FIAP

MBA+ | Artificial Intelligence & Machine Learning


AI

Aprendizado não supervisionado

K-means / EM / Regras de Associação

Prof. Dr. Vinicius F. Caridá

Algoritmos Hierárquicos

Hierarquia: Conceitos Básicos

Hierarquias são comumente usadas para organizar informação

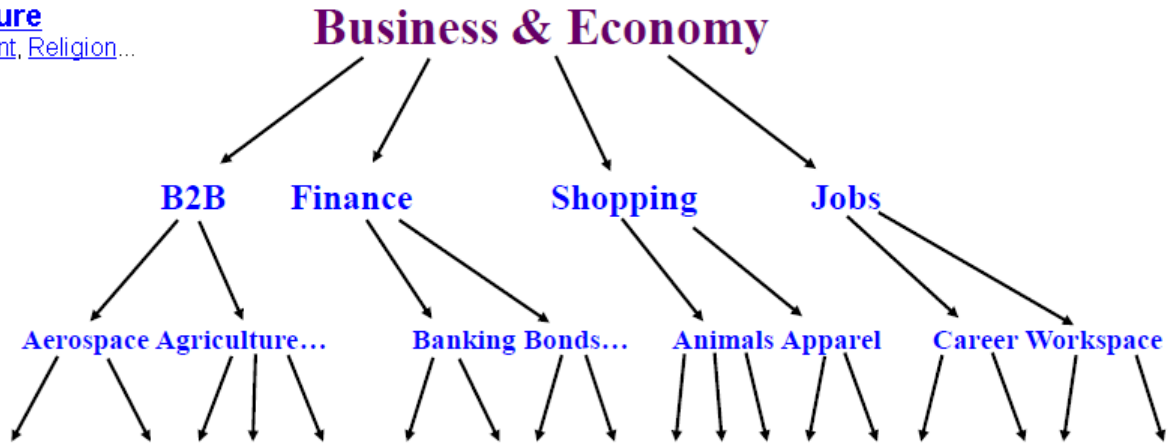
Web Site Directory - Sites organized by subject Suggest your site

Business & Economy
B2B, Finance, Shopping, Jobs...

Regional
Countries, Regions, US States...

Computers & Internet
Internet, WWW, Software, Games...

Society & Culture
People, Environment, Religion...



Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Hierarquia: Conceitos Básicos

Exemplo: árvores filogenéticas em biologia

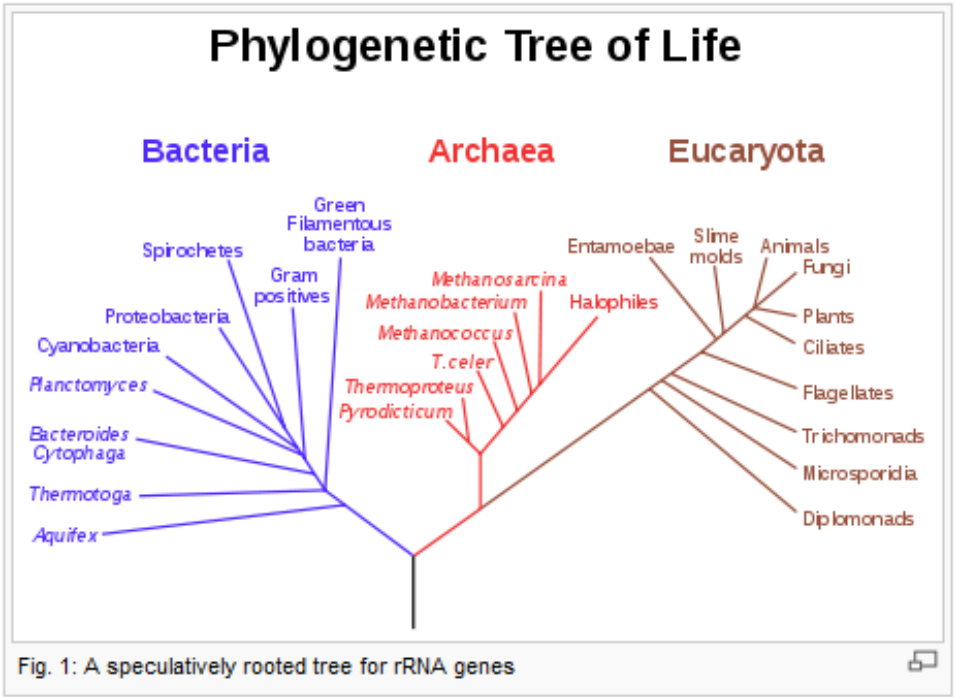
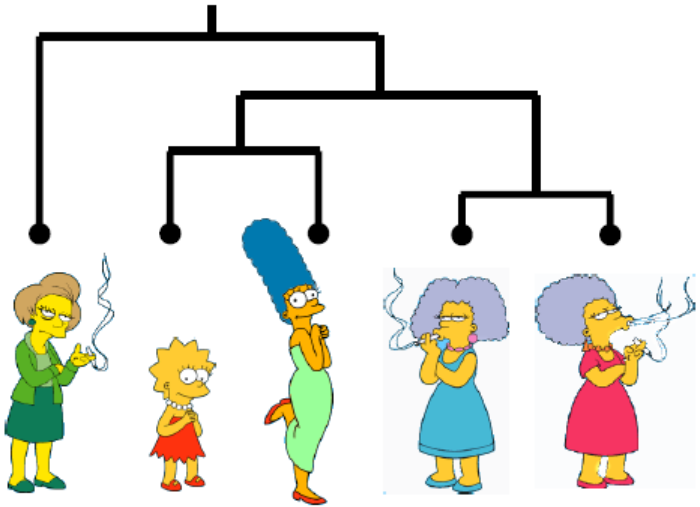


Fig. 1: A speculatively rooted tree for rRNA genes

http://en.wikipedia.org/wiki/Phylogenetic_tree

Métodos Clássicos para Agrupamento Hierárquico



Bottom-Up (aglomerativos):

- Iniciar colocando cada objeto em um *cluster*
- Encontrar o melhor par de *clusters* para unir
- Unir o par de *clusters* escolhido
- Repetir até que todos os objetos estejam reunidos em um só *cluster*

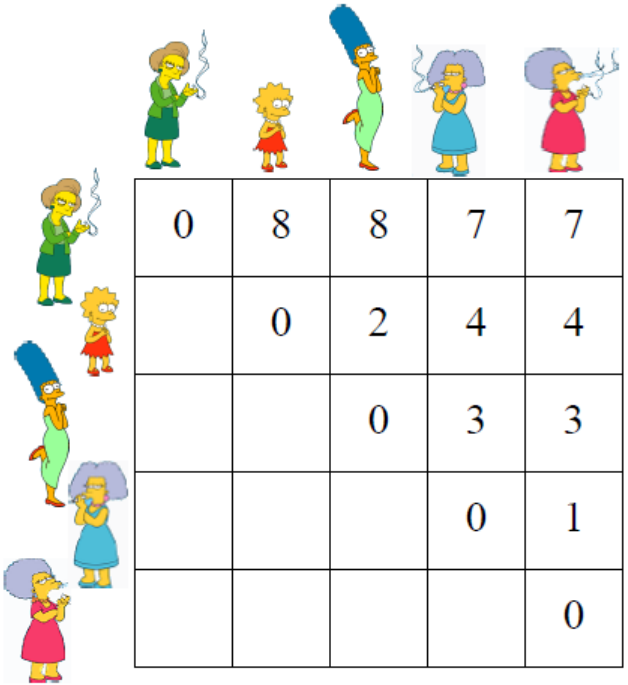
Top-Down (divisivos):

- Iniciar com todos objetos em um único *cluster*
- Sub-dividir o *cluster* em dois novos *clusters*
- Aplicar o algoritmo recursivamente em ambos, até que cada objeto forme um *cluster* por si só

Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Métodos Clássicos para Agrupamento Hierárquico

Algoritmos hierárquicos podem operar somente sobre uma matriz de distâncias.



0	8	8	7	7
	0	2	4	4
		0	3	3
			0	1
				0

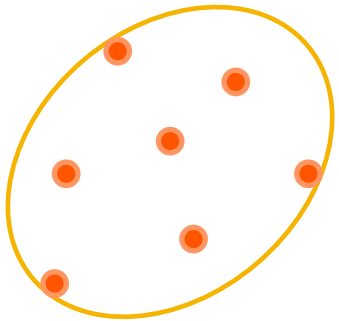
$D(\,,\,) = 1$ $D(\,,\,) = 8$

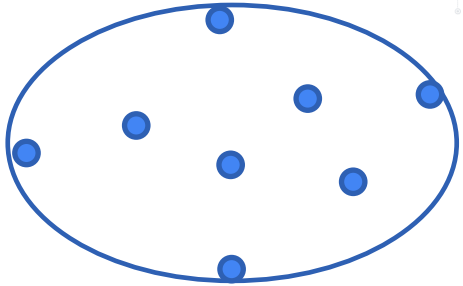
Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Como definir Inter-Cluster (Dis)similaridade

Matriz de Distância

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						





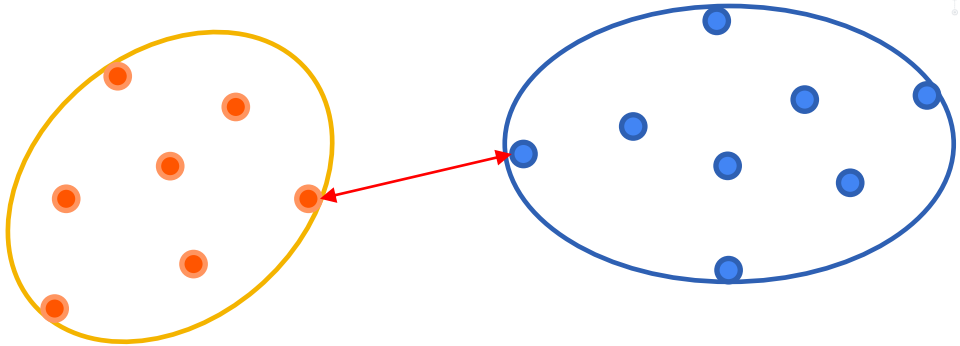
- MIN
- MAX
- Group Average
- ...

Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Como definir Inter-Cluster (Dis)similaridade

Matriz de Distância

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

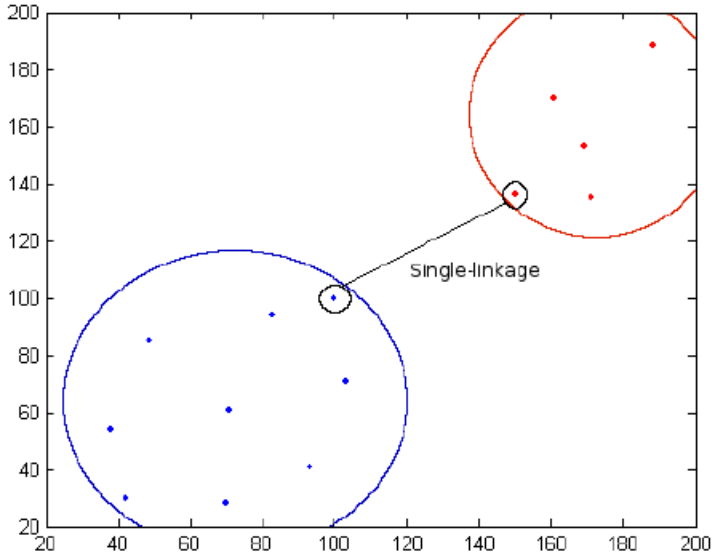


- MIN
- MAX
- Group Average
- ...

Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Single Linkage (Florek, 1951)

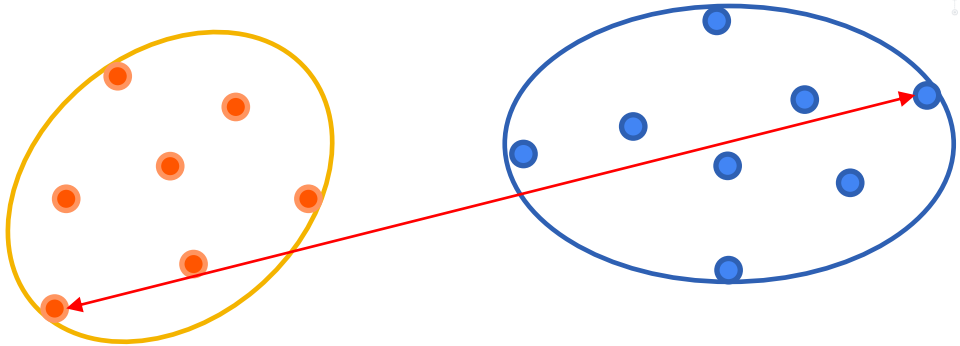
- Dissimilaridade entre clusters é dada pela **menor** dissimilaridade entre 2 objetos (um de cada cluster)
  - Originalmente baseado em Grafos: **menor** aresta entre dois vértices de subconjuntos distintos



Como definir Inter-Cluster (Dis)similaridade

Matriz de Distância

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

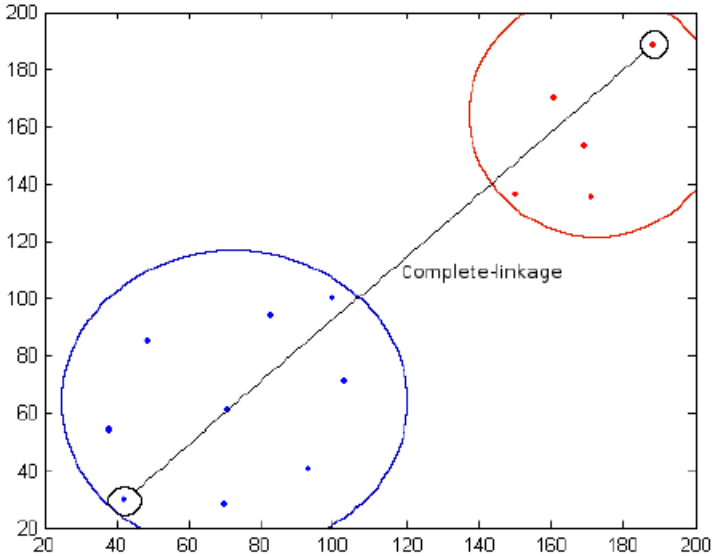


- MIN
- MAX
- Group Average
- ...

Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Complete Linkage (Sorensen, 1948)

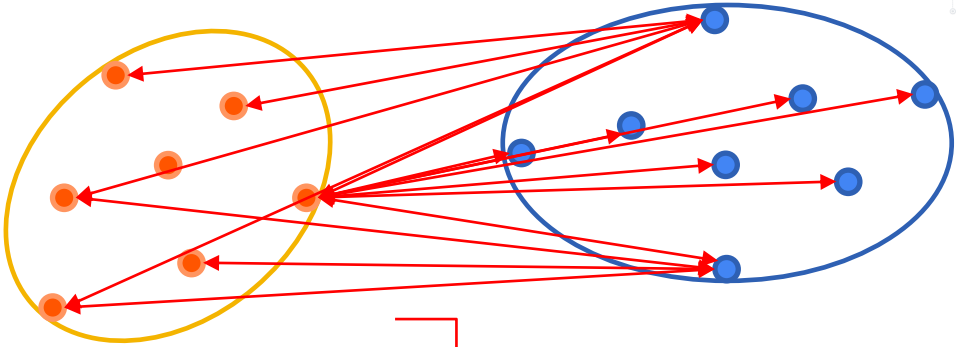
- Dissimilaridade entre clusters é dada pela **maior** dissimilaridade entre 2 objetos (um de cada cluster)
  - Originalmente baseado em Grafos: maior aresta entre dois vértices de subconjuntos distintos



Como definir Inter-Cluster (Dis)similaridade

Matriz de Distância

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						



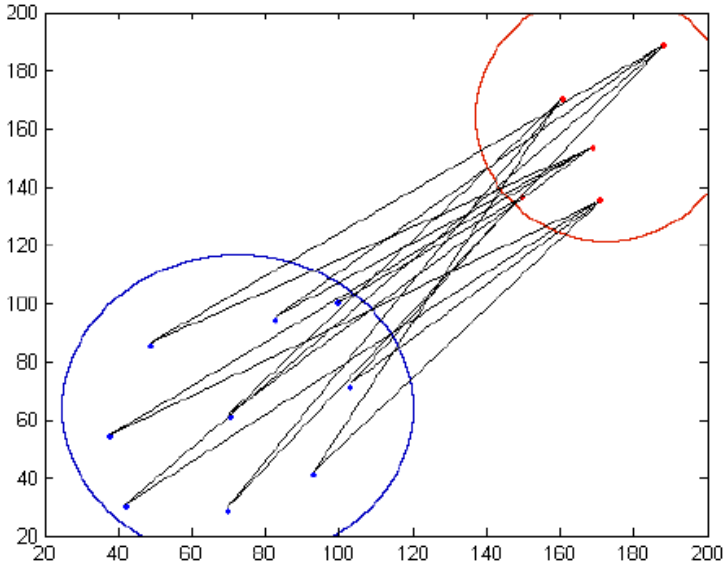
Todas as distâncias para-a-par

- MIN
- MAX
- Group Average
- ...

Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

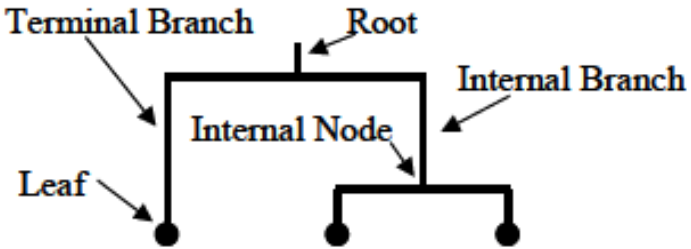
Complete Linkage (Sokal R and Michener C, 1958)

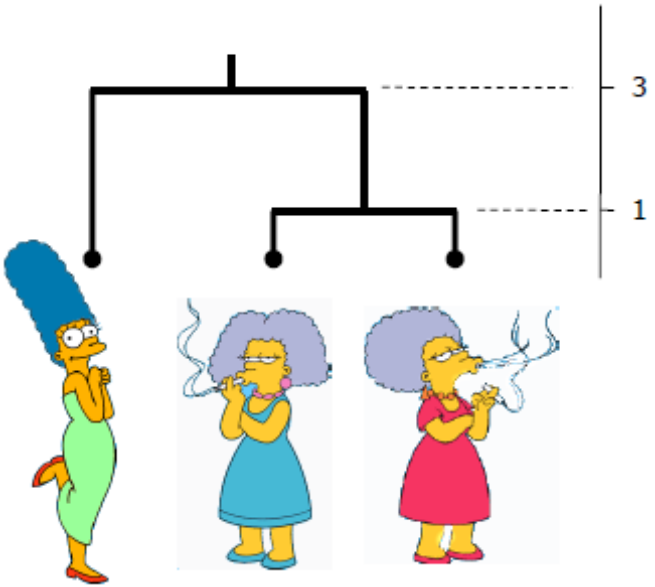
- Dissimilaridade entre clusters é dada pela **distância média** entre cada par de objetos (um de cada cluster)
- Também conhecido como UPGMA – Unweighted Pair Group Method using Arithmetic averages



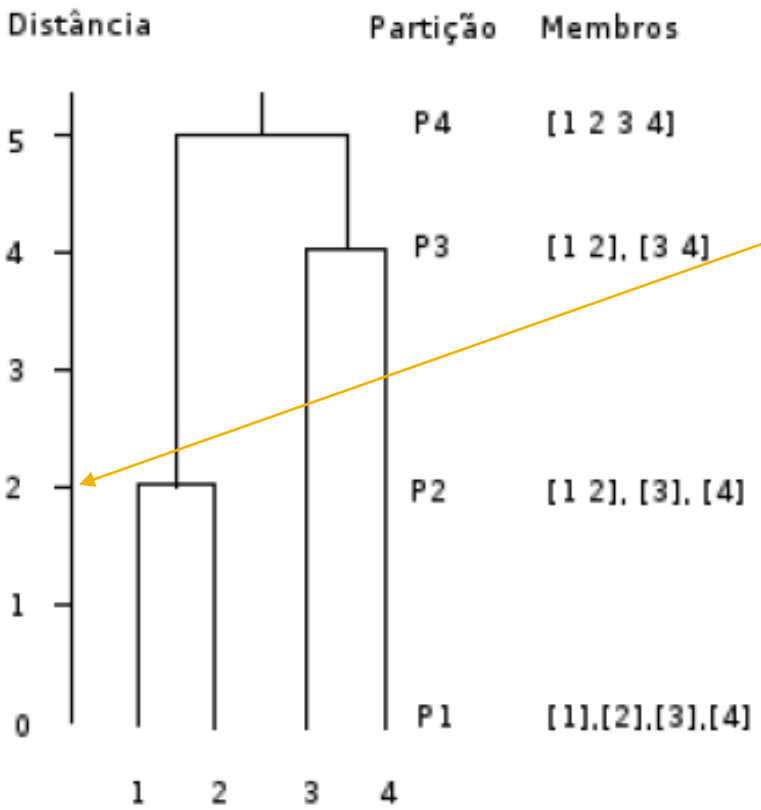
Dendrograma = Hierarquia + Dissimilaridade entre Clusters

A dissimilaridade entre dois clusters (possivelmente **singletons**) é representada como a altura do nó interno mais baixo compartilhado

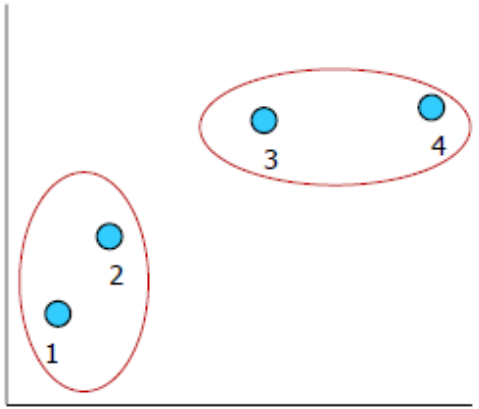




Dendrograma



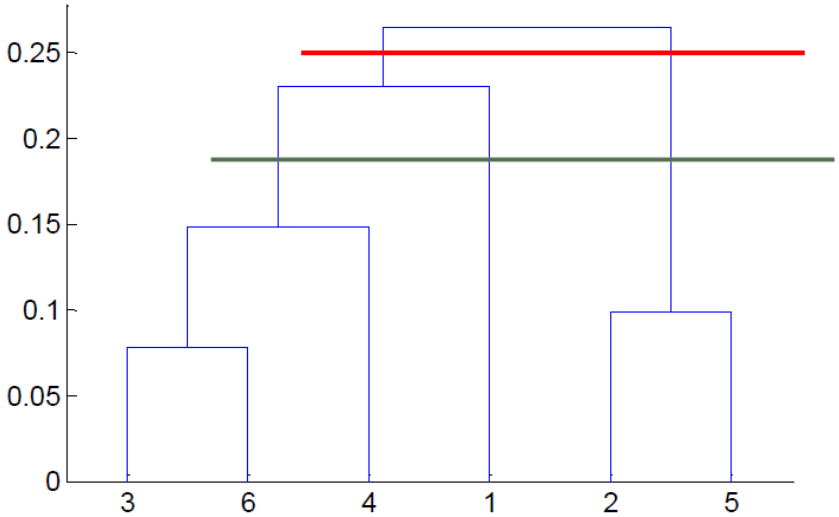
$$\mathbf{D} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{bmatrix} 0 & 2 & 7 & 13 \\ 2 & 0 & 5 & 10 \\ 7 & 5 & 0 & 4 \\ 13 & 10 & 4 & 0 \end{bmatrix}$$



Dendrograma -> Grupos

Partições são obtidas via **cortes** no dendrograma

- cortes horizontais
- no. de grupos da partição = no. de interseções

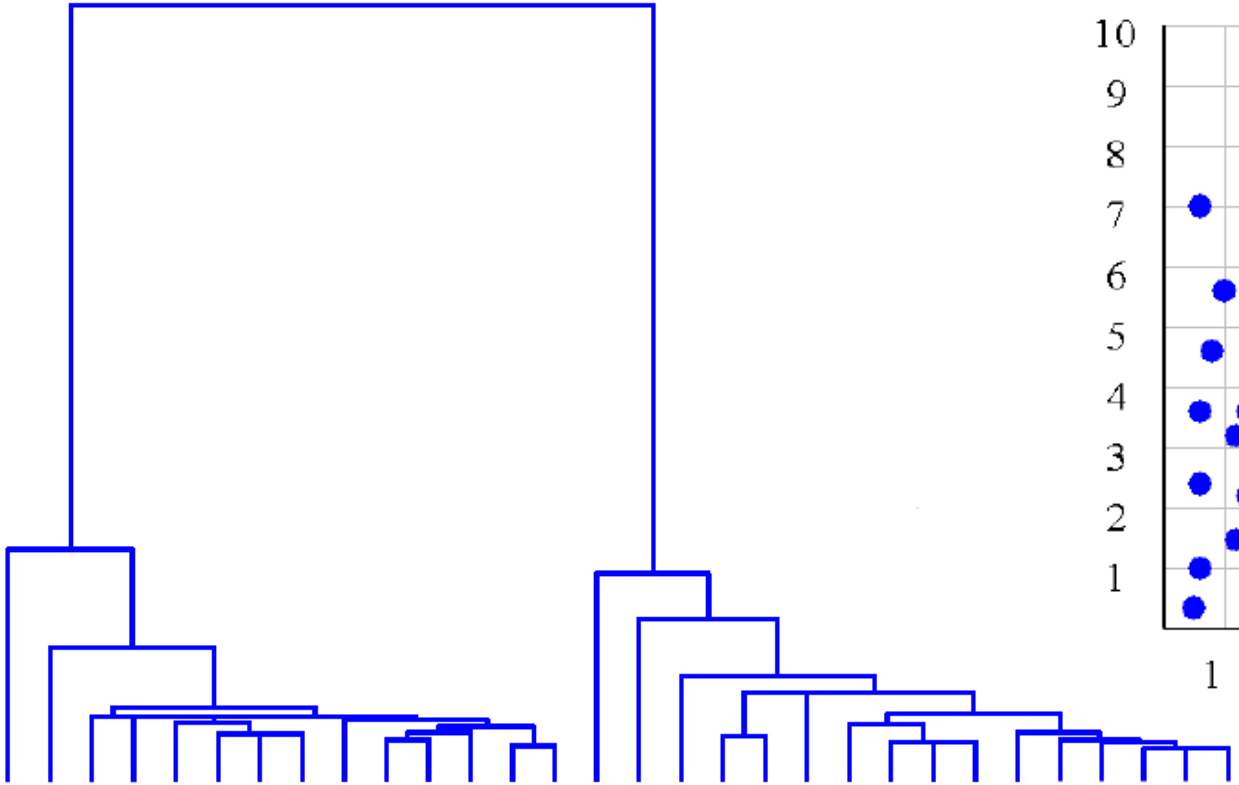


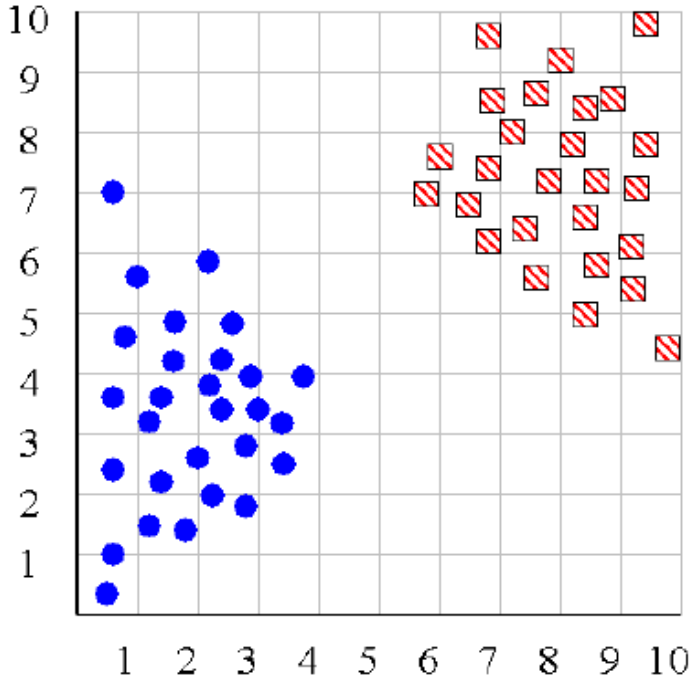
$$G_1 = \{ (x1, x3, x4, x6), (x2, x5) \}$$

$$G_2 = \{ (x1), (x3, x4, x6), (x2, x5) \}$$

Dendrograma -> Grupos

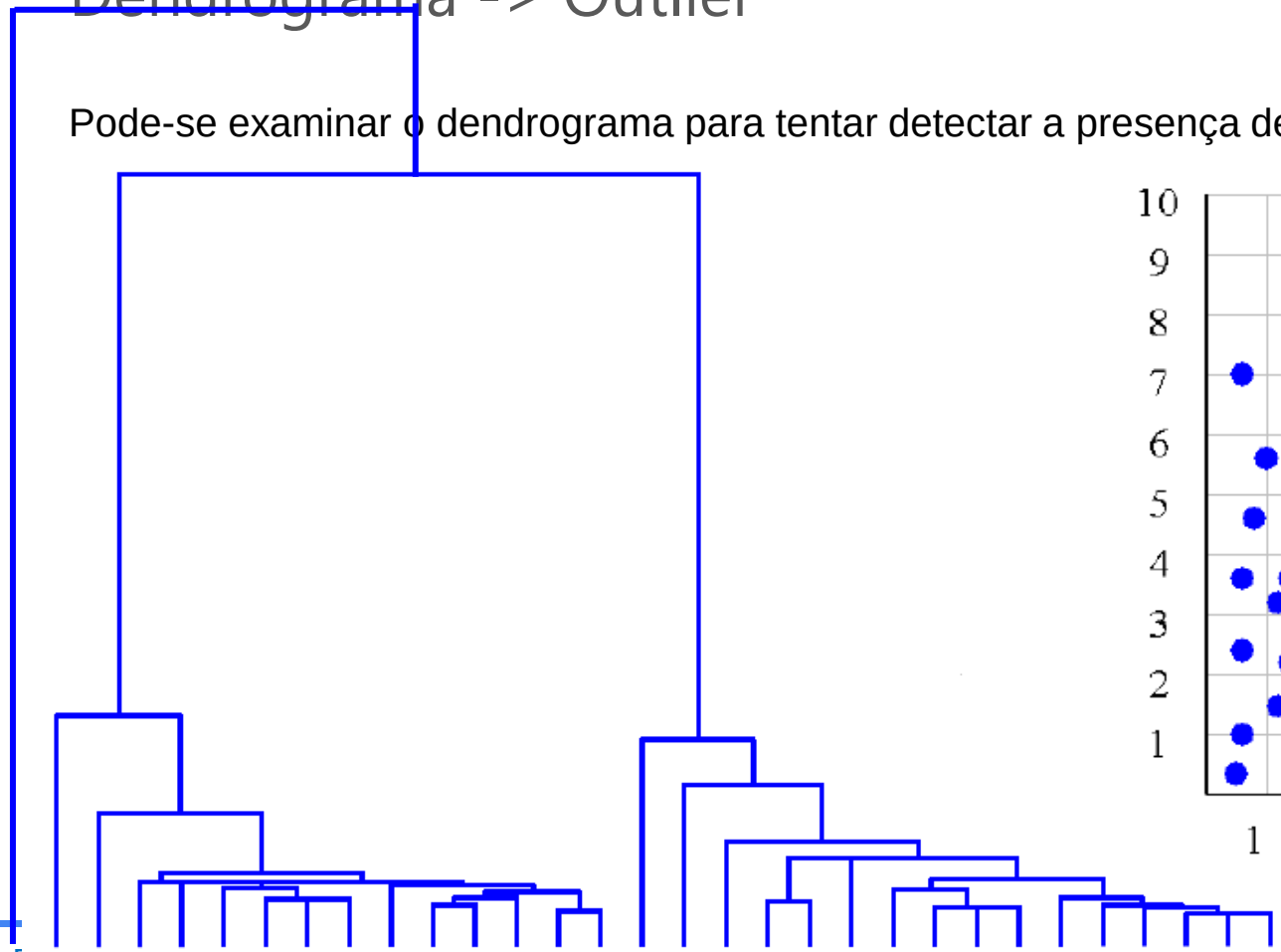
Pode-se examinar o dendrograma para tentar estimar o número mais natural de clusters.

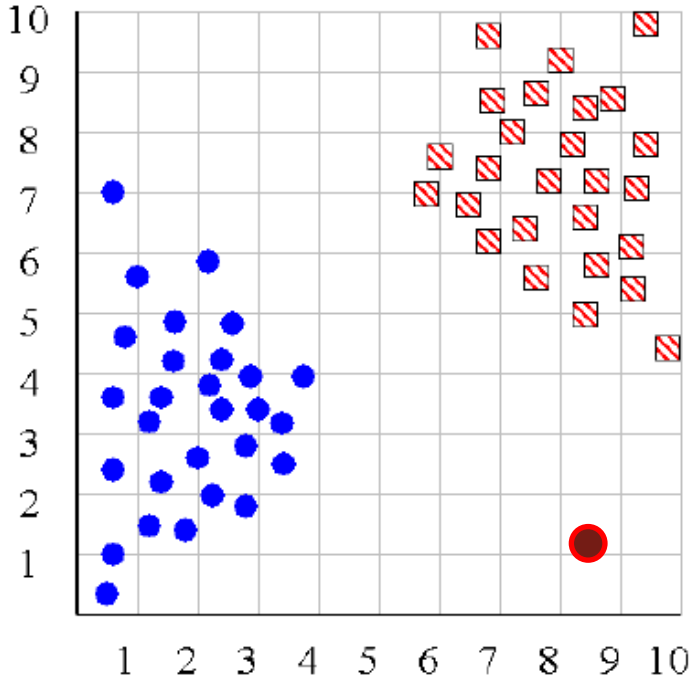




Dendrograma -> Outlier

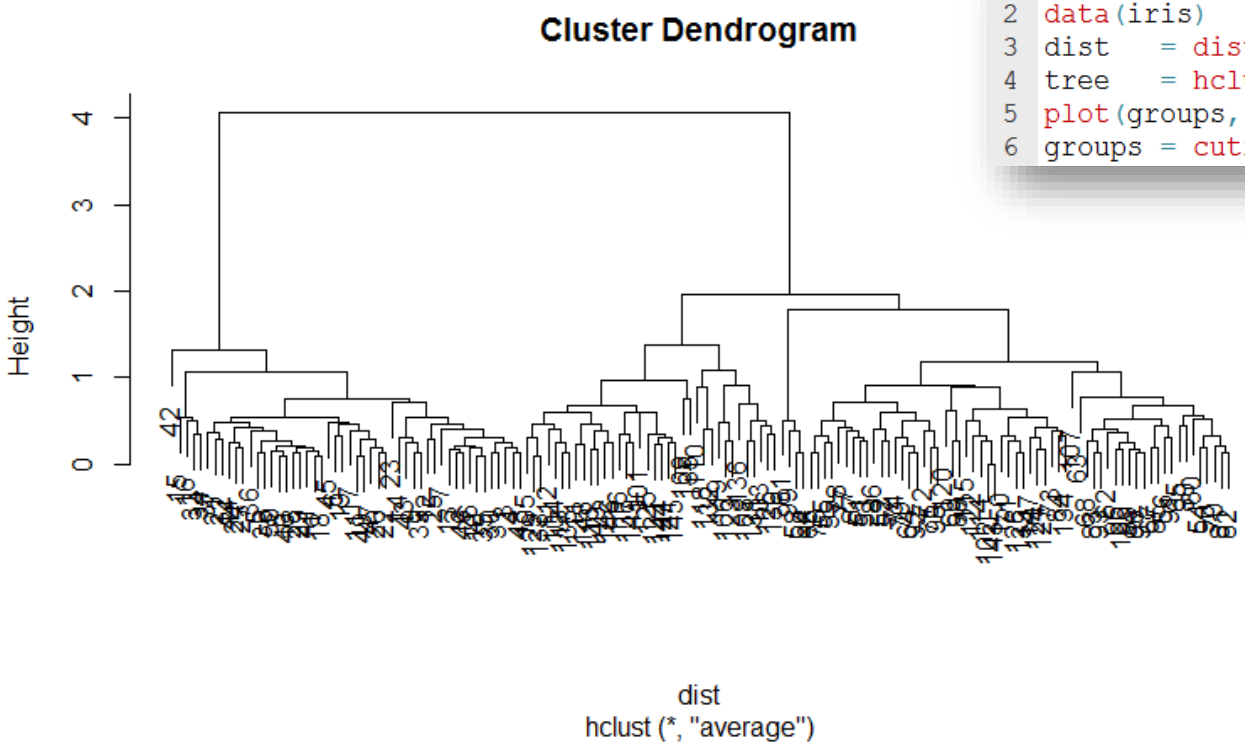
Pode-se examinar o dendrograma para tentar detectar a presença de outliers.





Hierárquico

```r
#Carrega os dados
data(iris)
dist    = dist(iris[1:4])
tree    = hclust(dist, method = "average")
plot(groups, col=unclass(iris$Species))
groups = cutree(tree, k=3)
```



Questions and Feedback

Thank you!

Obrigado !

Vinicius Fernandes Caridá
vfcarida@gmail.com

@Vinicius Caridá

@Vinicius Caridá
@machine learning Brasil

@vfcarida