

## MAE 5905: Introdução à Ciência de Dados

Pedro A. Morettin

Instituto de Matemática e Estatística  
Universidade de São Paulo  
[pam@ime.usp.br](mailto:pam@ime.usp.br)  
<http://www.ime.usp.br/~pam>

### Aula 12

18 de maio de 2023

# Sumário

- 1 Análise de Agrupamentos
- 2 Estratégias de agrupamento
- 3 Algoritmos hierárquicos
- 4 Algoritmos de partição: K-médias
- 5 AA-Tópicos Adicionais
- 6 Outras distâncias

## Preliminares

- Na Análise de Agrupamentos (AA), o objetivo é agrupar “pontos” pertencentes a determinado espaço em “grupos” de acordo com alguma medida de distância, de modo que pontos num mesmo grupo tenham uma pequena distância entre eles.
- A AA é, às vezes, chamada de segmentação de dados. O espaço mencionado acima pode ser um espaço Euclidiano, eventualmente de dimensão grande, ou pode ser um espaço não Euclidiano, por exemplo quando queremos agrupar documentos segundo certos tópicos.
- O problema da AA insere-se naquilo que chamamos anteriormente de aprendizado não supervisionado, ou seja, temos apenas um conjunto de variáveis preditoras (inputs), não há uma variável resposta, e o objetivo é descrever associações e padrões entre essas variáveis.
- Para alcançar nosso objetivo, podemos usar várias técnicas de agrupamento, e nesta aula iremos discutir algumas delas.

## Preliminares

A AA é usada em muitas áreas e aplicações, como:

- i) segmentação de imagens como em fMRI (imagens por ressonância magnética funcional), em que se pretende particionar a imagem em áreas de interesse. Veja, por exemplo, Sato et al. (2007);
- ii) bioinformática, por exemplo na análise de expressão de genes gerados de *microarrays* ou sequenciamento de DNA ou proteínas. Veja Hastie et al. (2009) ou Fujita et al. (2007), por exemplo.
- iii) reconhecimento de padrões, de objetos e caracteres, por exemplo identificação de textos escritos a mão em linguística;
- iv) redução (ou compressão) de grandes conjuntos de dados, a fim de escolher grupos de dados de interesse.

## AA-Exemplo 1

**Exemplo 1.** Consideremos as medidas das variáveis altura ( $X_1$ , em cm), peso ( $X_2$ , em kg), idade ( $X_3$ , em anos) e sexo ( $X_4$ , M ou F) em 12 indivíduos dispostas na Tabela 1.

Tabela 1: Dados de 12 indivíduos

Ind.	Altura	Peso	Idade	Sexo	Alt.Padr.	Peso Padr.	Id. Padr.
A	180	75	30	M	0,53	0,72	0,38
B	170	70	28	F	-0,57	-0,13	-0,02
C	165	65	20	F	-1,12	-0,97	-1,61
D	175	72	25	M	-0,02	0,21	-0,61
E	190	78	28	M	1,63	1,23	-0,02
F	185	78	30	M	1,08	1,23	0,38
G	160	62	28	F	-1,67	-1,48	-0,02
H	170	65	19	F	-0,57	-0,97	-1,81
I	175	68	27	M	-0,02	-0,47	-0,22
J	180	78	35	M	0,53	1,23	1,38
K	185	74	35	M	1,08	0,55	1,38
L	167	64	32	F	-0,90	-1,14	0,78
$\mu$	175,17	70,75	28,08	—	0	0	0
$\sigma$	9,11	5,91	5,02	—	1	1	1

## AA-Exemplo 1

- Nessa tabela, a média de cada variável é indicada por  $\mu$  e o desvio padrão por  $\sigma$ . As colunas 6, 7 e 8 trazem as variáveis padronizadas, ou seja,

$$Z_i = \frac{X_i - \mu_{X_i}}{\sigma_i}, \quad (1)$$

para  $i=1,2,3$

- Para a variável  $X_4$  (sexo), esta padronização, é claro, não faz sentido. A Figura 1 apresenta um gráfico de dispersão de  $X_3$  versus  $X_2$ .

## AA-Exemplo 1

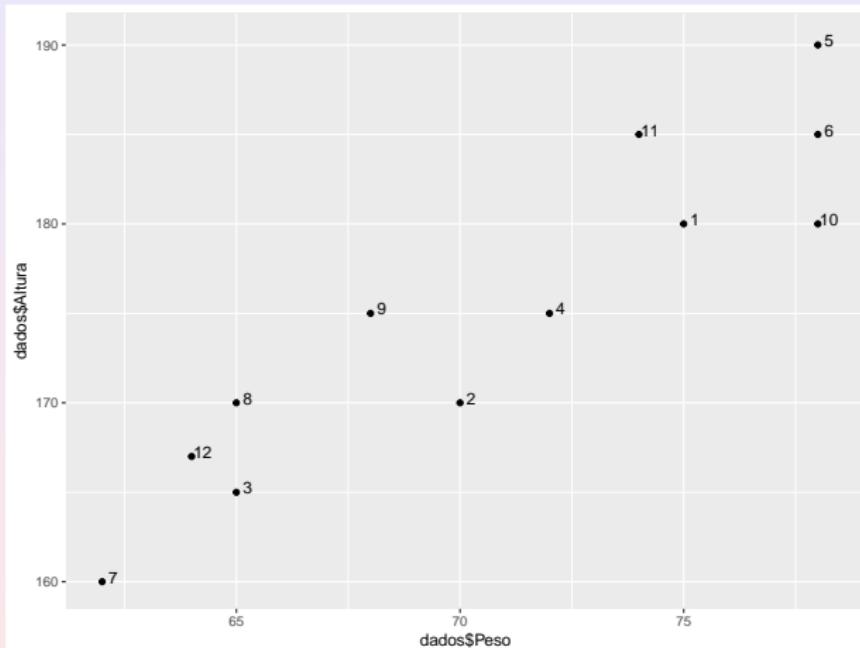


Figura 1: Gráfico de Altura *versus* Peso para os dados da Tabela 1, A=1, B=2 etc.

## AA-Exemplo 1

- O objetivo é agrupar pontos que estejam próximos. Para isso, definamos a **distância Euclidiana** entre dois pontos  $\mathbf{x} = (x_1, x_2)$  e  $\mathbf{y} = (y_1, y_2)$  como

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}.$$

- Nesse caso, a dimensão do espaço é dois. No caso geral de um espaço Euclidiano  $p$ -dimensional a distância é definida por

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + \dots + (x_p - y_p)^2}.$$

- Além da distância Euclidiana, podemos usar outras distâncias, como

$$d_1 = |x_1 - y_1| + \dots + |x_p - y_p| : \text{distância } L_1 \text{ ou Manhattan}, \quad (2)$$

$$d_2 = \max_{1 \leq i \leq p} |x_i - y_i| : \text{distância } L_\infty. \quad (3)$$

- Para espaços não Euclidianos, há outras definições de distância, como **Hamming, cosseno, Jaccard, edit** etc. (Ver Notas de Capítulo).

## AA–Exemplo 1

- Podemos obter a **matriz de similaridade** dos dados, segundo dada distância.
- Consideremos  $n$  indivíduos, para os quais observamos  $p$  variáveis. A matriz de similaridade será uma matriz  $n \times n$ ,  $D = [d(i,j)]_{i,j=1}^n$ , simétrica, com zeros na diagonal principal (Pode-se considerar uma matriz reduzida, de ordem  $(n - 1) \times n$ , para evitar os zeros).
- Na Tabela 2 temos essa matriz para os dados da Tabela 5, usando a distância Euclidiana. Algumas distâncias em ordem crescente são:

$$\begin{aligned}d(C, L) &= 2,24, \\d(A, J) &= 3,00, \\d(H, L) &= 3,16, \\d(D, I) &= 4,00, \\d(F, K) &= 4,00, \\d(B, H) &= 5,00, \\d(C, H) &= 5,00, \\d(E, F) &= 5,00, \\d(F, J) &= 5,00, \\d(A, K) &= 5,10.\end{aligned}$$

## AA-Exemplo 1

Essas distâncias (indicadas em negrito) nos dão uma ideia de como agrupar pontos.

Tabela 2: Matriz de similaridade, distância Euclidiana

	A	B	C	D	E	F	G	H	I	J	K
B	11,18										
C	18,03	7,07									
D	5,83	5,38	12,21								
E	10,44	21,54	28,18	16,16							
F	5,83	17,00	23,85	11,66	5,00						
G	23,85	12,81	5,83	18,03	34,00	29,68					
H	14,14	5,00	5,00	8,60	23,85	19,85	10,44				
I	8,60	5,39	10,44	4,00	18,03	14,14	16,16	5,83			
J	3,00	12,81	19,85	7,81	10,00	5,00	25,61	16,40	11,18		
K	5,10	15,52	21,93	10,20	6,40	4,00	27,73	17,49	11,66	6,40	
L	17,03	6,71	2,24	11,31	26, 93	22,80	7,28	3,16	8,94	19,10	20,59

## Tipos de algoritmos

Podemos dividir os algoritmos de agrupamento em três grupos (Hastie et al., 2009):

- a) **combinatórios**: trabalham diretamente com os dados, não havendo referência à sua distribuição;
- b) **baseados em modelos**: supõem que os dados sejam uma amostra aleatória simples de uma população, com uma densidade de probabilidade, que é uma mistura de densidades componentes, cada uma descrevendo um grupo;
- c) **bump hunters** : tratam de estimar, de modo não paramétrico, as modas da densidade. As observações mais próximas a cada moda definem os grupos.

## Tipos de algoritmos

Por sua vez, os algoritmos combinatórios podem ser classificados em dois grupos:

- i) **algoritmos hierárquicos**: que ainda podem ser subdivididos em **aglomerativos** e **divisivos**. No primeiro caso, iniciamos o procedimento de modo que cada ponto forma um grupo e vamos combinando grupos com base em suas proximidades, usando alguma definição de proximidade (como uma distância). Paramos quando fixamos um número de grupos ou obtemos grupos que não são desejáveis, por alguma razão. No segundo caso, partimos de um único grupo e por divisões sucessivas obtemos 1,2 etc. grupos. Neste texto, usaremos o método aglomerativo.
- ii) **algoritmos de partição** ( ou obtidos por associação de pontos): os grupos obtidos formam uma partição do conjunto total de pontos. Os pontos são considerados em alguma ordem e cada um deles é associado ao grupo no qual ele melhor se ajusta. O método chamado de **K-médias** pertence a esse grupo de algoritmos.

## Tipos de algoritmos

- Se estivermos num espaço Euclíadiano, quando usamos alguma distância entre os pontos, grupos podem ser caracterizados pelo seu **centróide** (média das coordenadas dos pontos).
- Num espaço não Euclíadiano, não existe a noção de centróide, e deveremos usar alguma outra maneira de caracterizar os grupos.
- Se o conjunto de pontos for muito grande, há o que se chama de **maldição da dimensionalidade** (*curse of dimensionality*). Em grandes dimensões, **quase** todos os pares de pontos têm a mesma distância entre si e quaisquer dois vetores são **quase** sempre ortogonais.

## AH–Exemplo 1

- Para ilustrar o método, vamos voltar ao Exemplo 1 e procuremos os pontos mais próximos, baseando-nos na Figura 1.
- Consideremos as variáveis  $X_1$  (altura) e  $X_2$  (peso) e o gráfico da Figura 1. Temos um espaço Euclidiano de dimensão 2 e a proximidade entre dois pontos medida pela distância Euclidiana (DE).
- Cada grupo será representado pelo seu centróide e a regra de agrupamento será, pois: calcular a distância Euclidiana entre os centróides de dois grupos quaisquer e escolher, para agrupar, os dois grupos com a menor DE.

## AH-Exemplo 1

- i) Inicialmente, cada ponto é considerado um grupo, que coincide com seu centróide.
- ii) Consultando a Tabela 2,  $C=(65, 165)$  e  $L=(64; 167, 64)$  são os mais próximos com DE  $d(C, L) = \sqrt{5,00} = 2,24$  e centróide  $c(C, L) = (64, 5; 166)$ . Esses dois pontos formam o primeiro agrupamento,  $\mathcal{G}_1 = \{C, L\}$ . A distância entre esses pontos, 2,24, será chamada **nível do agrupamento ou junção**.
- iii) A seguir, teríamos que recalcular as distâncias entre os pontos, considerando agora os grupos A, B, CL, D, E, F, G, H, I, J, K. A maioria das distâncias não muda, mudando somente as distâncias dos pontos ao centróide de C e L. Pela Tabela 2, a DE de  $A = (75, 180)$  a  $J = (78, 180)$  é 3, logo agrupamos A e J, formando o grupo  $\mathcal{G}_2 = \{A, J\}$ , com centróide  $c(\mathcal{G}_2) = (76, 5, 180)$  e nível 3,00.
- iv) Agora, DE entre  $D = (72, 175)$  e  $I = (68, 175)$  é 4,00, obtendo-se o agrupamento,  $\mathcal{G}_3 = \{D, I\}$ , com centróide  $c(\mathcal{G}_3) = (70, 175)$  e nível 4,00.

## AH–Exemplo 1

- v) A seguir, os pontos mais próximos são  $F = (78, 185)$  e  $K = (74, 185)$ , com DE igual a 4,00, obtendo-se  $\mathcal{G}_4 = \{F, K\}$ , centróide  $c(\mathcal{G}_4) = (76, 185)$  e nível 4,00.
- vi) Consideramos o ponto  $H = (65, 170)$ , cuja DE ao centróide de C e L é  $d(H, c(\mathcal{G}_1)) \approx 4,03$ . Portanto, obtemos outro agrupamento,  $\mathcal{G}_5 = \{C, H, L\}$ , com centróide  $c(\mathcal{G}_5) = (64, 7; 167, 3)$  e nível 4,03.
- vii) A seguir, agrupamos  $B = (70, 170)$  com  $\mathcal{G}_3 = \{D, I\}$ , notando que a DE entre B e o centróide desse grupo é 5,00, passando a ser esse valor o nível da junção. O novo grupo é  $\mathcal{G}_6 = \{B, D, I\}$ , com centróide  $c(\mathcal{G}_6) = (70; 173, 3)$ .
- viii) Agrupamos, agora, os grupos  $\mathcal{G}_2 = \{A, J\}$  e  $\mathcal{G}_4 = \{F, K\}$ , com DE entre seus centróides de 5,02, formando-se o grupo  $\mathcal{G}_7 = \{A, F, J, K\}$ , com centróide  $c(\mathcal{G}_7) = (76, 25; 182, 5)$  e nível 5,02.

## AH–Exemplo 1

- ix) Agregamos, a seguir, o ponto  $G = (62, 160)$  ao grupo  $\mathcal{G}_5 = \{C, H, L\}$ , obtendo-se o novo grupo  $\mathcal{G}_8 = \{C, G, H, L\}$ , com centróide  $c(\mathcal{G}_8) = (64; 165, 5)$  e nível 7,31.
- x) Finalmente, agrupamos o ponto  $E = (78, 190)$  ao grupo  $\mathcal{G}_7 = \{A, F, J, K\}$ , obtendo-se o grupo  $\mathcal{G}_9 = \{A, E, F, J, K\}$ , com centróide  $c(\mathcal{G}_9) = (76, 6; 184)$  e nível 7,50.

A Figura 2 mostra esses agrupamentos. Podemos prosseguir, agrupando-se dois desses grupos (aqueles que possuem a menor DE entre os respectivos centróides) e, finalmente, agrupar os dois grupos restantes.

## AH-Exemplo 1

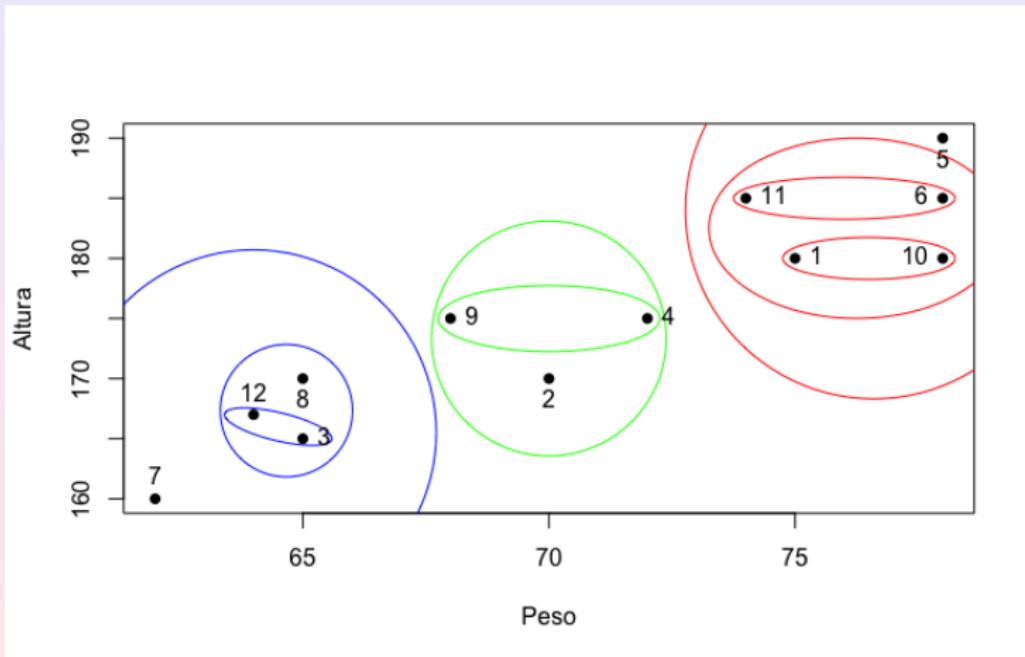


Figura 2: Agrupamentos obtidos para o Exemplo 1.

## AH-Dendrograma

- Um gráfico que sumariza o procedimento ([dendrograma](#)) está na Figura 3. No eixo vertical da figura colocamos os níveis, no horizontal, os pontos de modo conveniente. Nessa figura, usamos a distância Euclidiana.
- Na Tabela 3 temos um resumo do método hierárquico, usando distância Euclidiana e centróides, para agrupar os pontos, para o Exemplo 1.

Tabela 3: Resumo do procedimento de agrupamento para o Exemplo 1

Passo	Agrupamento	Nível
1	C, L	2,24
2	A, J	3,00
3	D, I	4,00
4	F, K	4,00
5	H, CL	4,03
6	B, DI	5,00
7	AJ, KF	5,02
8	G, CHL	7,31
9	E, AFJK	7,50

## AH-Dendrograma

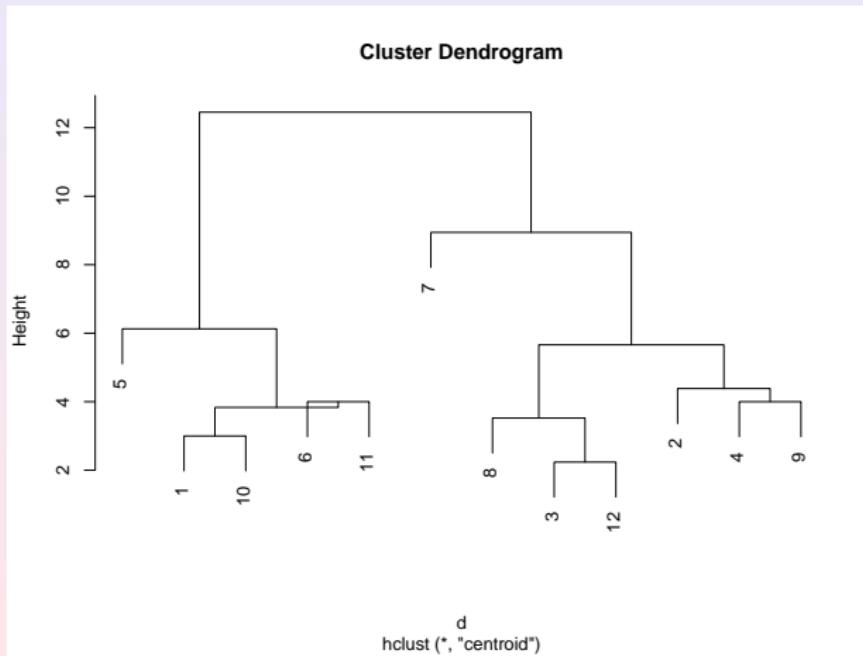


Figura 3: Dendrograma para o Exemplo 1.

## AH–Interpretação

- Interpretando o resultado, com esses três grupos, vemos que o primeiro contém as pessoas menos pesadas e mais baixas (4 pessoas), depois aquele que contém pessoas com pesos e alturas intermediárias (3 pessoas) e, finalmente, o grupo que contém as pessoas mais pesadas e mais altas (5 pessoas).
- Se esse for objetivo, podemos parar aqui. Se o objetivo é obter dois grupos, um com pessoas mais baixas e menos pesadas e, outro, com pessoas mais altas e mais pesadas, continuamos a agrupar mais uma vez, obtendo os grupos  $\mathcal{G}_{10} = \{B, C, D, G, H, I, L\}$  e  $\mathcal{G}_9$ .
- Um dos objetivos da construção de grupos é **classificar** um novo indivíduo em algum dos grupos obtidos. O problema da classificação está intimamente ligado ao problema de AA, e já foi tratado em capítulos anteriores.
- Um pacote do repositório R que pode ser usado é o **cluster**. Após carregar o pacote em sua área por meio de `library(cluster)`, temos que informar a distância (`euclidian`, `maximum`, `manhattan` etc.) a usar e o método de agrupamento (`centroid`, `average`, `median` etc.). O pacote contém várias funções para mostrar em que grupo estão as unidades, obter o dendrograma, fornecer a ordem para fazer o dendrograma etc.

## K-médias

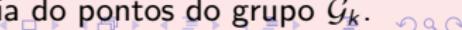
- O método de K-médias tem por objetivo partitionar os pontos em  $K$  grupos, de tal modo que a soma dos quadrados das distâncias dos pontos aos centros dos agrupamentos (**clusters**) seja minimizada. É um método baseado em centróides, como vimos anteriormente, e pertence à classe de algoritmos (ii) discutida antes, e requer que o espaço seja Euclidiano.
- Usualmente, o valor de  $K$  é conhecido e deve ser fornecido pelo usuário mas é possível obtê-lo por tentativa e erro.

O algoritmo mais comum é devido a Hartigan and Wong (1979) é usado como *default* em pacotes computacionais. Outros algoritmos são os de MacQueen (1967), Lloyd (1957) eForgy (1965).

- A função **kmeans** do pacote **cluster** pode ser utilizada.
- A ideia básica consiste em definir grupos em que a variação interna seja minimizada. Esta é, em geral, definida como a soma das DE ao quadrado entre pontos e o centróide correspondente:

$$W(\mathcal{G}_k) = \sum_{x_i \in \mathcal{G}_k} (x_i - \mu_k)^2, \quad (4)$$

em que  $x_i$  é um ponto no grupo  $\mathcal{G}_k$  e  $\mu_k$  é a média do pontos do grupo  $\mathcal{G}_k$ .



## K-médias

O algoritmo consiste nos seguintes passos:

- i) especifique K e selecione K pontos que pareçam estar em diferentes grupos;
- ii) considere esses pontos como os centróides iniciais desses grupos;
- iii) associe cada observação ao centróide mais próximo, baseado na distância Euclidiana entre essa e o centróide;
- iv) para cada um dos K grupos, recalcule o centróide após cada ponto ser incluído.
- v) iterativamente, minimize a soma total de quadrados dentro dos grupos, até que os centróides não mudem muito (o R usa 10 iterações como *default*). A soma total de quadrados dentro dos grupos é definida por

$$\sum_{j=1}^K W(\mathcal{G}_j) = \sum_{j=1}^K \sum_{x_i \in \mathcal{G}_j} (x_i - \mu_j)^2. \quad (5)$$

## K-médias: Exemplo 2

**Exemplo 2.** Consideremos 100 simulações de duas variáveis com distribuição normal, uma com média 0 e desvio padrão 0,3 e outra, com média 1 e desvio padrão também 0,3. Na Figura 4 apresentamos os dois grupos com os respectivos centros, resultantes da aplicação do algoritmo **K-means** com  $K = 2$ .

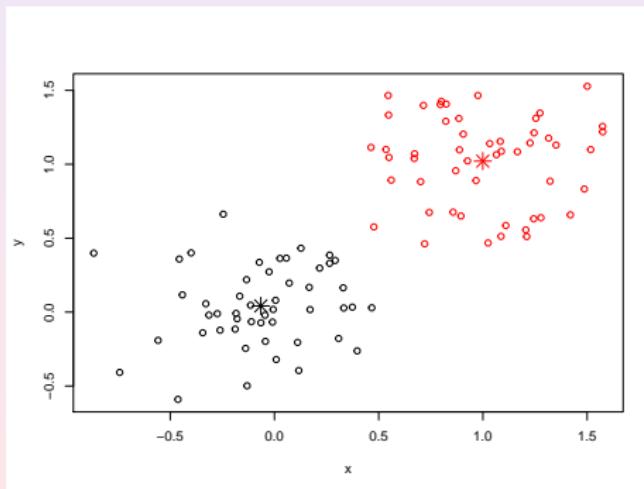


Figura 4: Uso do pacote `kmeans` para o exemplo simulado

## K-médias: Exemplo 3

- **Exemplo 3.**: Consideremos os dados da Tabela 1 e as variáveis Peso e Altura. Usando o resultado do procedimento hierárquico, suponha que tenhamos  $K = 3$  grupos.
- Usando a função `kmeans`, obtemos o gráfico da Figura 5. Nessa figura temos os três grupos (com tamanhos 3,4 e 5) em diferentes cores e os centros de cada grupo.
- Esses centróides são dados pelo programa,  $(63, 67; 164)$ ,  $(68, 75; 172, 5)$  e  $(76, 60; 184, 0)$ . As somas de quadrados dentro dos grupos são 30,67, 51,75 e 85,20, respectivamente. A soma de quadrados total entre grupos é 87,1.

## K-médias: Exemplo 3

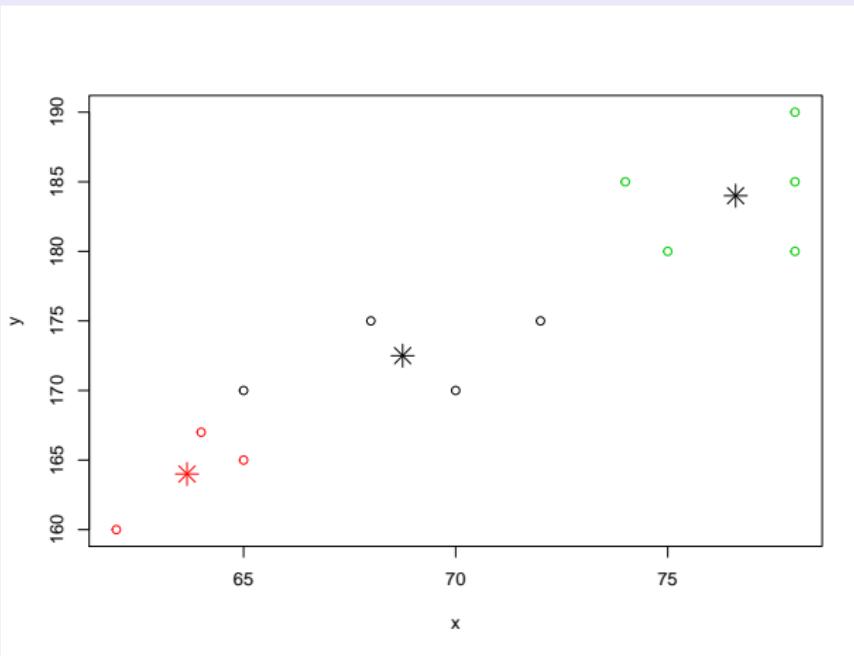


Figura 5: Uso do pacote `kmeans` para o Exemplo 3.

## K-médias: Exemplo 4

- **Exemplo 4:** Vamos considerar os dados de um conjunto de 4.000 motoristas encarregados de fazer entregas de determinados produtos. Há várias variáveis envolvidas, mas iremos considerar somente duas, nomeadamente,  $X_1$ : distância média percorrida por cada motorista (em milhas) e  $X_2$ : porcentagem média do tempo em que o motorista esteve acima do limite de velocidade por mais de 5 milhas por hora. Há dados do setor urbano e rural.
- Os dados podem ser obtidos de  
[https://raw.githubusercontent.com/datasets/introduction-to-k-means-Clustering/master/Data/data\\_1024.csv](https://raw.githubusercontent.com/datasets/introduction-to-k-means-Clustering/master/Data/data_1024.csv).
- Na Figura 6 apresentamos um diagrama de dispersão dos dados, segundo essas duas variáveis, mostrando claramente dois grupos distintos: grupo 1, contendo os motoristas que fazem entregas no setor urbano e grupo 2, contendo motoristas que fazem entregas no setor rural.

## K-médias: Exemplo 4

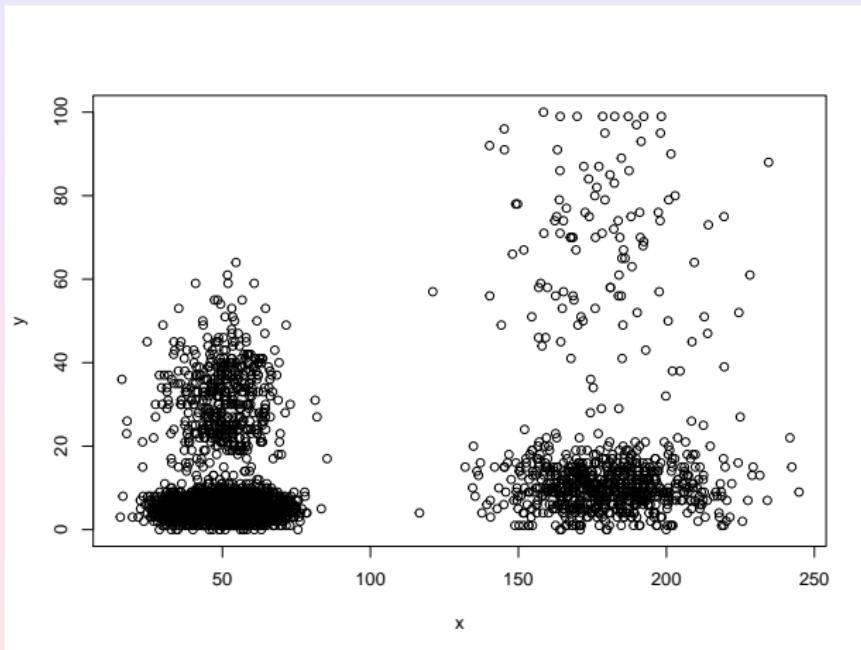


Figura 6: Gráfico de dispersão de  $X_1$  versus  $X_2$  para o Exemplo 4.

## K-médias: Exemplo 4

Utilizando o algoritmo K-médias com  $K = 2$ , obtemos:

Grupo 1: Centróide = (50,05, 8,83)

Grupo 2: Centróide = (180,02, 18,29)

Na Figura 7 temos os dois grupos representados.

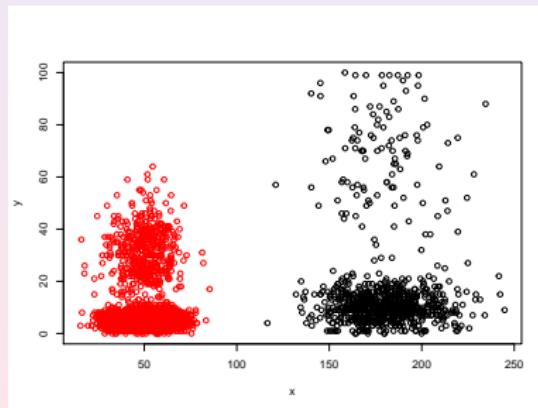


Figura 7: Grupos para o Exemplo 4 com  $K=2$ .

## K-médias: Exemplo 4

Se tomarmos  $K = 4$ , obtemos o gráfico da Figura 8. Agora, os motoristas são separados por aqueles que seguem ou não a velocidade limite, além da divisão zona urbana/rural.

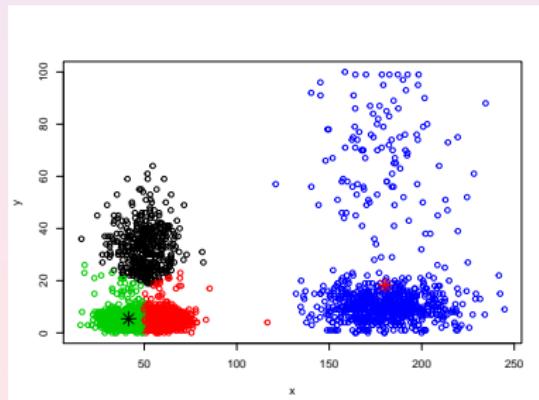
Os centróides são:

Grupo 1: (50,61, 33,06),

Grupo 2: (57,90, 5,28),

Grupo 3: (41,52, 5,40),

Grupo 4: (180,10, 18,31).



## K-médias: Exemplo 5

- **Exemplo 5:** Vamos, agora, considerar dados do Uber, na cidade de Nova Iorque (NYC). Esses dados podem ser obtidos no site

[www.kaggle.com/fivethirtyeight/  
uber-pickups-in-new-york-city/  
data](http://www.kaggle.com/fivethirtyeight/uber-pickups-in-new-york-city/data)

e contém cerca de 4,5 milhões de corridas do Uber de abril a setembro de 2014, além de outros dados do Uber de 2015 e outras de companhias.

- NYC tem 5 distritos: Brooklyn, Queens, Manhattan, Bronx e Staten Island. O conjunto de dados que vamos usar, de 2014, tem informação detalhada sobre a localização do início da corrida com as seguintes colunas:  
Date/Time: dia e hora do início da corrida;  
Lat: a latitude da localidade;  
Lon: a longitude da localidade;  
Base: o código da base da companhia afiliada àquela corrida.
- Os nomes dos arquivos são da forma `uber-raw-data-month.csv`, em que **month** deve ser substituído por `apr14`, `aug14`, `jul14`, `jun14`, `may14`, `sept14`. Em nosso exemplo, vamos usar somente os dados de abril de 2014.

## K-médias: Exemplo 5

- Para ler os dados usamos o comando:

```
> read.csv("https://raw.githubusercontent.com/fivethirtyeight/  
uber-tlc-foil-response/master/uber-trip-data/uber-raw-data-apr14.csv")
```

- Iremos usar o pacote **kmeans** do R e, dependendo do que se quer, outros pacotes, como **dplyr**, **VIM**, **lubridate** ou **ggmap**, poderão ser empregados.
- Com o comando **summary(apr14)**, obtemos:

```
summary(apr14)
      Date.Time           Lat             Lon
4/7/2014 20:21:00 :   97   Min.   :40.07   Min.   :-74.77
4/7/2014 20:22:00 :   87   1st Qu.:40.72   1st Qu.:-74.00
4/30/2014 17:45:00:   78   Median :40.74   Median :-73.98
4/30/2014 18:43:00:   70   Mean    :40.74   Mean   :-73.98
4/30/2014 19:00:00:   70   3rd Qu.:40.76   3rd Qu.:-73.97
4/30/2014 16:55:00:   67   Max.    :42.12   Max.   :-72.07
(Other)                  :564047
```

## K-médias: Exemplo 5

Para ver os dados correspondentes (até dia), temos:

```
head(apr14, n=10)
  Date.Time      Lat      Lon     Base Year Month Day
1 2014-04-01 00:11:00 40.7690 -73.9549 B02512 2014     4    1
2 2014-04-01 00:17:00 40.7267 -74.0345 B02512 2014     4    1
3 2014-04-01 00:21:00 40.7316 -73.9873 B02512 2014     4    1
4 2014-04-01 00:28:00 40.7588 -73.9776 B02512 2014     4    1
5 2014-04-01 00:33:00 40.7594 -73.9722 B02512 2014     4    1
6 2014-04-01 00:33:00 40.7383 -74.0403 B02512 2014     4    1
7 2014-04-01 00:39:00 40.7223 -73.9887 B02512 2014     4    1
8 2014-04-01 00:45:00 40.7620 -73.9790 B02512 2014     4    1
9 2014-04-01 00:55:00 40.7524 -73.9960 B02512 2014     4    1
10 2014-04-01 01:01:00 40.7575 -73.9846 B02512 2014     4    1
```

## K-médias: Exemplo 5

Finalmente, usando o `kmeans` e `ggmap` obtemos a Figura 34. Detalhes dos scripts necessários, estão na página do livro.

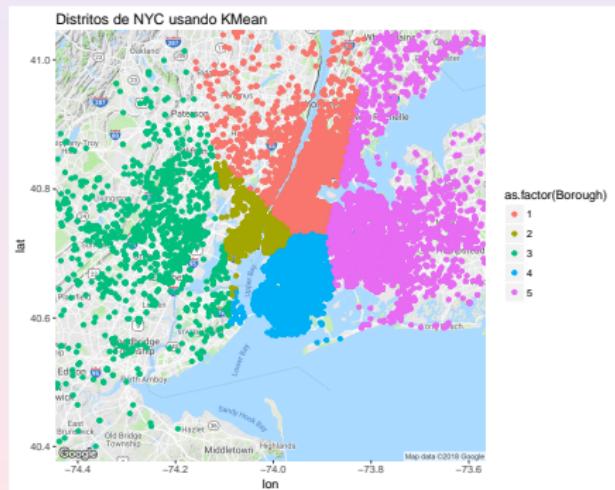


Figura 9: Grupos obtidos para o Exemplo 5.

## Matriz Cofenética

- Esta matriz, **S**, digamos, contém as distâncias entre os objetos a partir do dendograma. Por exemplo, a distância entre os pontos C e L é dada pelo nível em que os dois foram agrupados, nesse caso, 2,24.
- A seguir, verificamos a proximidade entre essa matriz e a matriz de proximidade, **D**.
- Essa proximidade é dada pelo coeficiente de correlação entre os valores de **D** e de **S**, chamado de **coeficiente de correlação cofenético**.
- No Exemplo 12.1, esse valor é 0,80, e pode ser considerado um valor adequado.

## AA–Outros algoritmos

Vimos exemplos com duas variáveis. Podemos ter mais variáveis, mas seu número deve ser menor que o número de indivíduos. Quando temos mais do que duas dimensões, alguns algoritmos foram propostos e são, basicamente, variantes de algoritmos hierárquicos e de *K*-means.

- Algoritmo BFR** [(Bradley, Fayyad and Reina, (1998)]. Esse algoritmo é uma variante do *K*-means, para o caso de um espaço Euclidiano de alta dimensão mas é baseado numa suposição muito forte: a forma dos agrupamentos segue uma distribuição normal, em cada dimensão, ao redor do respectivo centróide, e além disso as dimensões são independentes. Então, a forma dos grupos deve ser uma elipsóide, com os eixos paralelos aos eixos da dimensão, podendo eventualmente ser um círculo, mas não podem, por exemplo, terem os eixos do elipsóide oblíquos aos eixos da dimensão ou terem formas mais complicadas. Esse algoritmo não está contemplado no R.
- Algoritmo CURE** (de *Clustering Using Representatives*). Esse algoritmo usa procedimentos hierárquicos e os grupos podem ter quaisquer formas mas também não está disponível no R. No entanto, ele pode ser obtido no pacote **pyclustering**, que usa as linguagens Python e C++.

## AA–Outros algoritmos

- c) **Density-based algorithms** (DBSCAN): dado um conjunto de pontos em algum espaço, esse algoritmo agrupa pontos que estão dispostos em uma vizinhança com maior densidade, colocando como *outliers* os pontos que estão em regiões de baixa densidade. Veja Ester et al. (1996).

## AA-distância para strings

- Como salientamos anteriormente, para espaços não Euclidianos (ENE) há outras formas de distância.
- No caso de sequências (**strings**)  $x = x_1x_2 \cdots x_n$  e  $y = y_1y_2 \cdots y_n$ , uma distância conveniente é a distância **edit**, que dá o número de inserções e exclusões de caracteres que converterão  $x$  em  $y$ .
- Por exemplo, considere  $x = abcd$  e  $y = acde$ . A distância **edit** entre essas sequências é  $d_e(x, y) = 2$ , pois temos que excluir  $b$  e inserir  $e$  depois do  $d$ .
- Uma outra maneira de obter essa distância é considerar uma subsequência mais longa comum (SML) a  $x$  e  $y$ . No exemplo, é  $acd$ . Então, a distância **edit** é dada por

$$d_e(x, y) = \ell(x) + \ell(y) - 2\ell(\text{SML}),$$

em que  $\ell$  indica o comprimento de cada sequência. No exemplo,  
 $d_e(x, y) = 4 + 4 - 2 \times 3 = 2$ .

## AA-distância para strings

- Outra distância que pode ser usada é a **distância Hamming**, que dá o número de componentes em que dois vetores (de mesma dimensão) diferem. Por exemplo, se  $x = 110010$  e  $y = 100101$ , então a distância Hamming entre eles é  $d_H(x, y) = 4$ .
- Um problema em ENE é representar um grupo, pois não podemos, por exemplo, calcular o centróide de dois pontos. No exemplo acima, como obter uma sequência entre  $x$  e  $y$ ? Usando a distância *edit* poder-se-ia selecionar algo parecido ao centróide (**o grupóide**), escolhendo-se a sequência que minimiza, por exemplo, a soma das distâncias dessa com as outras sequências do grupo previamente selecionado.

## AA-algoritmos aglomerativos

Considere os algoritmos hierárquicos e dois grupos quaisquer,  $A$  e  $B$ , e a distância entre eles,  $d(A, B)$ . Como vimos, comumente usamos algoritmos aglomerativos que ainda podem ser divididos em (Hastie et al., 2009):

- a) algoritmos com ligação simples (**single linkage**), para os quais toma-se a distância mínima entre os pares, ou seja,

$$d_{SL}(A, B) = \min_{i \in A, j \in B} d(i, j). \quad (6)$$

Essa técnica é também conhecida como técnica do **vizinho mais próximo**;

- b) algoritmos com ligação completa (**complete linkage**), para os quais toma-se a máxima distância entre os pares:

$$d_{CL}(A, B) = \max_{i \in A, j \in B} d(i, j); \quad (7)$$

## AA-algoritmos aglomerativos

- c) algoritmos com ligação média (**group average**), que toma a distância média entre os grupos:

$$d_{GA}(A, B) = \frac{1}{N_A N_B} \sum_{i \in A} \sum_{j \in B} d(i, j). \quad (8)$$

Aqui,  $N_A$  e  $N_B$  indicam os números de observações em cada grupo.

Ligações simples produzem grupos com diâmetros grandes e ligações completas produzem grupos com diâmetros pequenos; agrupamentos por médias representam um compromisso entre esses dois extremos.

## Referências

- Hastie, T., Tibshirani, R. and Friedman, J. (2017). *The Elements of Statistical Learning*, 2nd Edition, Springer.
- James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017). *Introduction to Statistical Learning*. Springer.
- Morettin, P. A. e Singer, J. M. (2022). *Estatística e Ciência de Dados*. LTC: Rio de Janeiro.