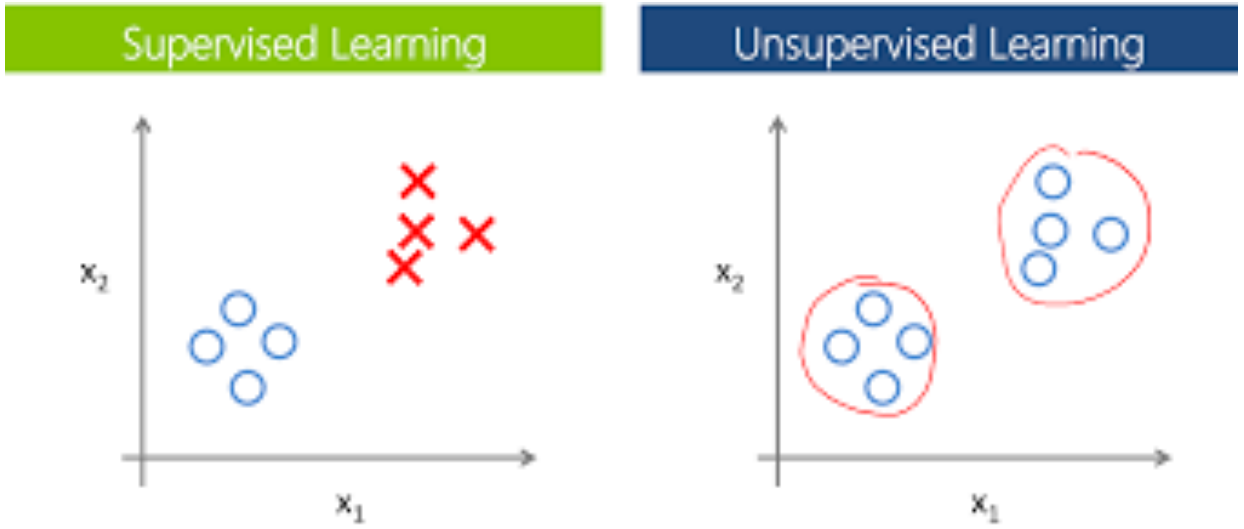
Clustering

Modelos Estatísticos para a Inteligência Artificial

Docente: Alexandra Oliveira
Email: aaoliveira@ipca.pt ou Alexandra.a.oliveira@gmail.com

Métodos de Aprendizagem Não Supervisionada - Clustering

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



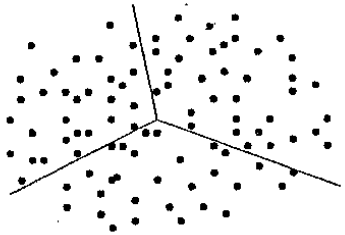
Análise de Clusters

- 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 interessantes hipóteses sobre associação entre variáveis

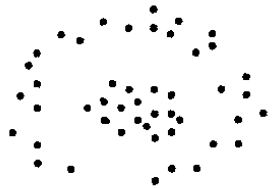
Análise de Clusters

- Formas de Clusters (ou grupos)

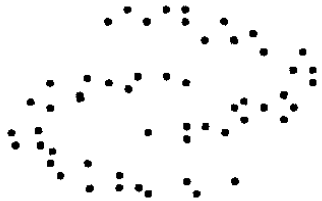
Grupos Coesos e bem separados



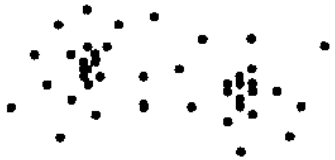
Grupo homogéneo sem clusters naturais



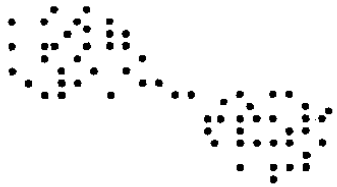
Grupo separados mas não coesos



Grupos separados mas sem coesão interna



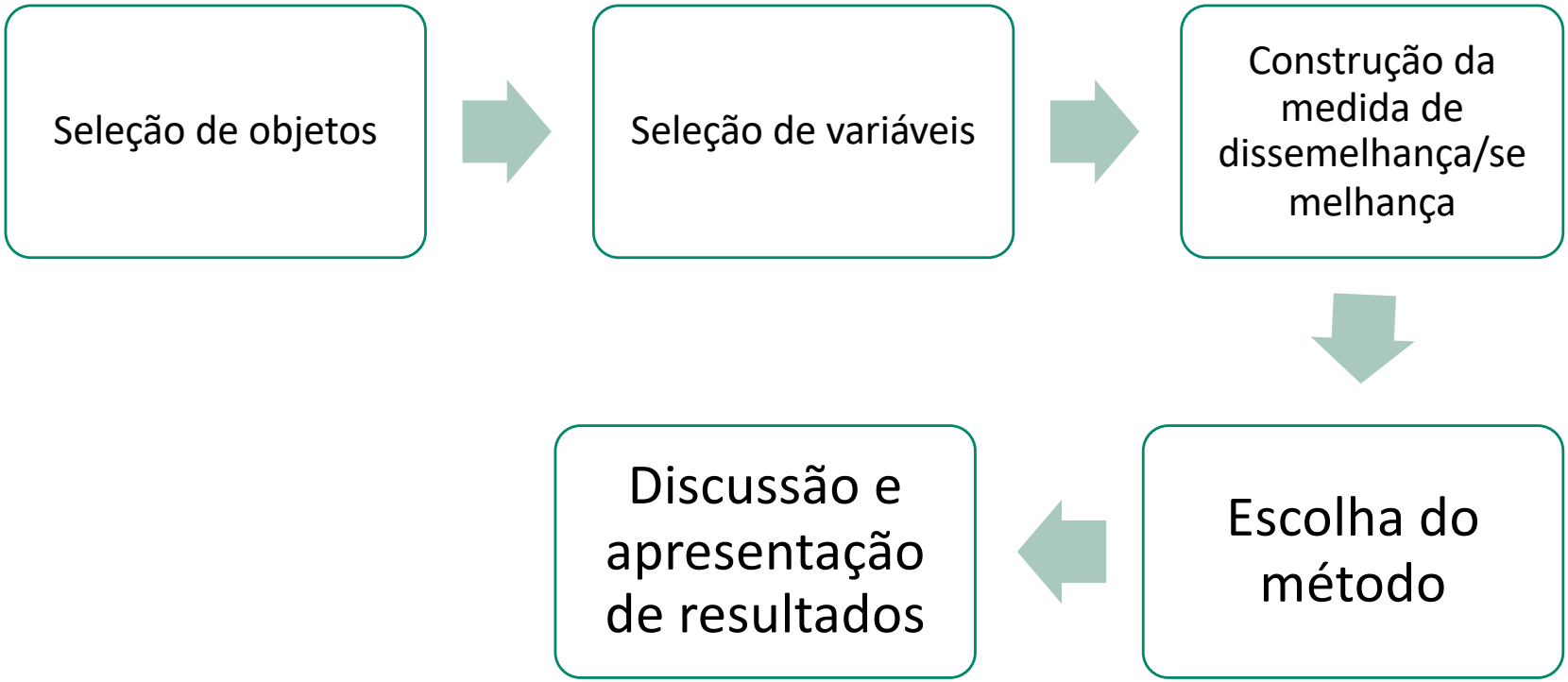
Zonas de grande densidade rodeadas por regiões de pequena densidade



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

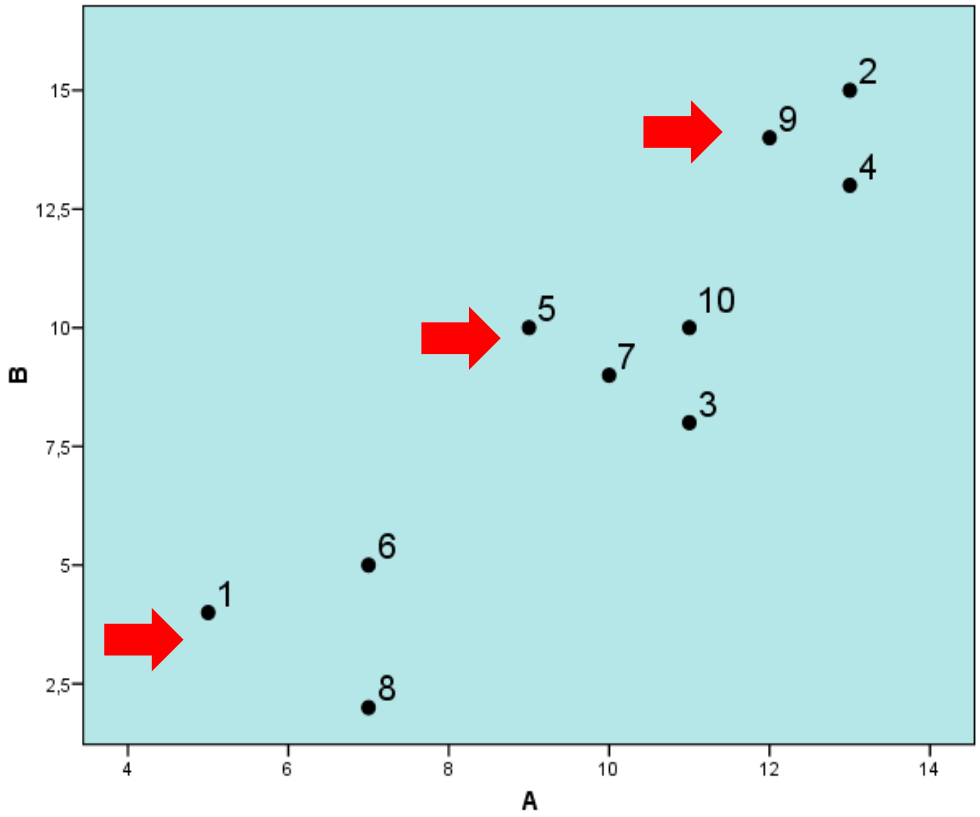
- Exemplo

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						

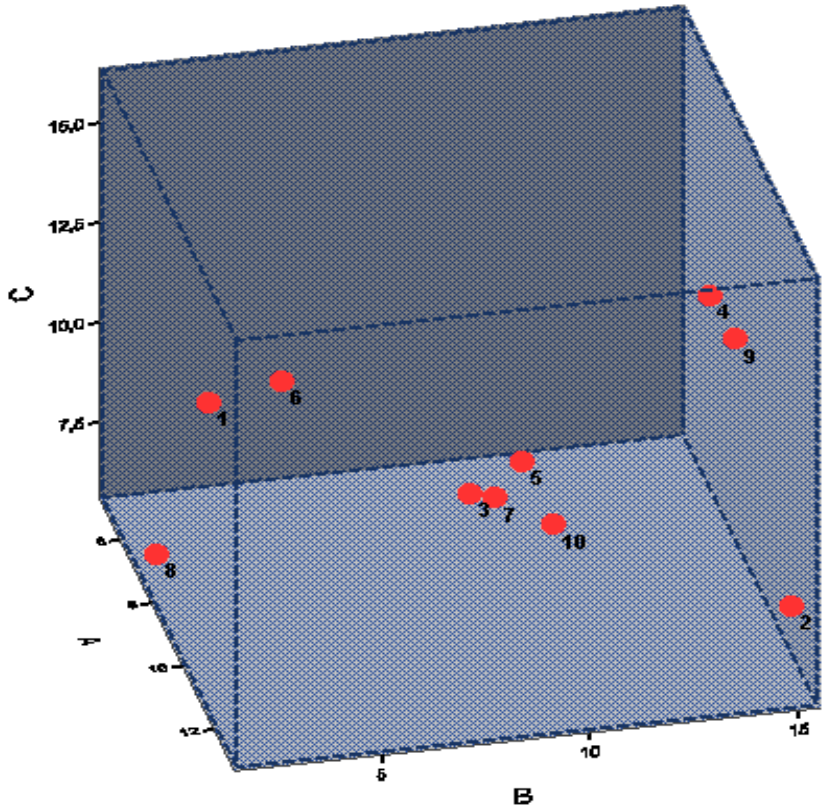
Análise de Clusters

- 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

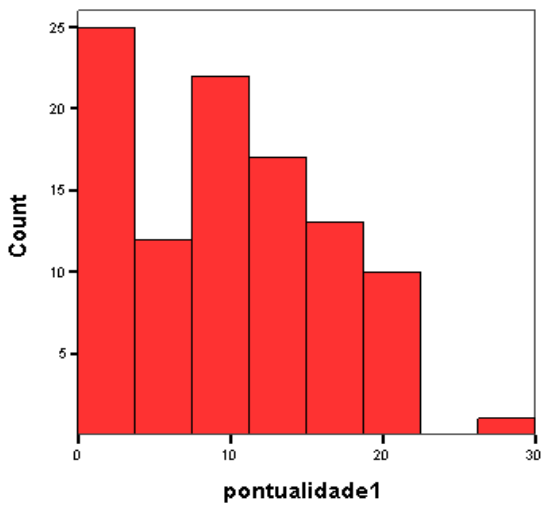
- Quanto menor a distância euclidiana menor é a dissemelhança (ou maior é a semelhança ou proximidade) entre indivíduos

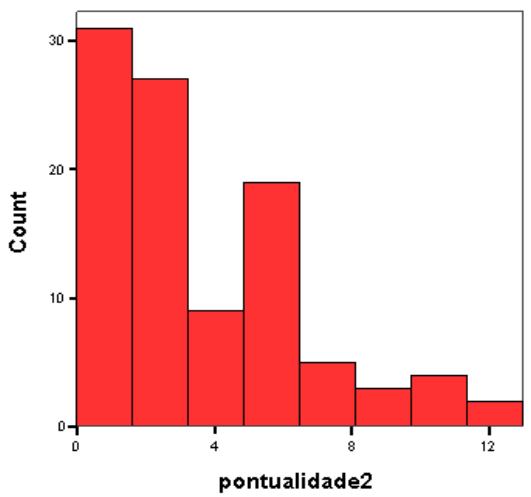
Análise de Clusters

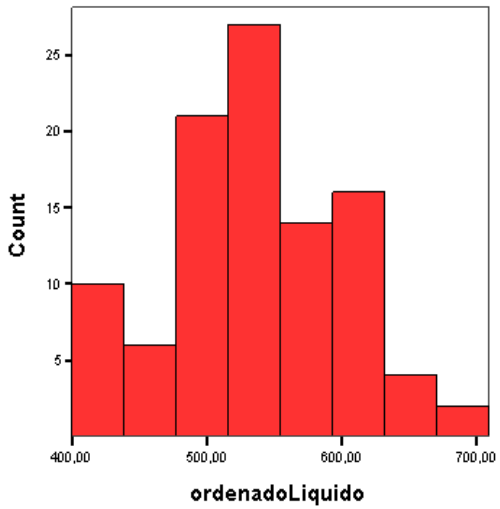
- Agrupar os sujeitos em clusters homogéneos
  - a partir das medidas de dissemelhança
  - de modo que dentro do mesmo cluster essas medidas sejam as menores possíveis
  - e entre clusters as maiores possíveis

Análise de Clusters

- Histograma – 1 variável



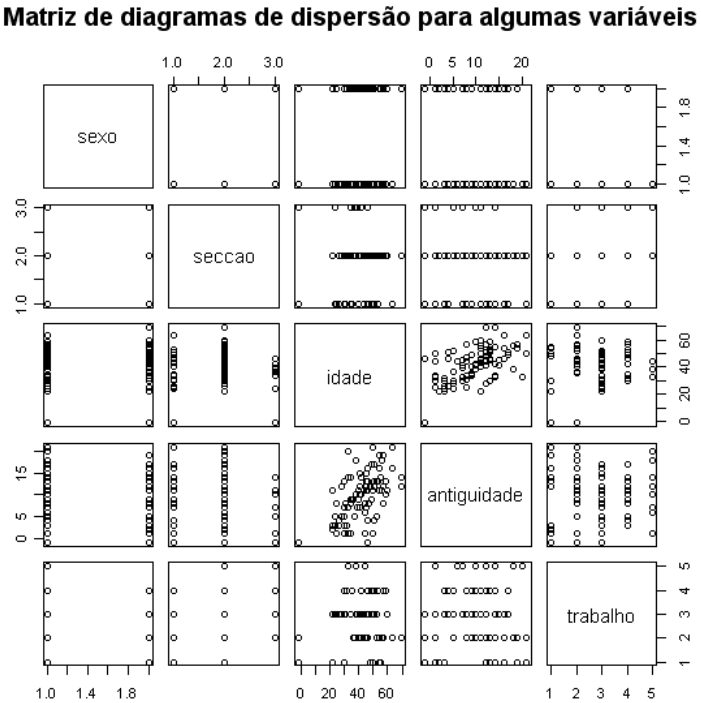




- A existência de várias modas é, em geral, reveladora da existência de clusters
- Existência de outros métodos para a representação gráfica, por ex, gráficos de barras, circulares e gráficos de caule-e-folhas

Análise de Clusters

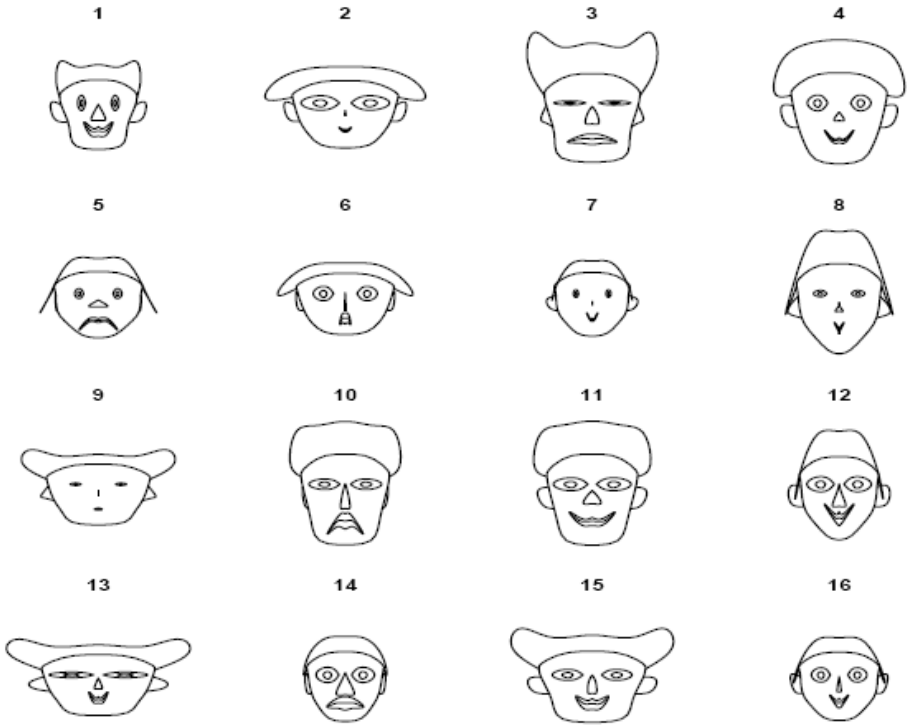
- Diagrama de Dispersão – 2 variáveis



- 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

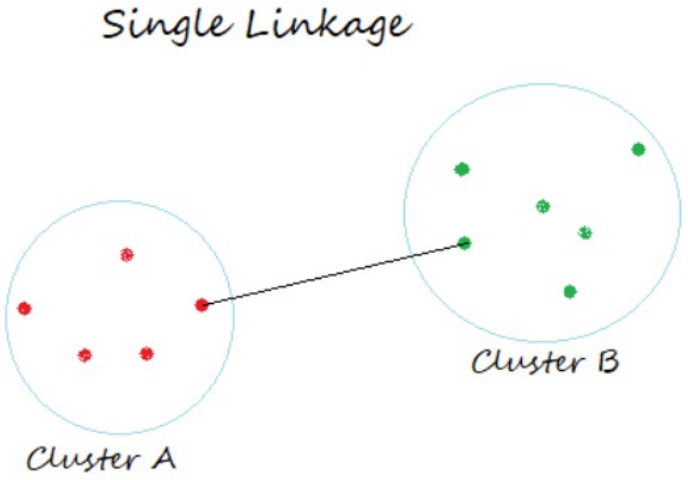


Análise de Clusters

- 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

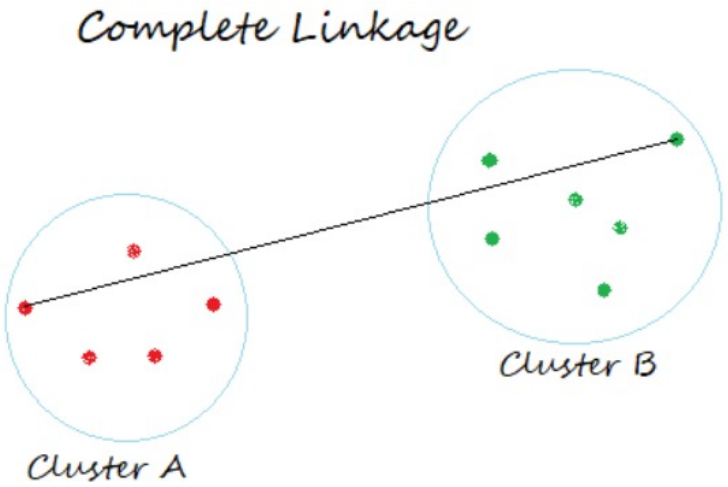
Análise de Clusters

- 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)



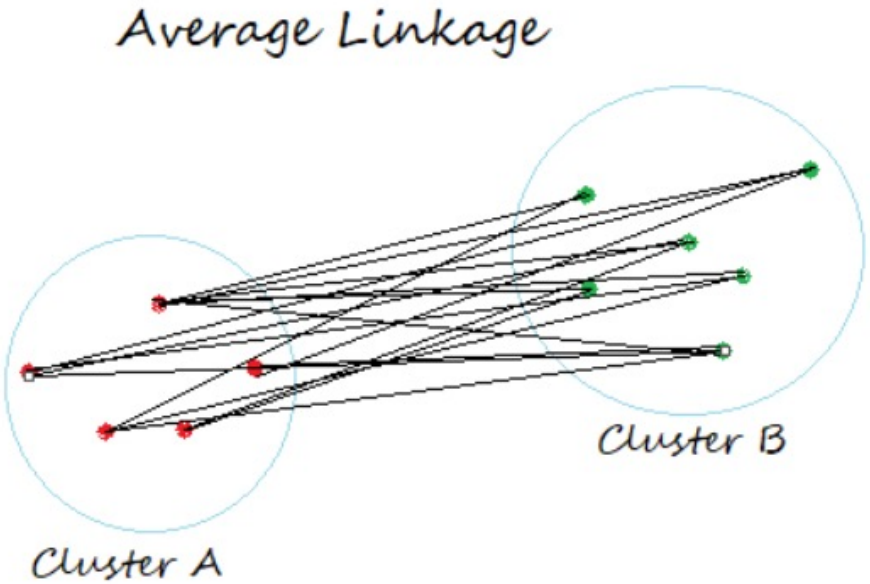
Análise de Clusters

- 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)



Análise de Clusters

- Métodos Hierárquicos - Método 1 – Aglomerativos
  - **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)

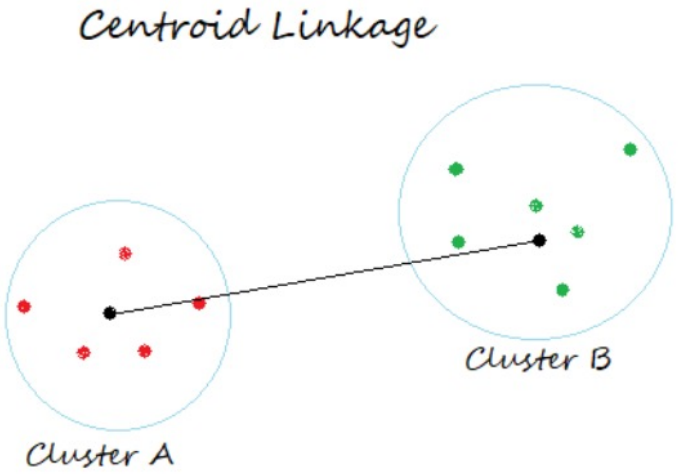


Análise de Clusters

- 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)



Análise de Clusters

- Métodos Hierárquicos - Método 1 - Aglomerativos
  - 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

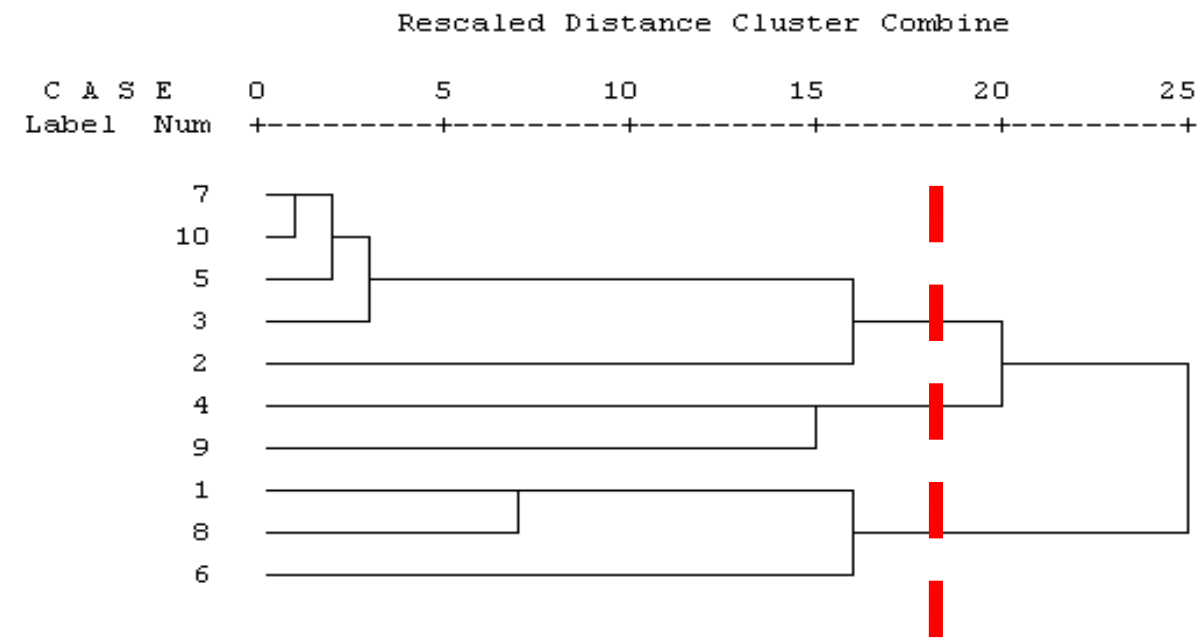
- Métodos Hierárquicos - Verificação e validação dos clusters
  - Construção de um diagrama de árvore hierárquica: **Dendrograma**
    - Contém parêntesis ligando dados e mostra a ordem pela qual os pontos estão assinalados para os grupos
    - Os comprimentos das suas ligações são proporcionais às distâncias 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 reescalonadas

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%)

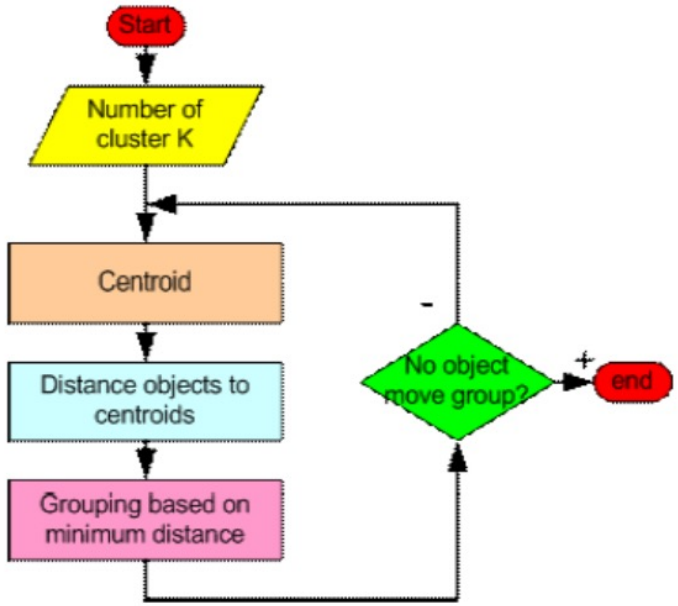
$$R-squared = \frac{SQC}{SQT} = \frac{\sum_{i=1}^{p}\sum_{j=1}^{k} n_{ij}\left(\overline{X}_{ij} - \overline{X}_{i}\right)^2}{\sum_{i=1}^{p}\sum_{j=1}^{k}\sum_{l=1}^{n_i}\left(X_{ijl} - \overline{X}\right)^2}$$

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 dissemelhança em cada passo do algoritmo
  - A inclusão de um indivíduo num cluster poderá não ser definitiva

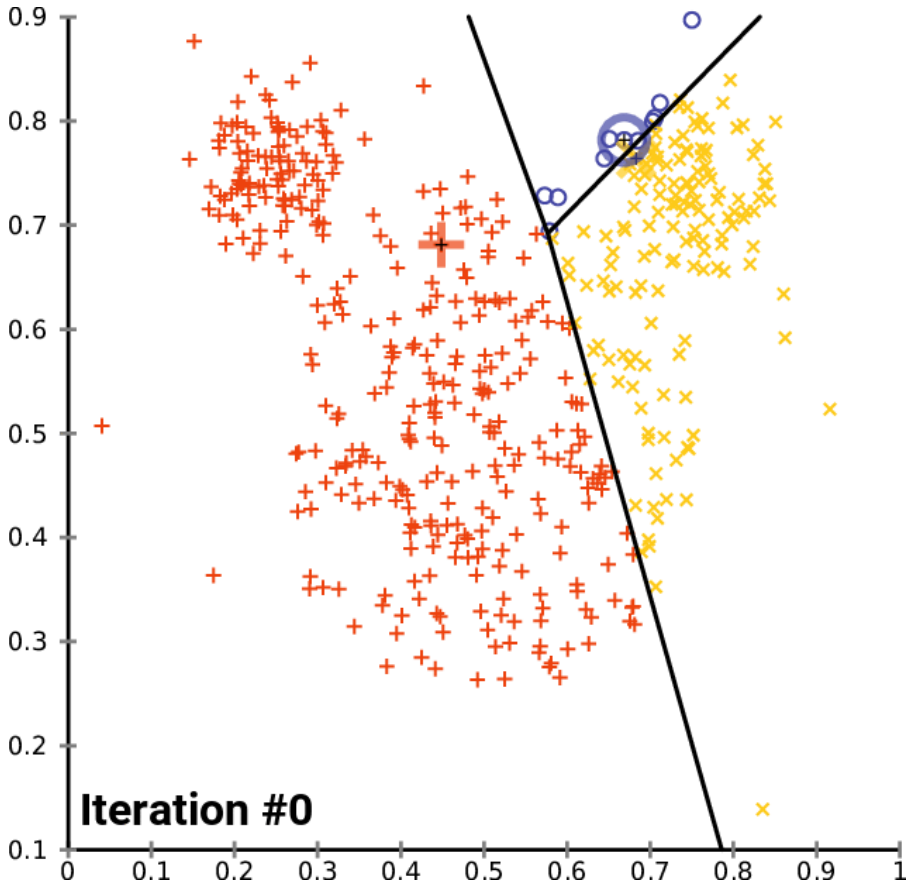
Análise de Clusters

- 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)



http://stanford.edu/class/ee103/visualizations/kmeans/kmeans.html
http://shabal.in/visuals/kmeans/4.html

Análise de Clusters



Determinação do número de clusters

- A package clValid do R permite comparar várias técnicas de clustering usando diversas métricas:
  - **Medidas internas:** utilizam informações intrínsecas dos dados para avaliar a qualidade do agrupamento. As medidas internas incluem a conectividade, o coeficiente de silhueta e o índice de Dunn.
  - **Medidas de estabilidade:** avalia a consistência de um resultado de agrupamento comparando-o com os agrupamentos obtidos após a remoção de cada coluna, um de cada vez.

  - **Escolhe-se o número de cluster que obtiver maior consenso entre os métodos.**

3.2 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

Clustering no R

- Vamos carregar a tradicional base de dados nativa do RStudio mtcars, que traz informações sobre 32 modelos de automóveis, sendo as respectivas variáveis que os descrevem:
- **mpg**: milhas por galão;
- **cyl**: número de cilindros;
- **disp**: número que representa o volume total no motor como um fator de circunferência do cilindro, profundidade e número total de cilindros;
- **hp**: potência;
- **drat**: relação do eixo traseiro;
- **wt**: peso (1.000 lbs);
- **qsec**: tempo de 1/4 de milha;
- **vs**: motor (0 = em forma de V; 1 = linha reta);
- **am**: transmissão (0 = automático; 1 = manual);
- **gear**: número de marchas na transmissão (3-4 automático; 4-5 manual);
- **carb**: número de carburadores;

Clustering no R

```r
#Carregamento dos dados
data("mtcars")
df=scale(mtcars)
head(df, n=3)
```

```
                mpg    cyl    disp      hp   drat      wt    qsec     vs   am
Mazda RX4     0.1509 -0.105 -0.5706 -0.5351 0.5675 -0.6104 -0.7772 -0.868 1.19
Mazda RX4 Wag 0.1509 -0.105 -0.5706 -0.5351 0.5675 -0.3498 -0.4638 -0.868 1.19
Datsun 710    0.4495 -1.225 -0.9902 -0.7830 0.4740 -0.9170  0.4260  1.116 1.19
                gear    carb
Mazda RX4     0.4236  0.7352
Mazda RX4 Wag 0.4236  0.7352
Datsun 710    0.4236 -1.1222
```

Clustering no R - kmeans

```r
# Clusterização k-means
set.seed(123)
km.res=kmeans(df, 4, nstart=25)
print(km.res)
```

```
Cluster means:
      mpg     cyl     disp      hp     drat       wt    qsec     vs      am
1 -0.8363  1.0149  1.02385  0.6925 -0.88975  0.90636 -0.3952 -0.868 -0.8141
2  1.3248 -1.2249 -1.10627 -0.9453  1.09821 -1.20087  0.3365  0.868  1.1899
3  0.1082 -0.5849 -0.44867 -0.6497 -0.04968 -0.02347  1.1855  1.116 -0.8141
4 -0.2639  0.3430 -0.05908  0.7601  0.44782 -0.22101 -1.2495 -0.868  1.1899
     gear    carb
1 -0.9318  0.1677
2  0.7624 -0.8126
3 -0.1573 -0.4146
4  1.2368  1.4781

Clustering vector:
          Mazda RX4       Mazda RX4 Wag          Datsun 710      Hornet 4 Drive
                  4                   4                   2                   3
  Hornet Sportabout             Valiant          Duster 360           Merc 240D
                  1                   3                   1                   3
           Merc 230            Merc 280           Merc 280C          Merc 450SE
                  3                   3                   3                   1
        Merc 450SL         Merc 450SLC  Cadillac Fleetwood Lincoln Continental
                  1                   1                   1                   1
  Chrysler Imperial            Fiat 128         Honda Civic      Toyota Corolla
                  1                   2                   2                   2
     Toyota Corona    Dodge Challenger         AMC Javelin          Camaro Z28
                  3                   1                   1                   1
   Pontiac Firebird           Fiat X1-9       Porsche 914-2        Lotus Europa
                  1                   2                   2                   2
     Ford Pantera L        Ferrari Dino       Maserati Bora          Volvo 142E
                  4                   4                   4                   2

Within cluster sum of squares by cluster:
[1] 23.08 19.04 21.29 23.40
 (between_SS / total_SS =  74.5 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

Clustering no R - kmeans

```
mtcars2=cbind(mtcars, cluster=km.res$cluster)
head(mtcars2)
```

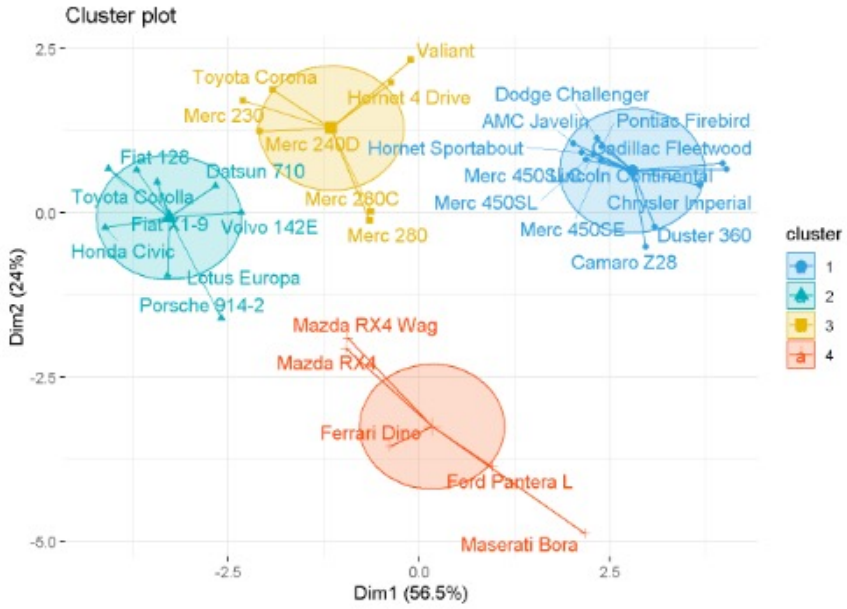
	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb	cluster
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1	2
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1	3
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2	1
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1	3

Clustering no R - kmeans

```
# Vizualizando os clusters

library(ggplot2)
library(factoextra)

fviz_cluster(km.res, data=mtcars2,
             palette = c("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),
             ellipse.type="euclid",
             star.plot=TRUE,
             repel=TRUE,
             ggtheme=theme_minimal()
             )
```



Clustering - hierarquico

```r
dista=dist(df, method="euclidean")
```

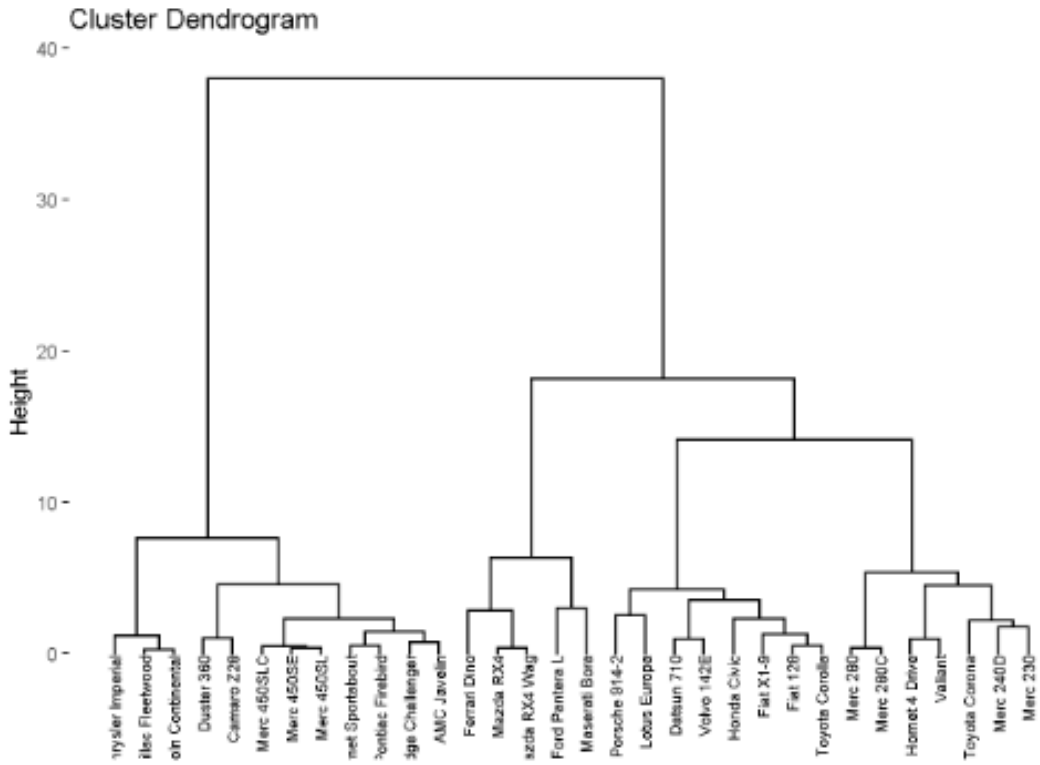
```r
as.matrix(dista)[1:3,1:3]
```

```
              Mazda RX4 Mazda RX4 Wag Datsun 710
Mazda RX4        0.0000         0.4076      3.243
Mazda RX4 Wag    0.4076         0.0000      3.176
Datsun 710       3.2431         3.1764      0.000
```

```r
dista.hc=hclust(d=dista, method="ward.D")
```

Clustering - hierarquico

```
library("factoextra")
fviz_dend(dista.hc, cex=0.5)
```



Clustering - clValid

```
clValid(obj, nClust, clMethods = "hierarchical",
        validation = "stability", maxitems = 600,
        metric = "euclidean", method = "average")
```

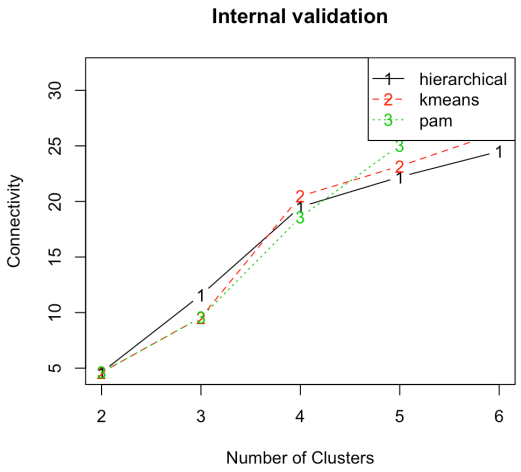
```
# Compute clValid
clmethods <- c("hierarchical","kmeans","pam")
intern <- clValid(exprs, nClust = 2:6,
              clMethods = clmethods, validation = "internal")
# Summary
summary(intern)
```

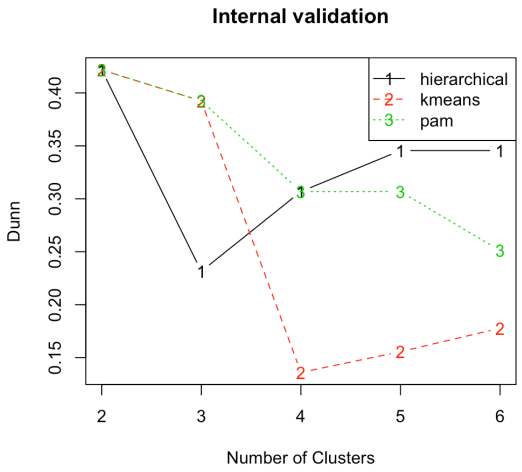
Clustering - clValid

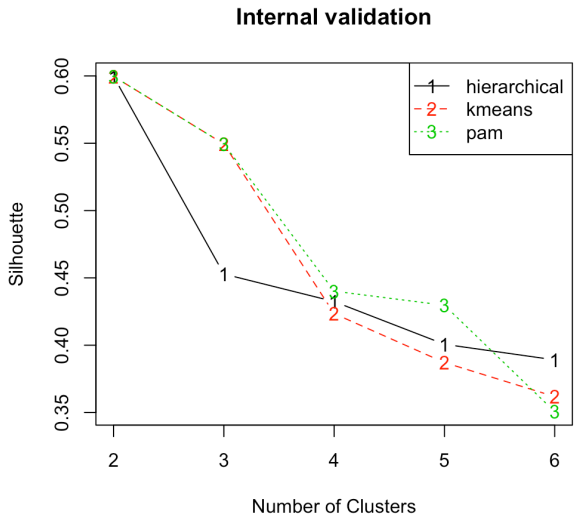
```
##
## Clustering Methods:
##  hierarchical kmeans pam
##
## Cluster sizes:
##  2 3 4 5 6
##
## Validation Measures:
##                                 2       3       4       5       6
##
## hierarchical Connectivity   4.6159 11.5865 19.5075 22.2075 24.5044
##              Dunn           0.4217  0.2315  0.3068  0.3456  0.3456
##              Silhouette     0.5997  0.4529  0.4324  0.4007  0.3891
## kmeans       Connectivity   4.6159  9.5607 20.4774 23.1774 26.2242
##              Dunn           0.4217  0.3924  0.1360  0.1556  0.1778
##              Silhouette     0.5997  0.5495  0.4235  0.3871  0.3618
## pam          Connectivity   4.6159  9.5607 18.5925 25.0631 31.8381
##              Dunn           0.4217  0.3924  0.3068  0.3068  0.2511
##              Silhouette     0.5997  0.5495  0.4401  0.4297  0.3506
##
## Optimal Scores:
##
##              Score  Method       Clusters
## Connectivity 4.6159 hierarchical 2
## Dunn         0.4217 hierarchical 2
## Silhouette   0.5997 hierarchical 2
```

Clustering – clValid

```
plot(intern)
```







Clustering - clValid

```
# Stability measures
clmethods <- c("hierarchical","kmeans","pam")
stab <- clValid(exprs, nClust = 2:6, clMethods = clmethods,
                validation = "stability")
# Display only optimal Scores
optimalScores(stab)
```

```
##          Score       Method Clusters
## APN 0.0000000 hierarchical        2
## AD  0.9642344          pam        6
## ADM 0.0000000 hierarchical        2
## FOM 0.3925939          pam        6
```

```
summary(stab)
plot(stab)
```

Bibliografia

- Tan, P., Steinbach, M. & Kumar, V. (2006). Introduction to Data Mining. Pearson Addison-Wesley.
- Adaptação de slides de: http://www-users.cs.umn.edu/~kumar/dmbook/index.php
- Adaptação de slides de: Gladys Castillo, Aprendizagem Computacional (Machine Learning), Universidade de Aveiro
- 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, Edições Sílabo, Lda, Abril, 2003. ISBN: 972-618-298-0
- Dawson-Saunders B, Trapp G (2004) Basic and Clinical Biostatistics, Fourth Edition. Prentice-Hall International Inc