Clusterização Hierárquica

EST0133 - Introdução à Modelagem de Big Data

Marcus Nunes
https://introbigdata.org/
https://marcusnunes.me/

Universidade Federal do Rio Grande do Norte

Introdução

Introdução

- Outra maneira de clusterizar dados
- Enquanto o k-means se utiliza de somas de quadrados e distância euclidiana, a clusterização hierárquica se baseia em outras distâncias e similaridades
- É outro algoritmo que também pode ser baseado no vizinho mais próximo
- Busca coesão interna (intra grupo) e isolamento externo (entre grupos)

Distância

Distância

- Distância é um conceito bem definido em matemática
- Uma distância em um espaço métrico X é qualquer função

$$d : X \times X \to [0, \infty)$$

que satisfaça as quatro condições abaixo para todo $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in X$:

i) $d(\boldsymbol{x}, \boldsymbol{y}) \geq 0$
ii) $d(\boldsymbol{x}, \boldsymbol{y}) = 0 \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}$
iii) $d(\boldsymbol{x}, \boldsymbol{y}) = d(y, x)$
iv) $d(\boldsymbol{x}, \boldsymbol{z}) \leq d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z})$

Distância

- Distância discreta: se $x = y$, então $d(x, y) = 0$. Caso contrário, $d(x,y) = 1$.
- Distância euclidiana: sejam $X = \mathbb{R}^k$ e sejam $x, y \in \mathbb{R}^k$ tais que $x = (x_1, \cdots, x_n)$ e $y = (y_1, \cdots, y_n)$. Assim,

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2} = \left[\sum_{i=1}^{n} (x_i - y_i)^2 \right]^{\frac{1}{2}}$$

Distância

- Distância de Manhattan:

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

- Distância de Minkowski (ou norma-p):

$$d(x, y) = \left[\sum_{i=1}^{n} |x_i - y_i|^p \right]^{\frac{1}{p}}$$

Distância

- Como Escolher a Distância a ser Usada?
- A distância euclidiana funciona muito bem em casos lineares, mas é sensível a outliers
- A distância Manhattan funciona melhor em casos com muitas observações aberrantes ou num espaço com muitas dimensões
- Outras distâncias podem ser usadas em outros casos
- Essa escolha é arbitrária

Ligação

Ligação

- Outro critério a ser utilizado na clusterização hierárquica é o método de ligação
- Assim como a distância escolhida, o método de ligação influenciará na clusterização final
- Também é uma escolha arbitrária

Ligação

Sejam $\boldsymbol{x} \in X$ e $\boldsymbol{y} \in Y$ vetores nos conjuntos X e Y e $d(\boldsymbol{x}, \boldsymbol{y})$ alguma distância entre estes vetores

- Simples: mínima distância ou vizinho mais próximo - $\min\{d(\boldsymbol{x}, \boldsymbol{y})\}$
- Completa: máxima distância ou vizinho mais distante - $\max\{d(\boldsymbol{x}, \boldsymbol{y})\}$
- Ward: mínima variância - $\min \|\boldsymbol{x} - \boldsymbol{y}\|^2$
- E muitos outros

Algoritmo

Algoritmo

1. Para a clusterização aglomerativa, comece com n nós de singletons (isto é, de um sujeito por nó)
2. Calcule a similaridade dois a dois entre todos os pares de nós
3. Junte os dois nós mais similares entre si em um nó
4. Repita os passos 2. e 3. até obter o número desejado de conglomerados

Algoritmo

- O pesquisador deve definir que tipo de medida de similaridade utilizar para esta análise
- É preciso decidir o melhor tipo de medida de distância/similaridade entre os sujeitos (distância euclidiana, correlação etc.) e método de ligação
- Em geral, os dados são apresentados em um dendrograma

Número de Clusters

Número de Clusters

- Assim como no caso do k-means, precisamos determinar o número correto de clusters
- É uma tarefa difícil, que é melhor executada com conhecimentos *a priori* sobre os dados
- Vamos ver como utilizar a estatística gap para escolher o número de clusters

Número de Clusters

- Seja W_k a estatística gap para k clusters
- W_k parte do pressuposto de que, se as unidades amostrais não formassem grupos, ela seria aproximadamente uniforme para qualquer número de grupos
- Se existem grupos nos dados, então W_k forma um cotovelo no número ótimo de grupos

Número de Clusters

- W_k é dada por

$$W_k = \sum_{i=1}^{k} \frac{1}{2n_r} D_r,$$

em que $D_r = \sum_{i,j \in C_r} d_{ij}$ e

- k é o número de clusters
- n_r é o número de observações em cada cluster C_r
- d_{ij} é a distância entre as observações i e j

Exemplo

Exemplo

- Utilizaremos os dados sobre prisões nos Estados Unidos
- Esse conjunto de dados traz as seguintes taxas:
  - `Murder`: prisões por assassinato (a cada 100.000 habitantes)
  - `Assault`: prisões por assalto (a cada 100.000 habitantes)
  - `UrbanPop`: percentual de população urbana
  - `Rape`: prisões por estupro (a cada 100.000 habitantes)
- Os dados são em nível estadual, coletados em 1973
- Veremos como combinar distância, métodos de ligação e número de clusters nas nossas análises

Exemplo

```
> library(tidyverse)
> library(ggdendro)
> library(factoextra)
>
> prisoes <- scale(USArrests)
>
> prisoes.euclidiana <- dist(prisoes, method = "euclidean")
>
> prisoes.manhattan  <- dist(prisoes, method = "manhattan")
```

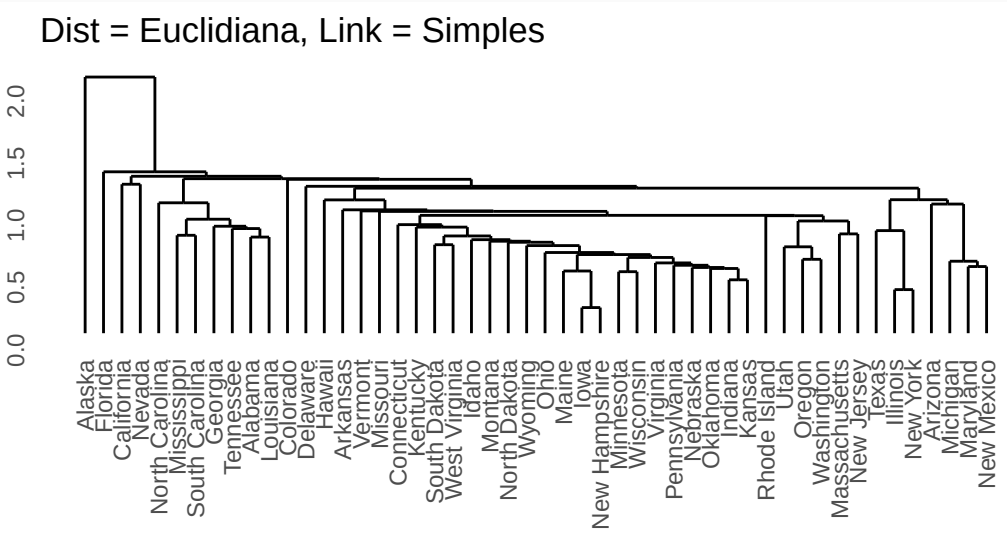
Exemplo

```
> ggdendrogram(hclust(prisoes.euclidiana, method = "single")) +
+   labs(title = "Dist = Euclidiana, Link = Simples")
>
> ggdendrogram(hclust(prisoes.euclidiana, method = "complete")) +
+   labs(title = "Dist = Euclidiana, Link = Completo")
>
> ggdendrogram(hclust(prisoes.euclidiana, method = "ward.D")) +
+   labs(title = "Dist = Euclidiana, Link = Ward")
```

Exemplo

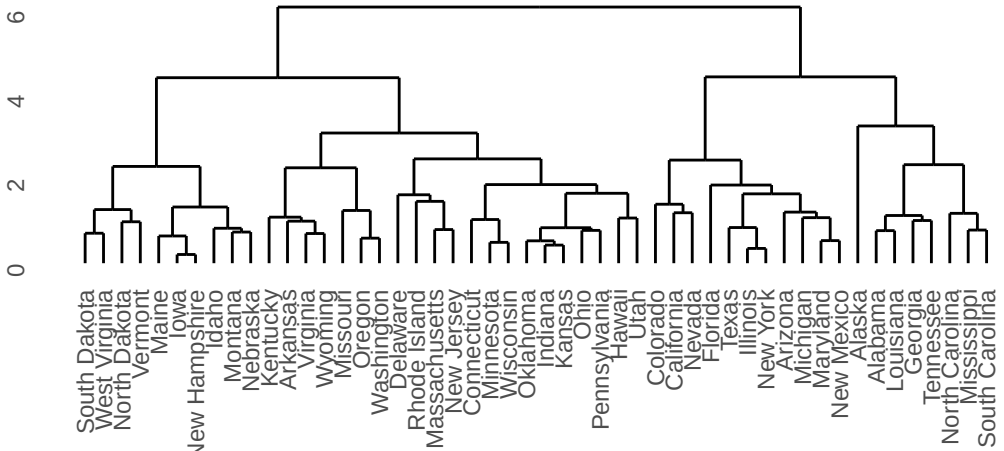
```
> ggdendrogram(hclust(prisoes.manhattan, method = "single")) +
+   labs(title = "Dist = Manhattan, Link = Simples")
>
> ggdendrogram(hclust(prisoes.manhattan, method = "complete")) +
+   labs(title = "Dist = Manhattan, Link = Completo")
>
> ggdendrogram(hclust(prisoes.manhattan, method = "ward.D")) +
+   labs(title = "Dist = Manhattan, Link = Ward")
```

Exemplo

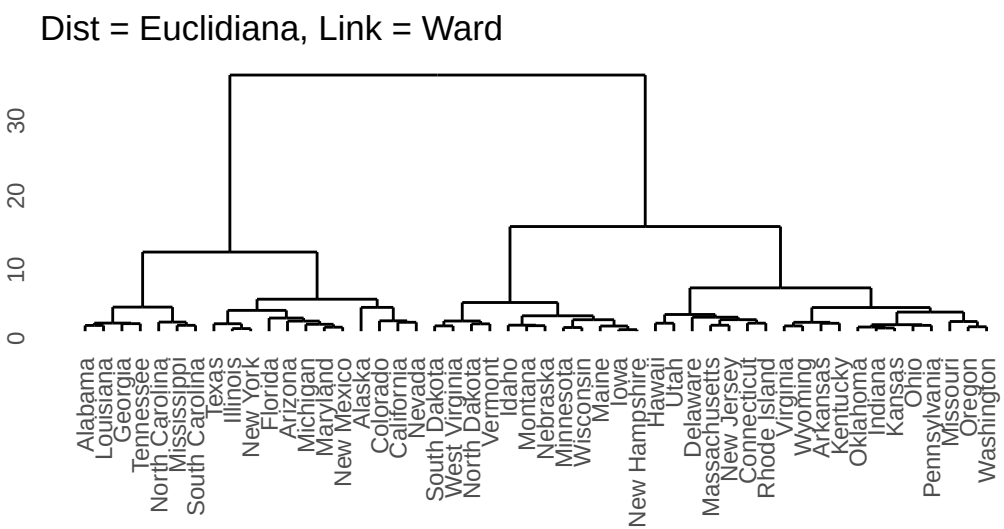

Dist = Euclidiana, Link = Simples
0.0
0.5
1.0
1.5
2.0
Alaska
Florida
California
Nevada
North Carolina
Mississippi
South Carolina
Georgia
Tennessee
Alabama
Louisiana
Colorado
Delaware
Hawaii
Arkansas
Vermont
Missouri
Connecticut
Kentucky
South Dakota
West Virginia
Idaho
Montana
North Dakota
Wyoming
Ohio
Maine
Iowa
New Hampshire
Minnesota
Wisconsin
Virginia
Pennsylvania
Nebraska
Oklahoma
Indiana
Kansas
Rhode Island
Utah
Oregon
Washington
Massachusetts
New Jersey
Texas
Illinois
New York
Arizona
Michigan
Maryland
New Mexico

Exemplo

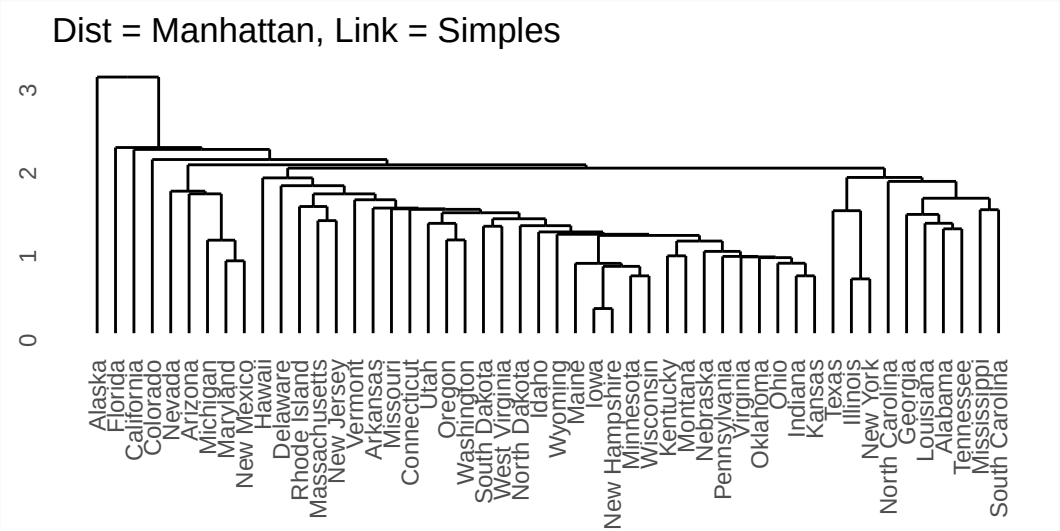
Dist = Euclidiana, Link = Completo



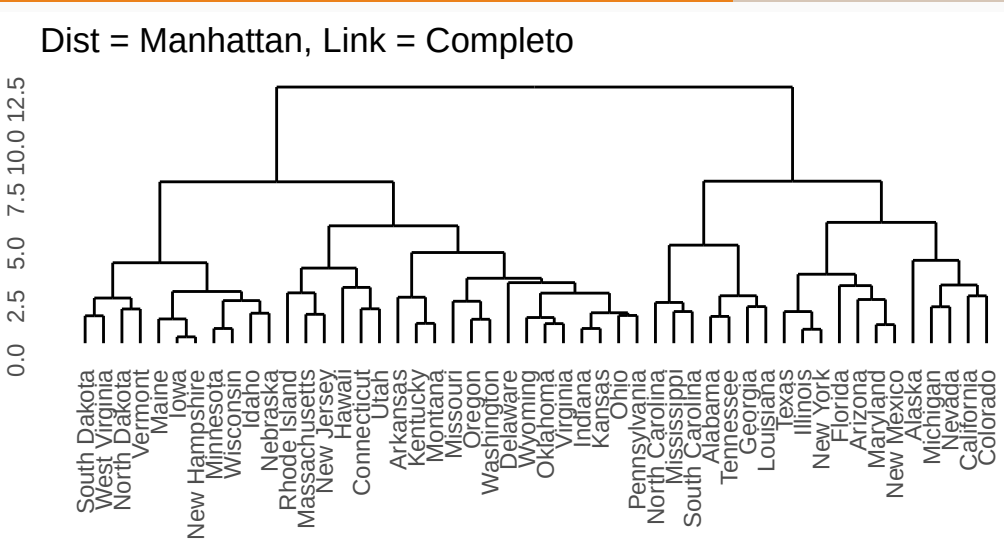
Exemplo


Dist = Euclidiana, Link = Ward
0
10
20
30
Alabama
Louisiana
Georgia
Tennessee
North Carolina
Mississippi
South Carolina
Texas
Illinois
New York
Florida
Arizona
Michigan
Maryland
New Mexico
Alaska
Colorado
California
Nevada
South Dakota
West Virginia
North Dakota
Vermont
Idaho
Montana
Nebraska
Minnesota
Wisconsin
Maine
Iowa
New Hampshire
Hawaii
Utah
Delaware
Massachusetts
New Jersey
Connecticut
Rhode Island
Virginia
Wyoming
Arkansas
Kentucky
Oklahoma
Indiana
Kansas
Ohio
Pennsylvania
Missouri
Oregon
Washington

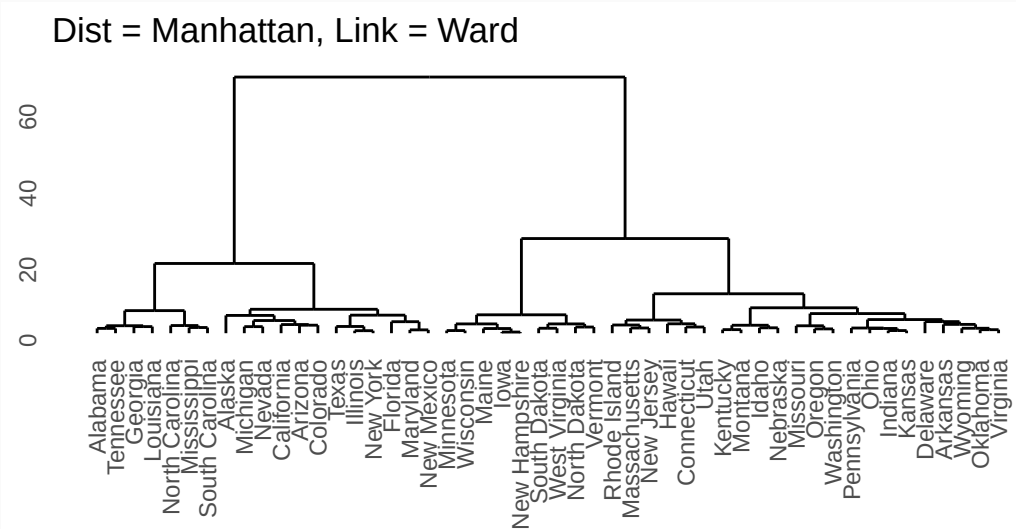
Exemplo


Dist = Manhattan, Link = Simples
0
1
2
3
Alaska
Florida
California
Colorado
Nevada
Arizona
Michigan
Maryland
New Mexico
Hawaii
Delaware
Rhode Island
Massachusetts
New Jersey
Vermont
Arkansas
Missouri
Connecticut
Utah
Oregon
Washington
South Dakota
West Virginia
North Dakota
Idaho
Wyoming
Maine
Iowa
New Hampshire
Minnesota
Wisconsin
Kentucky
Montana
Nebraska
Pennsylvania
Virginia
Oklahoma
Ohio
Indiana
Kansas
Texas
Illinois
New York
North Carolina
Georgia
Louisiana
Alabama
Tennessee
Mississippi
South Carolina

Exemplo


Dist = Manhattan, Link = Completo
0.0
2.5
5.0
7.5
10.0
12.5
South Dakota
West Virginia
North Dakota
Vermont
Maine
Iowa
New Hampshire
Minnesota
Wisconsin
Idaho
Nebraska
Rhode Island
Massachusetts
New Jersey
Hawaii
Connecticut
Utah
Arkansas
Kentucky
Montana
Missouri
Oregon
Washington
Delaware
Wyoming
Oklahoma
Virginia
Indiana
Kansas
Ohio
Pennsylvania
North Carolina
Mississippi
South Carolina
Alabama
Tennessee
Georgia
Louisiana
Texas
Illinois
New York
Florida
Arizona
Maryland
New Mexico
Alaska
Michigan
Nevada
California
Colorado

Exemplo


Dist = Manhattan, Link = Ward
0
20
40
60
Alabama
Tennessee
Georgia
Louisiana
North Carolina
Mississippi
South Carolina
Alaska
Michigan
Nevada
California
Arizona
Colorado
Texas
Illinois
New York
Florida
Maryland
New Mexico
Minnesota
Wisconsin
Maine
Iowa
New Hampshire
South Dakota
West Virginia
North Dakota
Vermont
Rhode Island
Massachusetts
New Jersey
Hawaii
Connecticut
Utah
Kentucky
Montana
Idaho
Nebraska
Missouri
Oregon
Washington
Pennsylvania
Ohio
Indiana
Kansas
Delaware
Arkansas
Wyoming
Oklahoma
Virginia

Exemplo

```
> euclidiana_single <- function(x, k) {
+   list(cluster = cutree(hclust(dist(x),
+                                method = "single"),
+                         k=k))
+   }
>
> fviz_nbclust(prisoes,
+              FUNcluster = euclidiana_single,
+              method = "gap_stat",
+              nboot = 100)
```

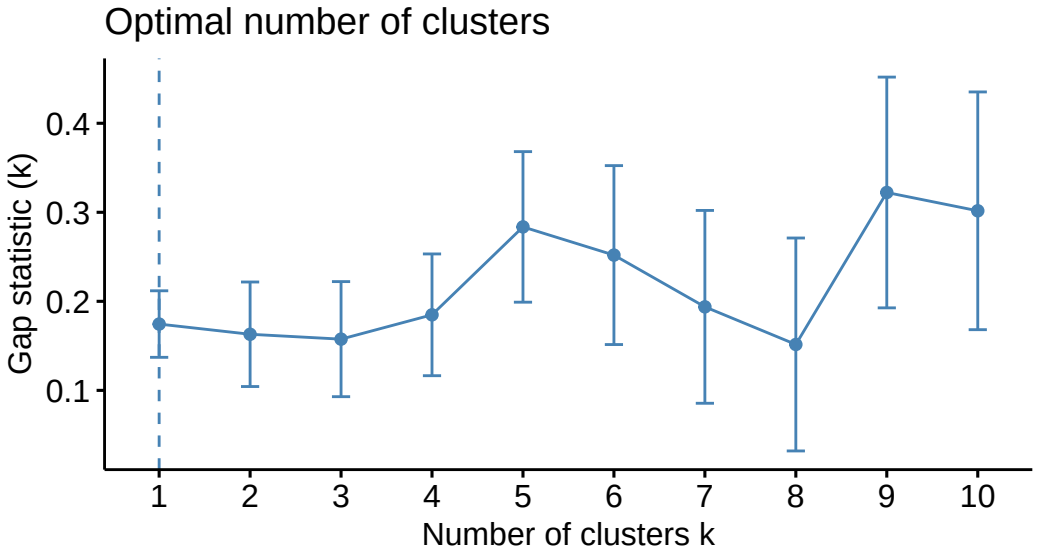
Exemplo

```
> euclidiana_complete <- function(x, k) {
+   list(cluster = cutree(hclust(dist(x),
+                                method = "complete"),
+                         k=k))
+ }
>
> fviz_nbclust(prisoes,
+              FUNcluster = euclidiana_complete,
+              method = "gap_stat",
+              nboot = 100)
```

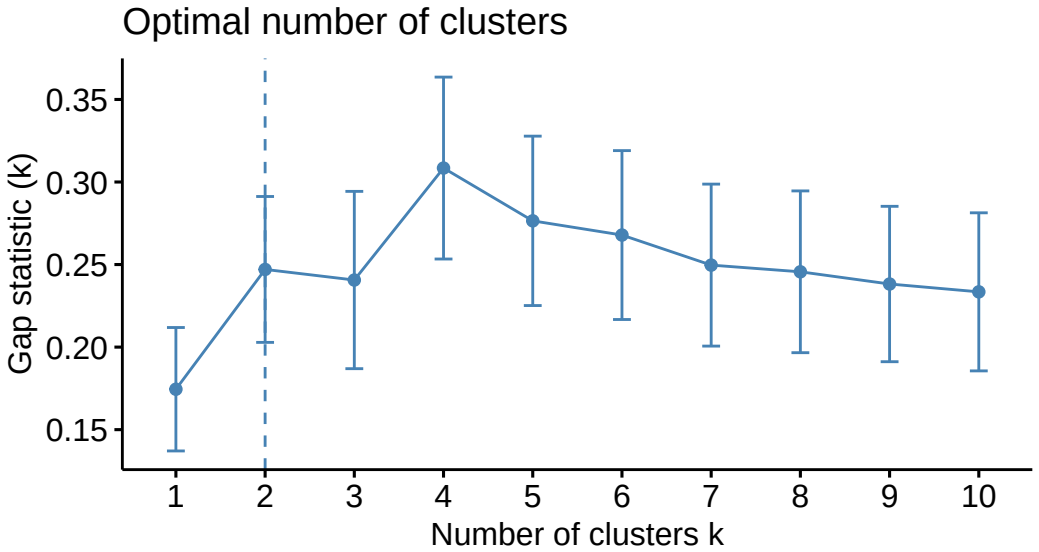
Exemplo

```
> euclidiana_ward <- function(x, k) {
+   list(cluster = cutree(hclust(dist(x),
+                                method = "ward.D"),
+                         k=k))
+ }
>
> fviz_nbclust(prisoes,
+              FUNcluster = euclidiana_ward,
+              method = "gap_stat",
+              nboot = 100)
```

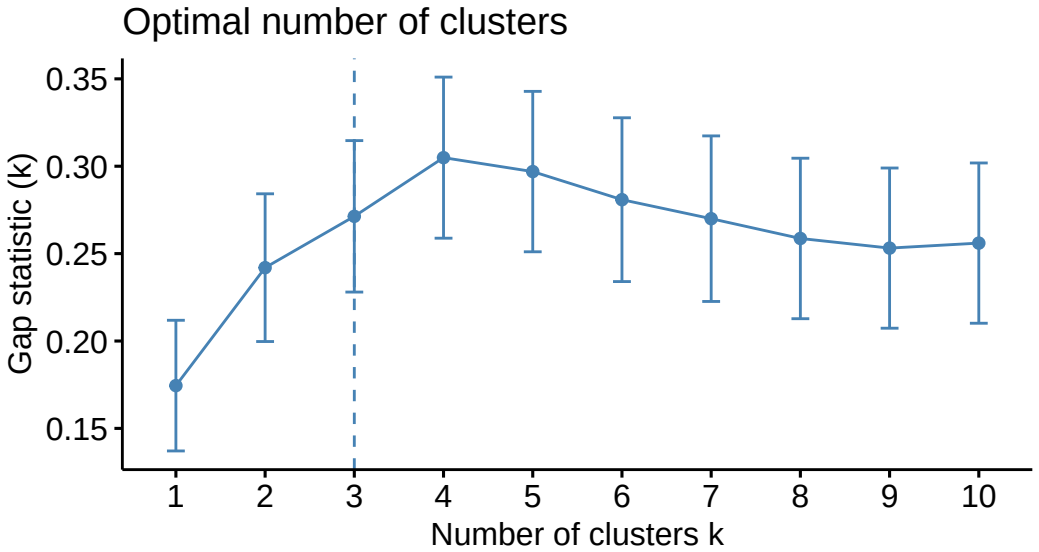
Exemplo


Optimal number of clusters
Gap statistic (k)
0.1
0.2
0.3
0.4
1
2
3
4
5
6
7
8
9
10
Number of clusters k

Exemplo


Optimal number of clusters
Gap statistic (k)
0.35
0.30
0.25
0.20
0.15
1
2
3
4
5
6
7
8
9
10
Number of clusters k

Exemplo


Optimal number of clusters
Gap statistic (k)
0.35
0.30
0.25
0.20
0.15
1
2
3
4
5
6
7
8
9
10
Number of clusters k

Exemplo

- O número correto de clusters vai ser definido de acordo com o interesse ou conhecimento do pesquisador
- A distância a ser utilizada depende dos dados
- Esta é outra ferramenta de análise exploratória

Exercícios

Exercícios

1. O arquivo `AlimentacaoReinoUnido.txt` mostra o consumo de diversos alimentos no Reino Unido em 1997. Importe este conjunto de dados para o R.
2. Faça a clusterização hierárquica deste conjunto de dados utilizando a distância euclidiana.
3. Faça a clusterização hierárquica deste conjunto de dados utilizando a distância de Manhattan.
4. Há algum padrão interessante nos dados? Qual ou quais?
5. Refaça a análise para o conjunto de dados com a transposta da matriz original de dados. O que mudou? O que é possível perceber com esta nova análise?

Exercícios

6. Repita os exercícios do slide anterior para o conjunto de dados `heptatlo`

Clusterização Hierárquica

EST0133 - Introdução à Modelagem de Big Data

Marcus Nunes
https://introbigdata.org/
https://marcusnunes.me/

Universidade Federal do Rio Grande do Norte