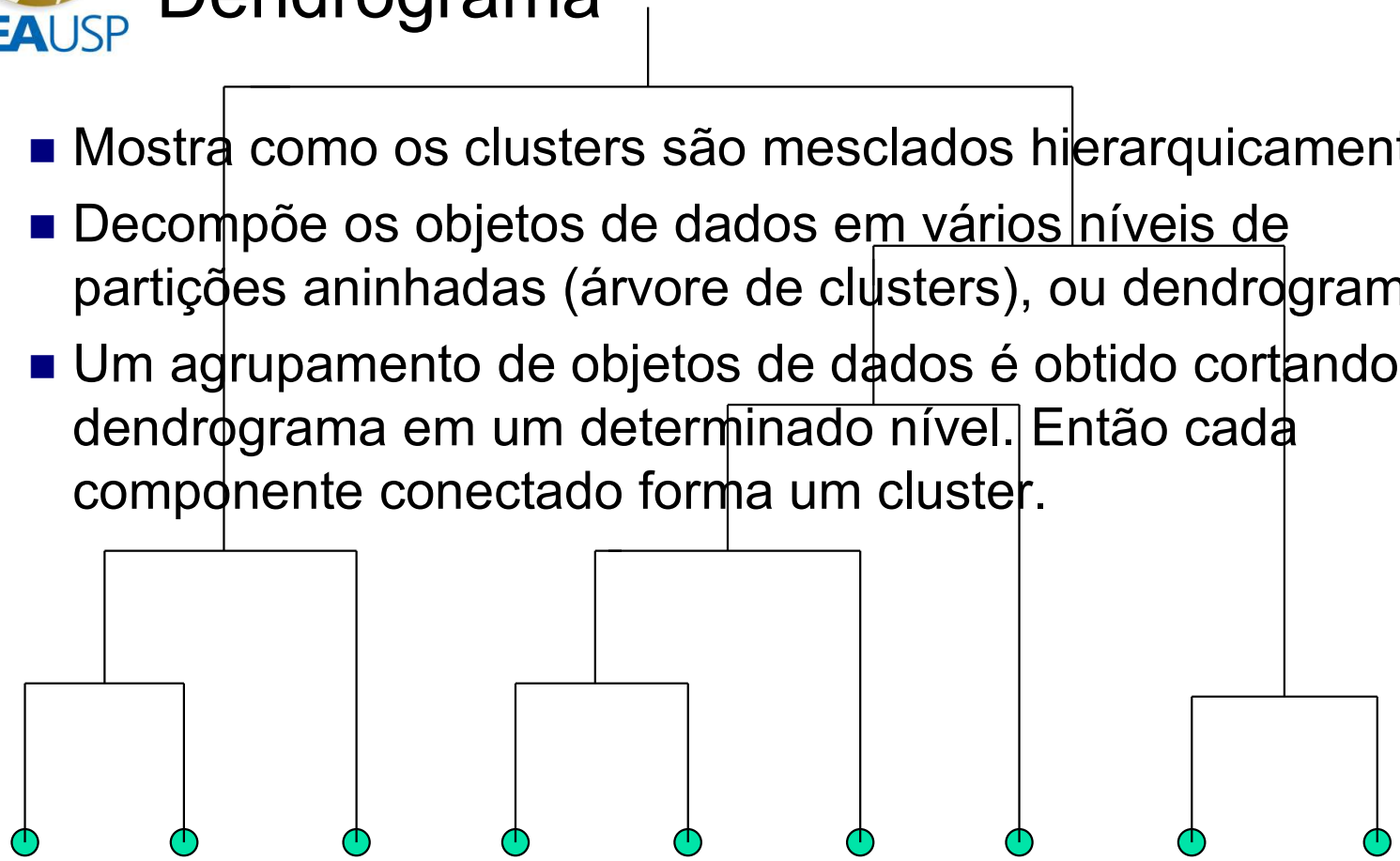
- Produz uma árvore de clusters (nós)
- Inicialmente cada objeto é um cluster (folha)
- Recursivamente mescla nós que possuem a menor dissimilaridade.
- Critério: menor distância, máxima distância, distância média, distância central.
- Eventualmente todos os nós pertencem ao mesmo cluster (raiz).



54

Dendrograma

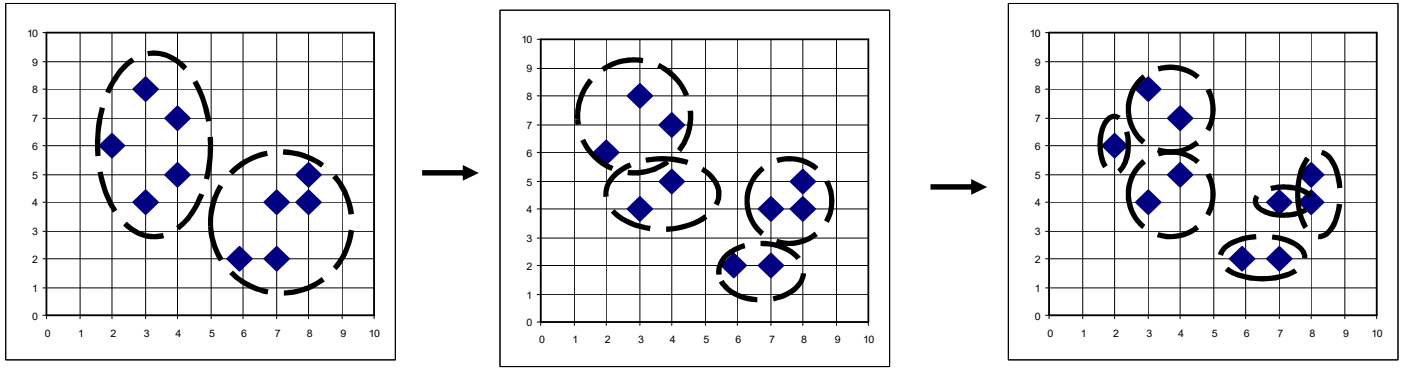
- Mostra como os clusters são mesclados hierarquicamente.
- Decompõe os objetos de dados em vários níveis de partições aninhadas (árvore de clusters), ou dendrograma.
- Um agrupamento de objetos de dados é obtido cortando o dendrograma em um determinado nível. Então cada componente conectado forma um cluster.



55

DIANA (Análise Divisiva)

- Ordem inversa do Aglomerativo (AGNES).
- Inicia com o cluster raiz contendo todos os objetos.
- Recursivamente divide em subclusters.
- Eventualmente cada cluster contém um único objeto.



56

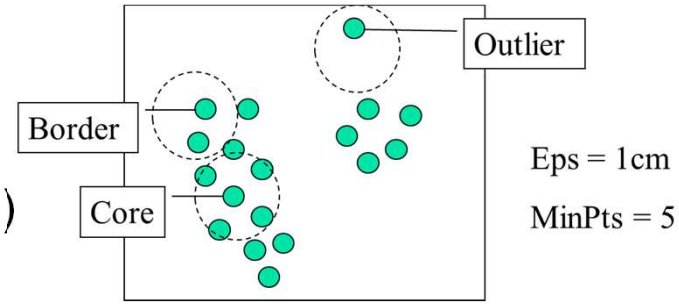
Outros Métodos Hierárquicos

- Fraquezas do método aglomerativo de agrupamento:
  - Não escalam bem: complexidade de tempo.
  - Não se pode desfazer o que foi feito previamente.
- Integração de métodos hierárquicos com métodos baseados em distância:
  - **BIRCH**: usa CF-tree (Cluster Feature Tree) e incrementalmente ajusta a qualidade dos subclusters.
  - **CURE**: seleciona objetos bem dispersos a partir do conjunto e, em seguida, os contrai em direção ao centro do conjunto por uma fração especificada.

57

Agrupamento baseado em Densidade

- Cluster baseado na densidade (critério de cluster local), tais como pontos conectados à densidade
- Destaques:
  - Descobre agrupamentos de formato arbitrário
  - Trata ruído
  - Uma única varredura
  - Necessita parâmetros de densidade como condição de finalização
- Algoritmos representativos:
  - DBSCAN (éster et al., 1996)
  - DENCLUE (Hinneburg & Keim, 1998)



58

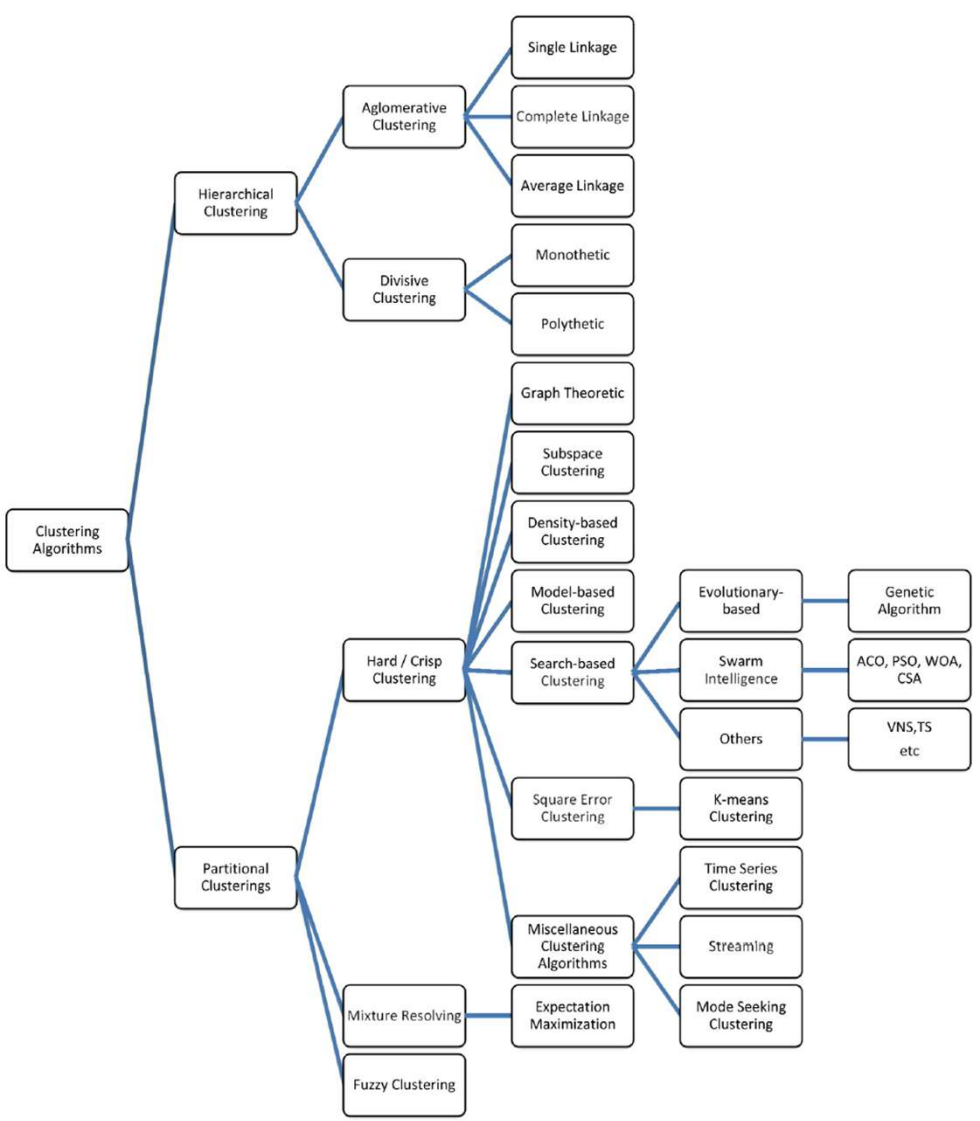


Fig. 1. Taxonomy of clustering algorithms.

59