

Artificial Intelligence/ Inteligência Artificial

Lecture 11: Unsupervised Learning/ Aprendizagem Não Supervisionada (adaptado de Faria, 2018 e Castillo 2011)

Luís Paulo Reis

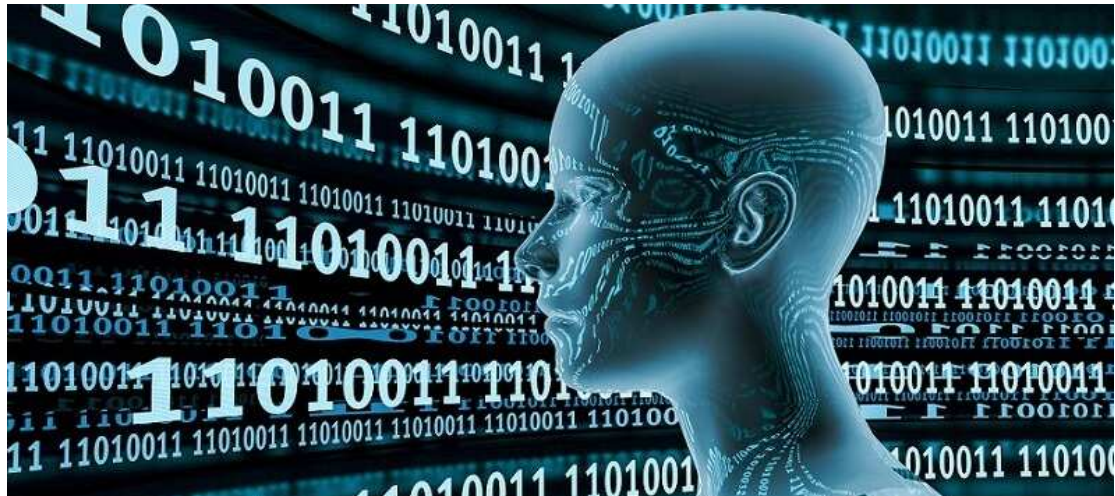
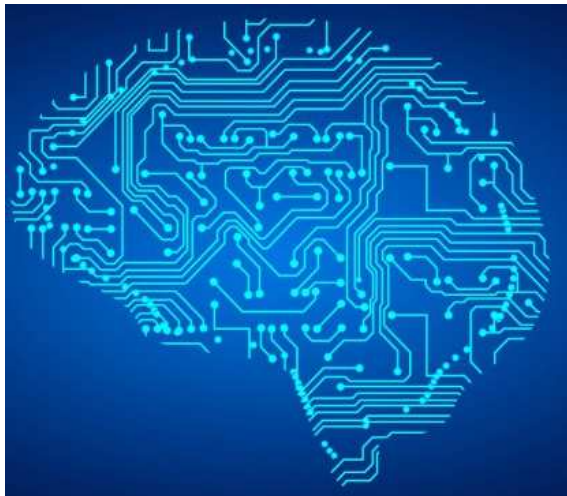
lpreis@fe.up.pt

Director of LIACC – Artificial Intelligence and Computer Science Lab.
Associate Professor at DEI/FEUP – Informatics Engineering Department,
Faculty of Engineering of the University of Porto, Portugal
President of APPIA – Portuguese Association for Artificial Intelligence



Machine Learning

- Machine learning is a field of artificial intelligence that gives computer systems the **ability to "learn"** (e.g., progressively **improve performance** on a specific task) from data/results of their actions, without being explicitly programmed

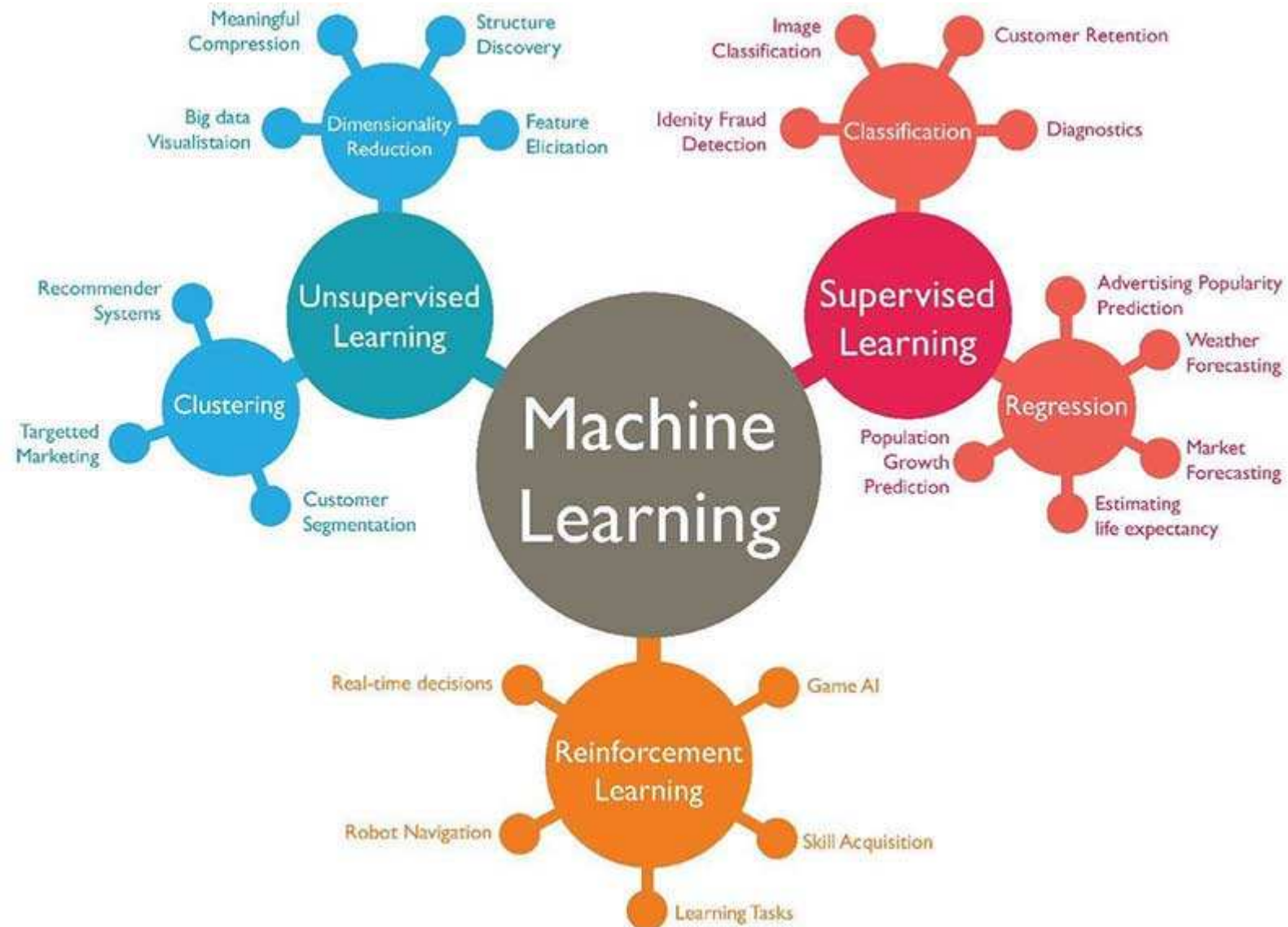


Machine Learning Tasks/Types

Machine Learning (ML) Tasks:

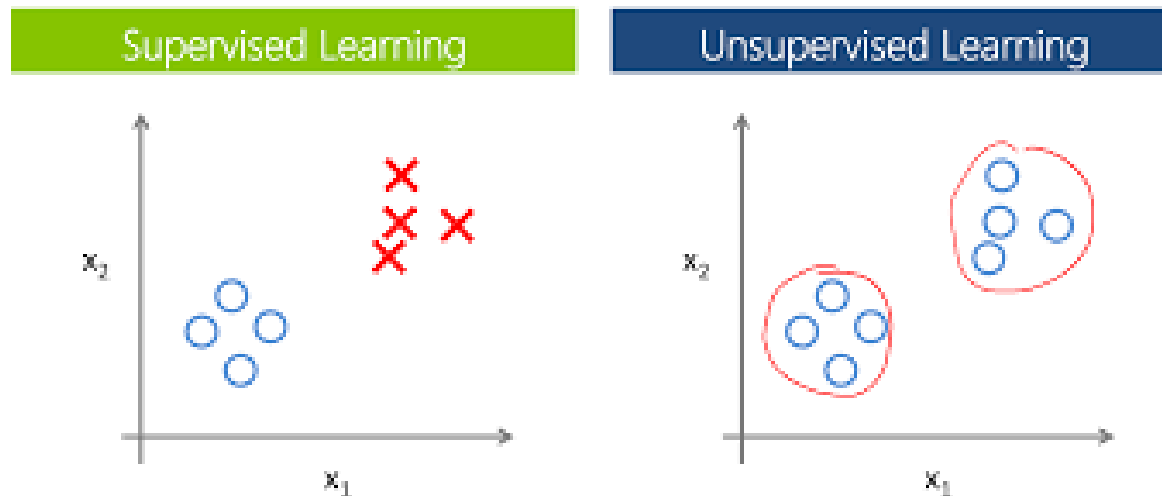
- **Supervised learning:** Example inputs and desired outputs are available/given by a "teacher", and the goal is to learn how to map inputs to outputs (possibility semi-supervised)
- **Reinforcement learning:** Data (in form of rewards and punishments) are given only as feedback to the computer/agent actions in a dynamic environment
- **Unsupervised learning:** No labels/outputs are given to the learning algorithm, leaving it on its own to find structure in its input

Machine Learning



Métodos de Aprendizagem Não Supervisionada

- Modelos Descritivos
- Descrever a informação
- Encontrar padrões nos dados
- Efetuada com base em observação e descoberta



Regras de Associação

- Dado um conjunto de transações, encontrar regras que prevejam a ocorrência de um item baseado nas ocorrências de outros itens na transação

Market-Basket transactions

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Exemplo de Regras de Associação

$\{\text{Diaper}\} \rightarrow \{\text{Beer}\},$
 $\{\text{Milk, Bread}\} \rightarrow \{\text{Eggs, Coke}\},$
 $\{\text{Beer, Bread}\} \rightarrow \{\text{Milk}\},$

Implicação significa co-ocorrência,
não causalidade!

Regras de Associação

- Uma expressão de implicação da forma $X \rightarrow Y$ onde X e Y são conjunto de itens
- **conjunto de itens** (*itemset*): coleção de um ou mais itens
 k -*itemset* = *itemset* com k itens
- **contagem de suporte** (σ): frequência de ocorrência de um conjunto de itens
- **suporte** (s): fração das transações que contêm X e Y
- **confiança** (c): mede quão frequentemente itens em Y aparecem em transações que contém X
- **conjunto de itens frequentes**: se suporte é maior ou igual a um limite *minsup*

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

$\{\text{Milk, Diaper}\} \Rightarrow \text{Beer}$

$$s = \frac{\sigma(\text{Milk, Diaper, Beer})}{|T|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk, Diaper, Beer})}{\sigma(\text{Milk, Diaper})} = \frac{2}{3} = 0.67$$

Regras de Associação

- Exemplos

<i>TID</i>	<i>Items</i>
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

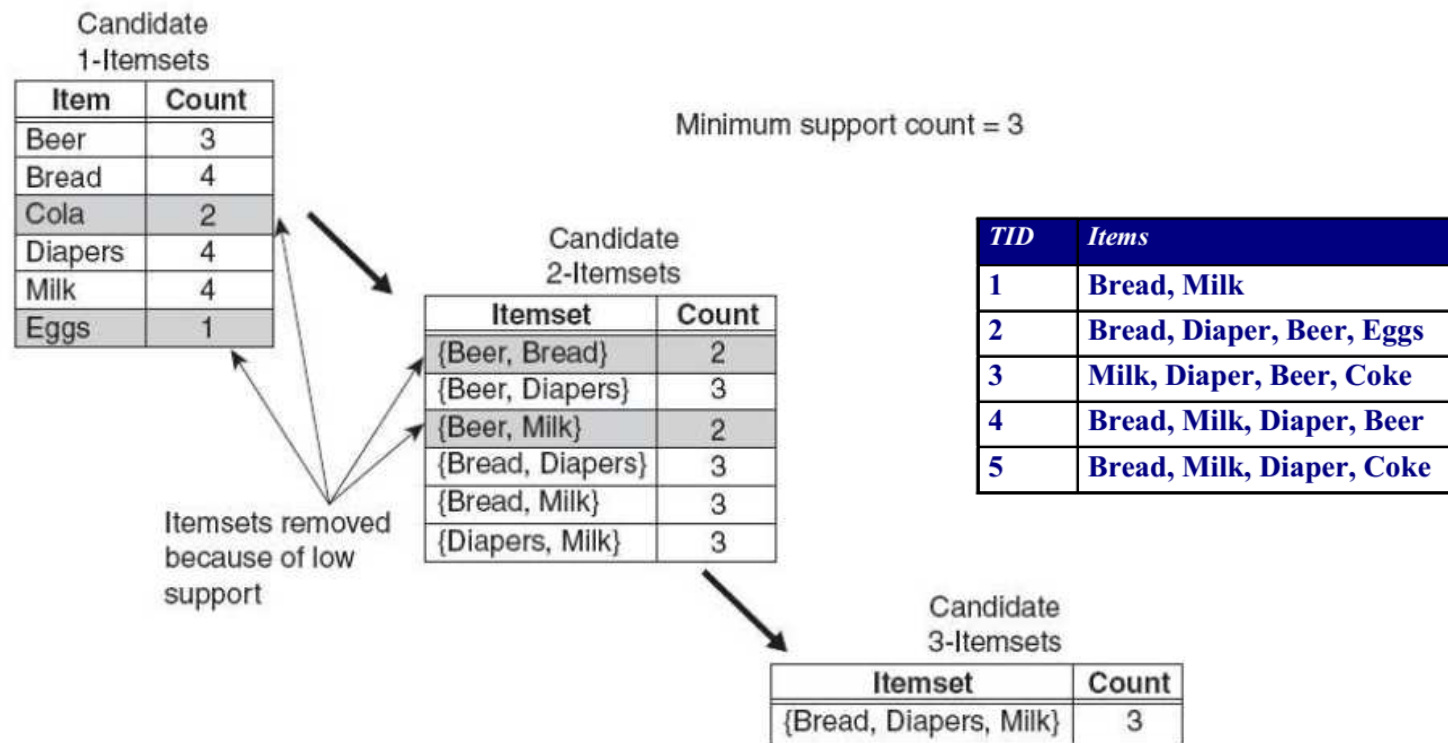
$\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\}$ (s=0.4, c=0.67)
 $\{\text{Milk, Beer}\} \rightarrow \{\text{Diaper}\}$ (s=0.4, c=1.0)
 $\{\text{Diaper, Beer}\} \rightarrow \{\text{Milk}\}$ (s=0.4, c=0.67)
 $\{\text{Beer}\} \rightarrow \{\text{Milk, Diaper}\}$ (s=0.4, c=0.67)
 $\{\text{Diaper}\} \rightarrow \{\text{Milk, Beer}\}$ (s=0.4, c=0.5)
 $\{\text{Milk}\} \rightarrow \{\text{Diaper, Beer}\}$ (s=0.4, c=0.5)

Regras de Associação

- Aprendizagem de Regras de Associação
 - Dado um conjunto de transações T , o objetivo é encontrar todas as regras que tenham:
 - suporte $\geq \text{minsup}$ (threshold)
 - confiança $\geq \text{minconf}$ (threshold)
 - Abordagem em dois passos:
 - **Geração de Conjuntos de Itens Frequentes** com suporte minsup
 - **Gerar regras** com alta confiança a partir de cada conjunto de itens frequente
 - Algoritmo mais popular: algoritmo **Apriori** (Agrawal & Srikant, 1994)
 - **Princípio Apriori**: se um conjunto de itens é frequente, então todos os seus subconjuntos também devem ser frequentes

Regras de Associação

- Geração do Conjunto de Itens Frequentes usando o Algoritmo Apriori



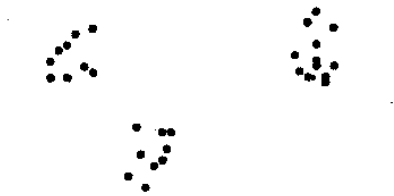
Análise de Clusters

Clustering

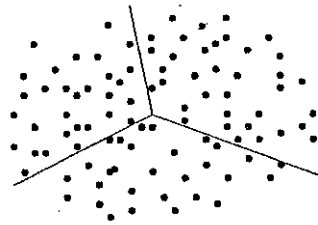
- Processo de classificação
- Divisão do conjunto inicial de dados em vários subconjuntos de dados ou o conjunto inicial de variáveis em vários subconjuntos de variáveis
- Meio informal de avaliar a dimensionalidade dos dados
- Identificar outliers nas observações
- Sugerir hipóteses interessantes sobre associação entre variáveis

Análise de Clusters

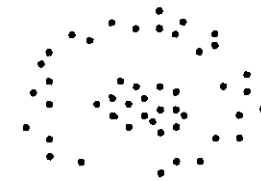
- Formas de Clusters (ou grupos)



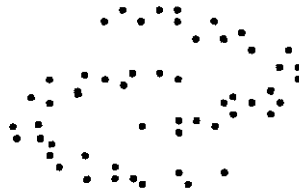
Grupos Coesos e bem separados



Grupo homogéneo sem clusters naturais



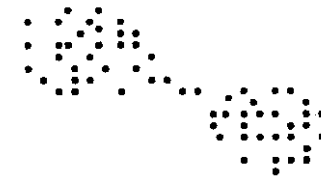
Grupo separados mas não coesos



Grupos separados mas sem coesão interna



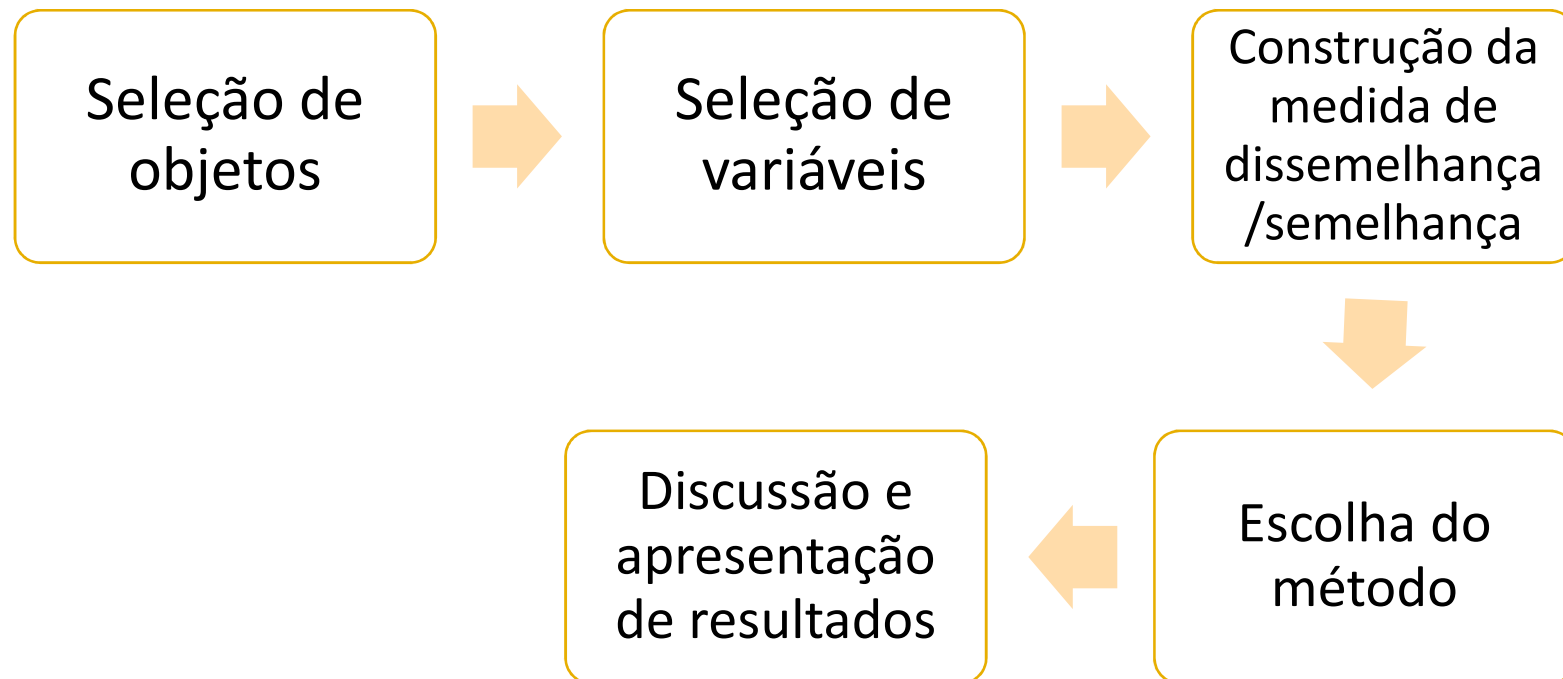
Zonas de grande densidade rodeadas por regiões de pequena densidade



Grupos totalmente coesos mas não separados

Análise de Clusters

Fases de uma análise de clusters



Análise de Clusters

- Medidas de Semelhança e Dissemelhança
- Sujeitos ou itens
 - Agrupados segundo tipos de distância métrica
- Variáveis
 - Agrupadas através de medidas de correlação ou associação

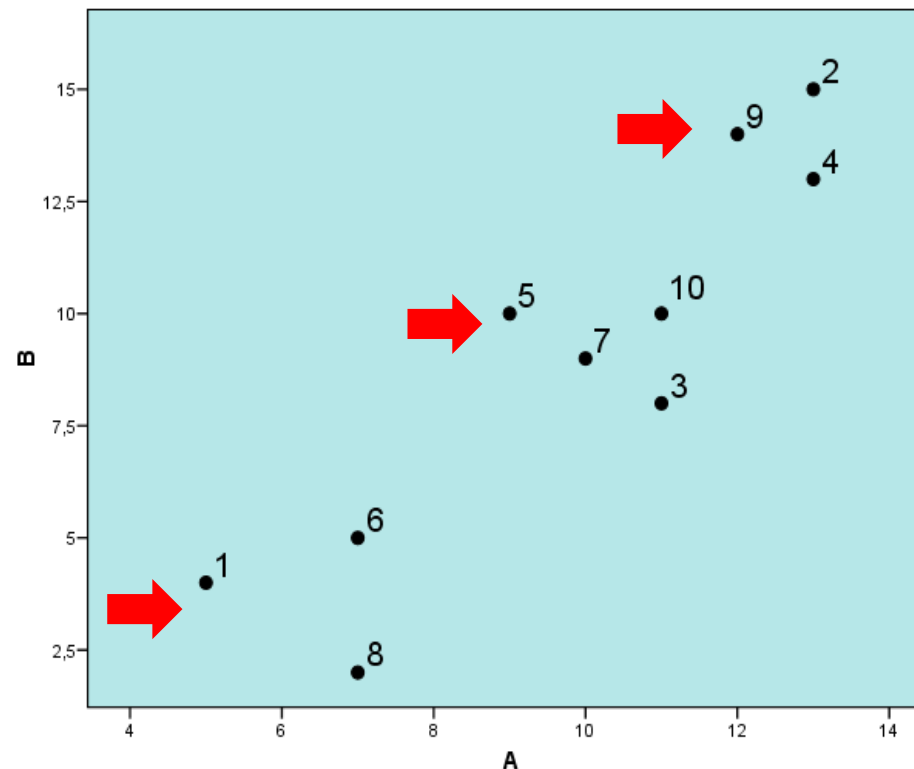
Análise de Clusters

- Exemplo
 - Identificar grupos de indivíduos para os quais possa ser recomendado um acompanhamento médico específico

	sujeito	A	B	C	D	var	
1	1	5	4	8	6		
2	2	13	15	8	6		
3	3	11	8	10	10		
4	4	13	13	16	9		
5	5	9	10	9	6		
6	6	7	5	10	1		
7	7	10	9	9	8		
8	8	7	2	6	4		
9	9	12	14	14	4		
10	10	11	10	9	8		
11							

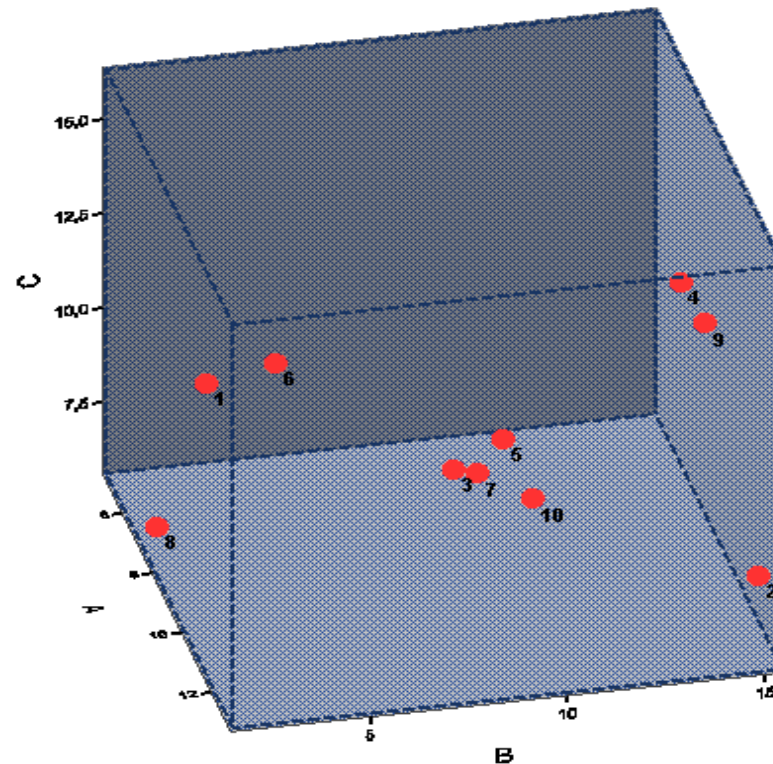
Análise de Clusters

- Diagrama de Dispersão do Exemplo
 - Recorrendo às variáveis A e B



Análise de Clusters

- Diagrama de Dispersão do Exemplo
 - Recorrendo às variáveis A, B e C



Análise de Clusters

- Diagrama de Dispersão do Exemplo
 - Para mais do que 3 variáveis não é possível visualizar
- Recorrer a medidas de semelhança (ou proximidade) e/ou medidas de dissemelhança (ou distância) entre sujeitos

Análise de Clusters

- Distância Euclidiana
- Distância Minkowski
- Distância de Mahalanobis
- Medida de Semelhança do Co-seno
- Coeficiente de Jaccard, de Russel & Rão e Medidas de Associação Binária
- Medidas de Semelhança para Variáveis

Análise de Clusters

- Matriz de Dissemelhança do Exemplo

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1	0,0	13,6	7,2	12,0	7,2	2,2	7,1	2,8	12,2	8,5
S2	13,6	0,0	7,3	2,0	6,4	11,7	6,7	14,3	1,4	5,4
S3	7,2	7,3	0,0	5,4	2,8	5,0	1,4	7,2	6,1	2,0
S4	12,0	2,0	5,4	0,0	5,0	10,0	5,0	12,5	1,4	3,6
S5	7,2	6,4	2,8	5,0	0,0	5,4	1,4	8,2	5,0	2,0
S6	2,2	11,7	5,0	10,0	5,4	0,0	5,0	3,0	10,3	6,4
S7	7,1	6,7	1,4	5,0	1,4	5,0	0,0	7,6	5,4	1,4
S8	2,8	14,3	7,2	12,5	8,2	3,0	7,6	0,0	13,0	8,9
S9	12,2	1,4	6,1	1,4	5,0	10,3	5,4	13,0	0,0	4,1
S10	8,5	5,4	2,0	3,6	2,0	6,4	1,4	8,9	4,1	0,0

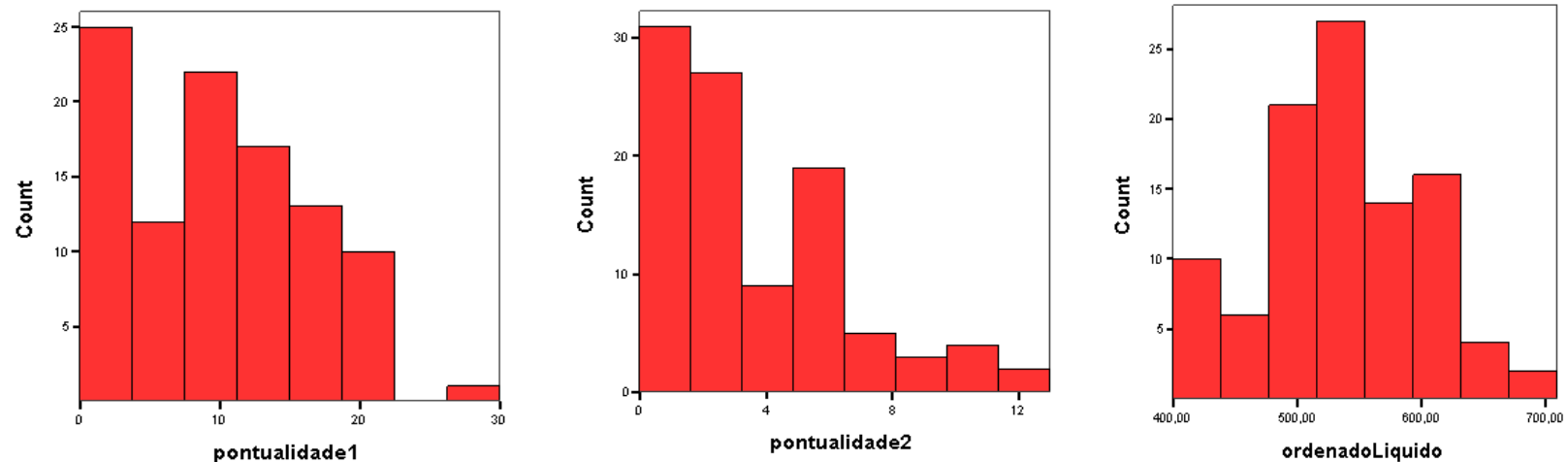
- Quanto menor a distância euclidiana menor é a dissemelhança (ou maior é a semelhança ou proximidade) entre indivíduos

Análise de Clusters

- Agrupar os sujeitos em clusters homogêneos
 - a partir das medidas de dissemelhança
 - de modo que dentro do mesmo cluster essas medidas sejam as menores possíveis
 - e entre clusters as maiores possíveis

Análise de Clusters

Histograma – 1 variável



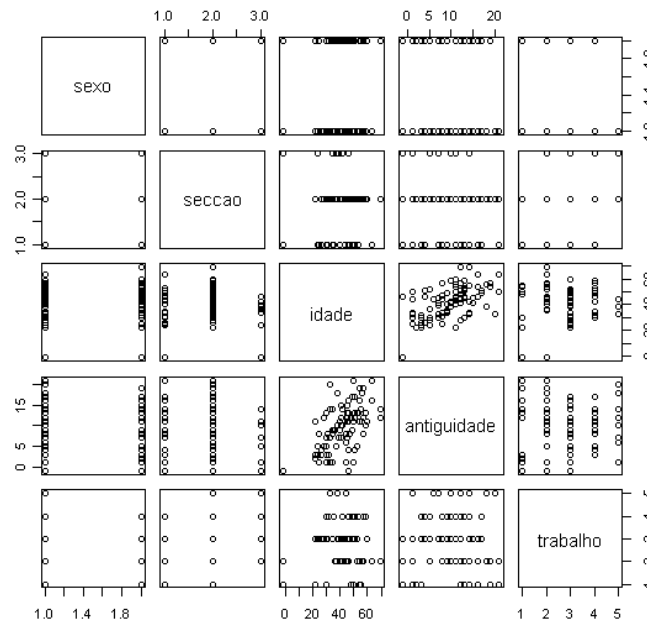
A existência de várias modas é, em geral, reveladora da existência de clusters

Existência de outros métodos para a representação gráfica, por ex, gráficos de barras, circulares e gráficos de caule-e-folhas

Análise de Clusters

Diagrama de Dispersão – 2 variáveis

Matriz de diagramas de dispersão para algumas variáveis

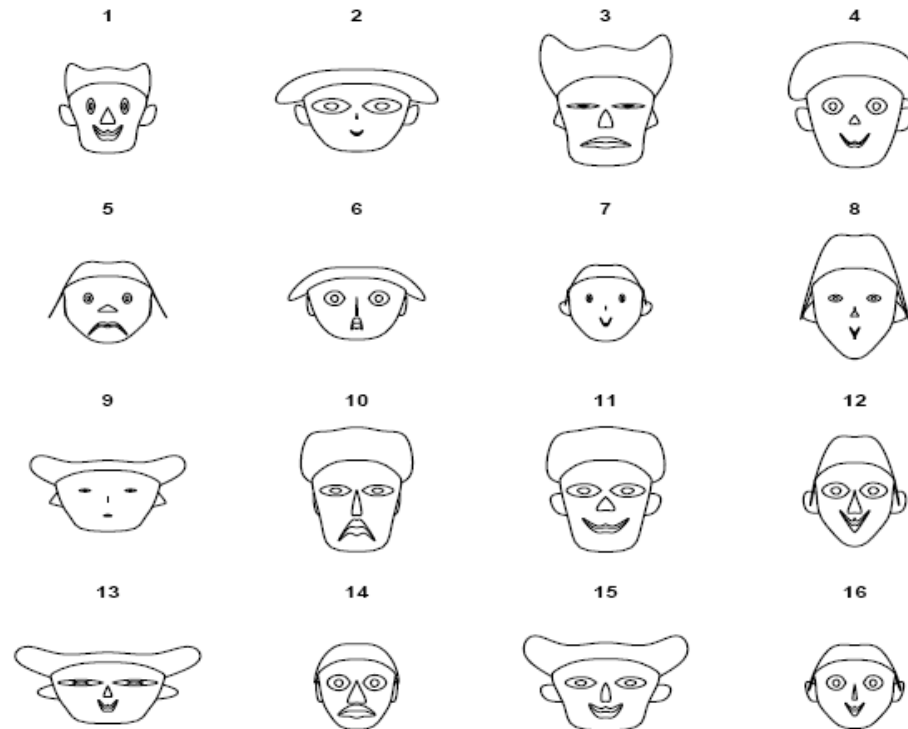


Consideração de todos os pares de variáveis numa tentativa de análise global
Tarefa complicada e confusa se o número de variáveis for elevado

Análise de Clusters

Caras de Chernoff

A cada variável é associada um aspecto particular da face de uma pessoa



Análise de Clusters

Métodos Hierárquicos - Formam uma hierarquia caracterizada pelo facto de dados dois grupos ou são disjuntos ou um deles está contido no outro

- **Método 1 - Aglomerativos**
 - Recorrem a passos sucessivos de agregação dos sujeitos considerados individualmente (cada sujeito é um cluster)
 - Em seguida vão sendo agrupados de acordo com as suas proximidades
- **Método 2 - Divisivos**
 - Todos os sujeitos são à partida agrupados num único Cluster
 - Depois são divididos em subgrupos de acordo com as suas medidas de distância

Análise de Clusters

Métodos Hierárquicos - Método 1 - Aglomerativos

- É necessário encontrar um modo de definir as distâncias entre o cluster com mais de um indivíduo (ou variável) e os restantes
- Menor distância (single linkage ou nearest neighbor) – após a formação do primeiro cluster, a distância deste aos restantes é a menor das distâncias de cada um dos elementos constituintes deste cluster a cada um dos restantes indivíduos (ou variáveis)

Análise de Clusters

Métodos Hierárquicos - Método 1 - Aglomerativos

- Maior distância (complete linkage ou farthest neighbor) – após a formação do primeiro cluster, a distância deste aos restantes é a maior das distâncias de cada um dos elementos constituintes deste cluster a cada um dos restantes indivíduos (ou variáveis)
- Distância média entre clusters (average linkage between groups) – após a formação do primeiro cluster, a distância deste aos restantes é a média das distâncias de cada um dos elementos constituintes deste cluster a cada um dos restantes indivíduos (ou variáveis)

Análise de Clusters

Métodos Hierárquicos - Método 1 - Aglomerativos

- Distância média dentro dos clusters (average linkage within groups) – semelhante à distância média entre clusters, mas com variabilidade dentro os clusters a menor possível
- Distância mediana (median linkage) - após a formação do primeiro cluster, a distância deste aos restantes é a mediana das distâncias de cada um dos elementos constituintes deste cluster a cada um dos restantes indivíduos (ou variáveis)

Análise de Clusters

Métodos Hierárquicos - Método 1 - Aglomerativos

- Método do centróide – o novo cluster formado é representado por um ponto cujas coordenadas são a média dos indivíduos que fazem parte do cluster para cada uma das variáveis (ou seja, pelo centróide)
- Método de Ward – não são calculadas distâncias e os clusters são formados de forma a minimizar a soma dos quadrados dos erros

Análise de Clusters

Métodos Hierárquicos - Método 1 – Aglomerativos

Tipo de método hierárquico a utilizar?

Por default o método da menor distância (Single linkage)

- Implementado em vários softwares por omissão
- Tende a maximizar a conectividade entre clusters
- Tendência para criar um menor número de clusters do que o método da máxima distância (Complete linkage)

Análise de Clusters

Métodos Hierárquicos - Método 1 – Aglomerativos

Tipo de método hierárquico a utilizar?

- Método da máxima distância
 - Tendência para minimizar a distância entre clusters em cada passo
 - Tendência para produzir clusters compactos
 - Outros métodos tendem a apresentar características intermédias entre os dois métodos anteriores
- Não existe um melhor processo de agregação hierárquica é aconselhável a utilização de vários métodos em simultâneo

Análise de Clusters

Métodos Hierárquicos - Verificação e validação dos clusters

Construção de um diagrama de árvore hierárquica: **Dendrograma**

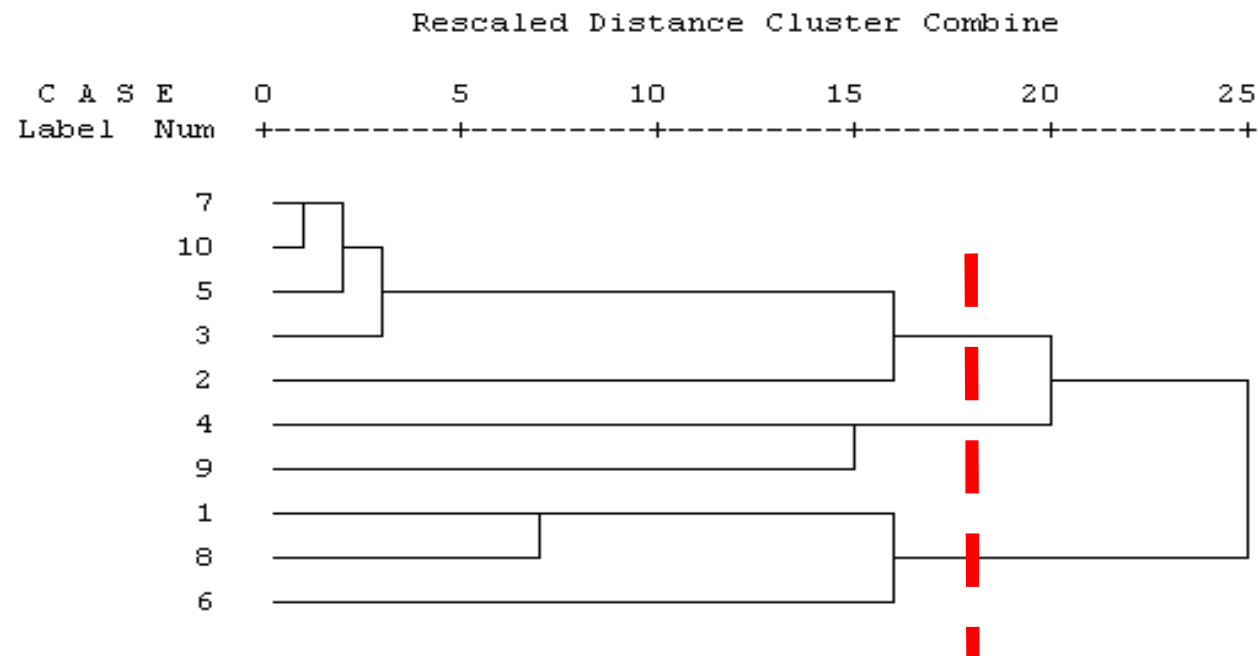
- Contém parêntesis ligando dados e mostra a ordem pela qual os pontos estão assinalados para os grupos
- Os comprimentos das suas ligações são proporcionais às distâncias entre os pontos e grupos

Análise de Clusters

- Métodos Hierárquicos - Verificação e validação dos clusters

* * * * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S * * * * *

Dendrogram using Single Linkage



As distâncias
(coeficientes)
foram
reescaladas

Análise de Clusters

- Métodos Hierárquicos - Verificação e validação dos clusters
- Quantos clusters se deve reter?
 - Através da análise do dendrograma:
 - 2 clusters (1, 8, 6) e (9, 4, 2, 3, 5, 10, 7)
 - Mais natural a divisão em 3 clusters, uma vez que o grupo formado por (4, 9) poderá ser separado

Análise de Clusters

- Métodos Hierárquicos - Verificação e validação dos clusters
 - Métodos heurísticos para avaliar a solução de clusters e o número de clusters
 - Distância entre clusters
 - Se a distância entre clusters é pequena estes devem ser agregados
 - Construída à custa da tabela de semelhança

Análise de Clusters

- Métodos Hierárquicos - Verificação e validação dos clusters
 - Critério do R quadrado - É uma medida de percentagem da variabilidade total que é retida em cada uma das soluções dos clusters
 - Se o número de clusters é um a variabilidade entre clusters é zero
 - Se o número de clusters é igual ao número de sujeitos, a variabilidade entre clusters é um que é a variabilidade total

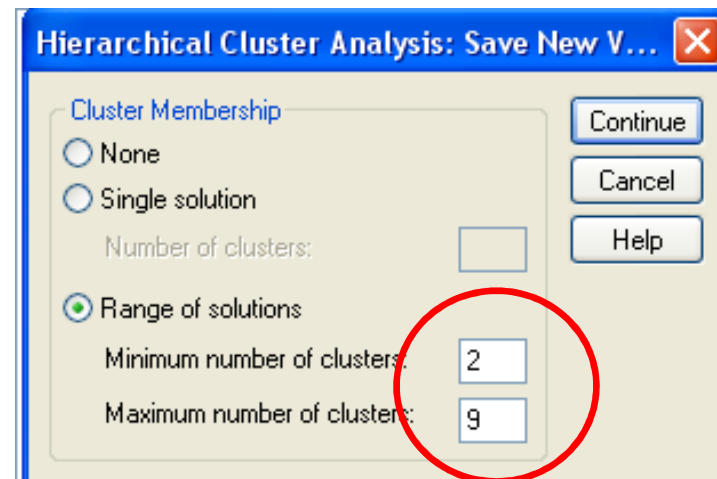
Análise de Clusters

- Métodos Hierárquicos - Verificação e validação dos clusters
- Objetivo: Encontrar o número mínimo de clusters que retenha uma percentagem significativa de variabilidade total (por exemplo superior a 80%)

$$R - squared = \frac{SQC}{SQT} = \frac{\sum_{i=1}^p \sum_{j=1}^k n_{ij} (\bar{X}_{ij} - \bar{X}_i)^2}{\sum_{i=1}^p \sum_{j=1}^k \sum_{l=1}^{n_i} (X_{ijl} - \bar{X})^2}$$

Análise de Clusters

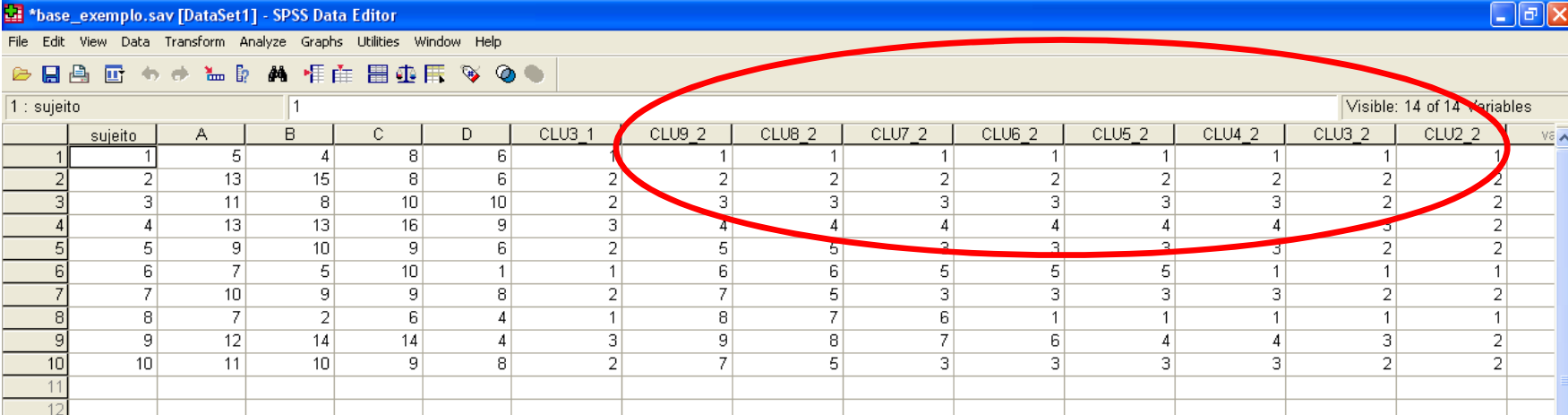
- Critério do R quadrado
- Os cálculos podem ser realizados com o auxílio da Anova one-way do SPSS



- As novas variáveis registam a presença de cada indivíduo aos diferentes clusters

Análise de Clusters

- Critério do R quadrado



*base_exemplo.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

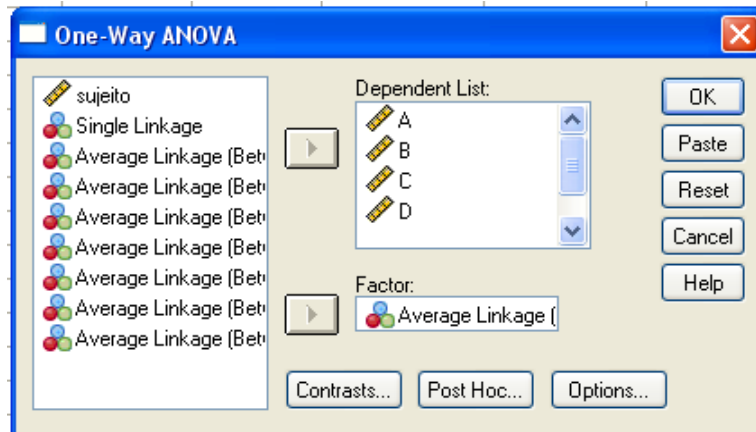
1 : sujeito 1 Visible: 14 of 14 Variables

	sujeito	A	B	C	D	CLU3_1	CLU9_2	CLU8_2	CLU7_2	CLU6_2	CLU5_2	CLU4_2	CLU3_2	CLU2_2	VS
1	1	5	4	8	6	1	1	1	1	1	1	1	1	1	
2	2	13	15	8	6	2	2	2	2	2	2	2	2	2	
3	3	11	8	10	10	2	3	3	3	3	3	3	2	2	
4	4	13	13	16	9	3	4	4	4	4	4	4	3	2	
5	5	9	10	9	6	2	5	5	2	3	3	3	2	2	
6	6	7	5	10	1	1	6	6	5	5	5	1	1	1	
7	7	10	9	9	8	2	7	5	3	3	3	3	2	2	
8	8	7	2	6	4	1	8	7	6	1	1	1	1	1	
9	9	12	14	14	4	3	9	8	7	6	4	4	3	2	
10	10	11	10	9	8	2	7	5	3	3	3	3	2	2	
11															
12															

Análise de Clusters

- Soma de Quadrados

Para cada uma das variáveis dependentes



ANOVA						
		Sum of Squares	df	Mean Square	F	Sig.
A	Between Groups	67,100	8	8,388	16,775	,187
	Within Groups	,500	1	,500		
	Total	67,600	9			
B	Between Groups	169,500	8	21,188	42,375	,118
	Within Groups	,500	1	,500		
	Total	170,000	9			
C	Between Groups	78,900	8	9,863	.	.
	Within Groups	,000	1	,000		
	Total	78,900	9			
D	Between Groups	65,600	8	8,200	.	.
	Within Groups	,000	1	,000		
	Total	65,600	9			

Análise de Clusters

- Soma de Quadrados
 - A **Soma de Quadrados dos Clusters** para cada variável é dada por Sum of Squares Between Groups e somando estas para todas as variáveis dependentes obtém-se $SQC = 381,1$
 - De modo idêntico obtém-se a **Soma dos Quadrados Totais** para todas as variáveis $SQT = 382,1$
 - $R\text{-Square} = 381,1 / 382,1 = 0,997$

Análise de Clusters

- Critério do R quadrado

De forma análoga para as restantes variáveis

Uma solução aceitável será entre 3 e 4 clusters

Nº Clusters	R-Square
1	0
2	0,566307
3	0,752159
4	0,848862
5	0,901204
6	0,941769
7	0,962706
8	0,986042
9	0,997383

Análise de Clusters

- Métodos Hierárquicos – No RapidMiner

The screenshot displays the RapidMiner Studio interface. The main workspace shows a workflow with two processes: 'Read Excel' and 'AgglomerativeClust...'. The 'Parameters' panel on the right shows settings for the 'Root (Process)' with 'logverbosity' set to 'warning'. The 'Repository' panel on the left lists various processes, including '04_AgglomerativeHierarchicalClustering'. The 'Result History' panel at the bottom shows a 'Hierarchical Cluster Model (AgglomerativeClustering)' with a dendrogram visualization. The dendrogram shows the hierarchical clustering of data points, with a color scale on the right indicating the distance between clusters.

Análise de Clusters

- Métodos Não Hierárquicos
 - Agrupar indivíduos (não variáveis)
 - Número de clusters definido inicialmente pelo analista
 - Facilidade de aplicação em matrizes de grande dimensão
 - Não é necessário calcular e armazenar uma nova matriz de dissimilaridade em cada passo do algoritmo
 - A inclusão de um indivíduo num cluster poderá não ser definitiva

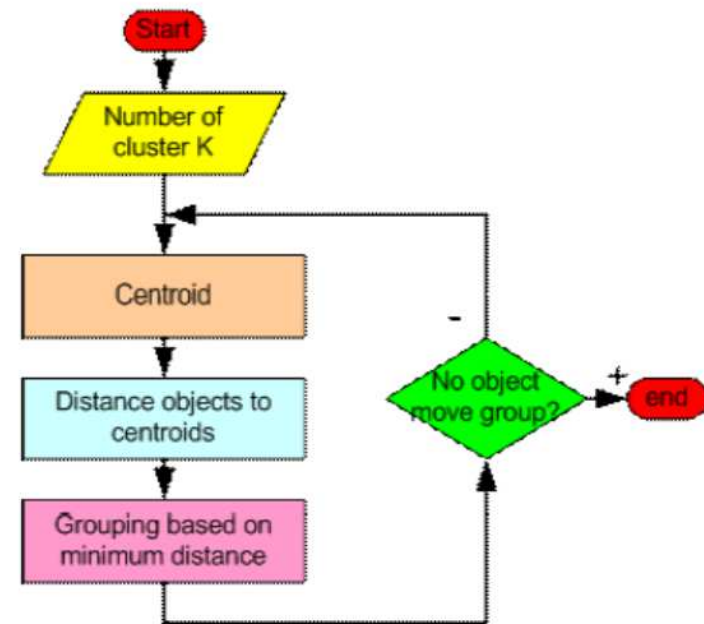
Análise de Clusters

K-Means

1º Partição inicial dos indivíduos em k clusters pré-definidos

2º Cálculo dos centróides para cada um dos k clusters e cálculo da distância euclidiana dos centróides a cada indivíduo

3º Agrupar os indivíduos aos clusters cujos centróides se encontram mais próximos e voltar ao passo 2 até que não ocorra variação significativa na distância mínima de cada sujeito da base de dados a cada um dos centróides dos k clusters (ou até que o número máximo de interações ou critério de convergência seja alcançado)



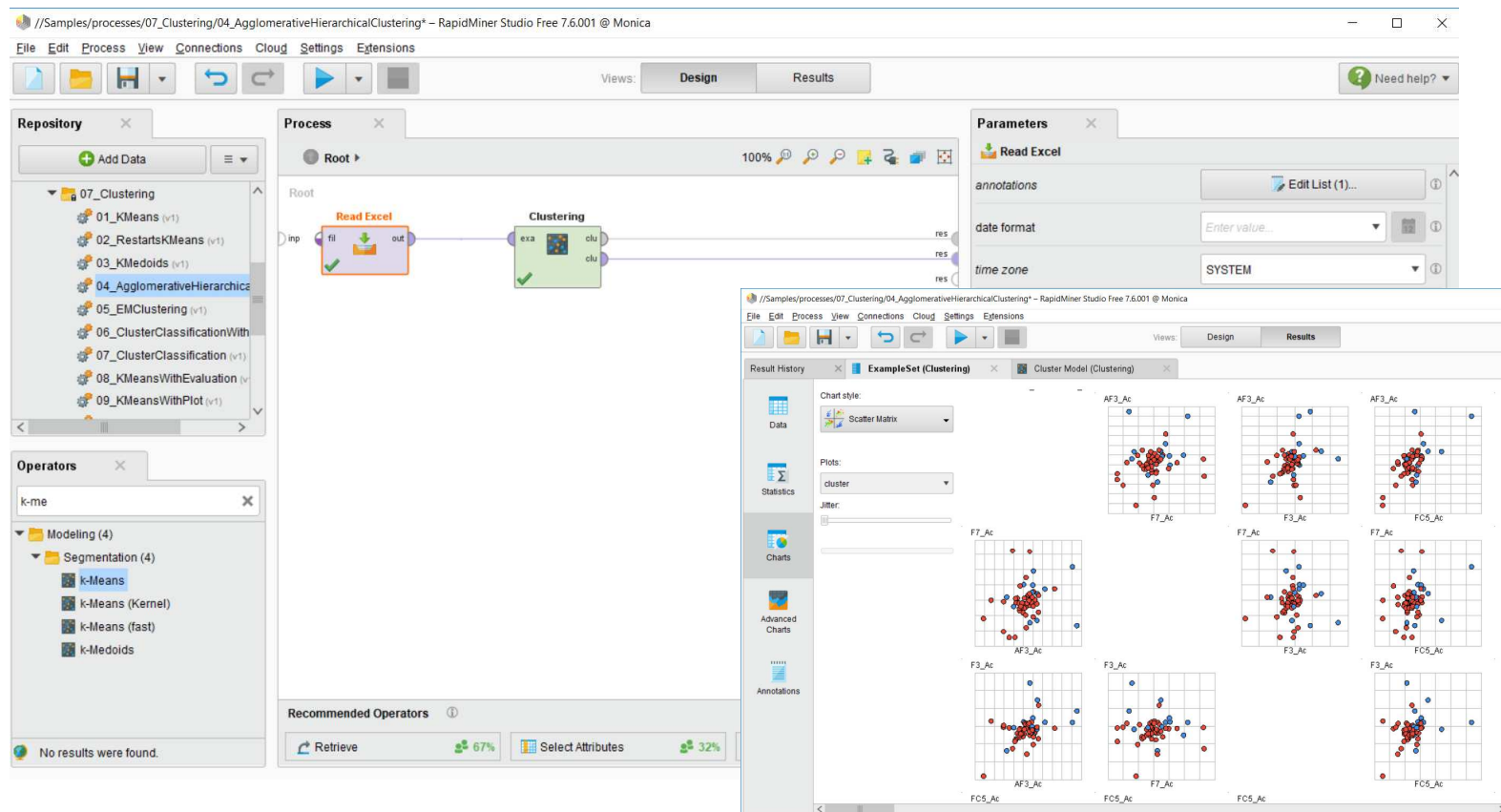
Exemplos:

<http://stanford.edu/class/ee103/visualizations/kmeans/kmeans.html>

<http://shabal.in/visuals/kmeans/4.html>

Análise de Clusters

- K-Means – No RapidMiner



Análise de Clusters

Conclusões

- A classificação dos indivíduos em cada um dos clusters é geralmente mais rigorosa nos métodos não hierárquicos
- É aconselhável iniciar a análise de clusters com métodos hierárquicos para explorar e proceder com o K-means para refinar e interpretar a solução de clusters
- A análise de clusters deve ser fundamentada com outras análises, por exemplo, a análise discriminante para obter probabilidades de erro associadas às conclusões obtidas

Bibliografia

- Tan, P., Steinbach, M. & Kumar, V. (2006). Introduction to Data Mining. Pearson Addison-Wesley.
- Adaptação de slides de “Introduction to Data Mining”, Pang-Ning Tan, Michael Steinbach, Vipin Kumar: <http://www-users.cs.umn.edu/~kumar/dmbook/index.php>
- Adaptação de slides de: Gladys Castillo, Aprendizagem Computacional (Machine Learning), Universidade de Aveiro, 2008
- Adaptação de slides de: B. Mónica Faria, Extração de Conhecimento, Politécnico do Porto, 2018
- Adaptação de slides de: <http://www-users.cs.umn.edu/~kumar/dmbook/index.php>
- Bergeron, B. (2003). Bioinformatics computing: the complete, practical guide to bioinformatics for life scientists. New Jersey: Prentice Hall.
- Santos, M. F. & Azevedo, C. (2005). Data mining: descoberta de conhecimento em bases de dados. Lisboa: FCA
- Hill M., Hill A. (2007) Investigação por Questionário, Edições Sílabo, 2ª Edição
- Maroco, J., Análise Estatística – com utilização do SPSS, Ed. Sílabo, Lda, Abril, 2003.
- Dawson-Saunders B, Trapp G (2004) Basic and Clinical Biostatistics, 4a Ed. Prentice-Hall Int. Inc
- RapidMiner: Data Science Platform, 2017, <https://rapidminer.com/>

Artificial Intelligence/ Inteligência Artificial

Lecture 11: Unsupervised Learning/ Aprendizagem Não Supervisionada (adaptado de Faria, 2018 e Castillo 2011)

Luís Paulo Reis

lpreis@fe.up.pt

Director of LIACC – Artificial Intelligence and Computer Science Lab.
Associate Professor at DEI/FEUP – Informatics Engineering Department,
Faculty of Engineering of the University of Porto, Portugal
President of APPIA – Portuguese Association for Artificial Intelligence

