Análise de Agrupamento

Minicurso de Machine Learning · I CiDAMO

Prof. Cesar Taconeli
taconeli@ufpr.br

Prof. Walmes Zeviani
walmes@ufpr.br

Departamento de Estatística
Universidade Federal do Paraná

13 de fevereiro, 2020

Contextualização

Programação do Curso

1. Introdução, regressão linear e polinomial - Prof. César Augusto Taconeli
2. Validação cruzada, overfitting e underfitting - Prof. Abel Soares Siqueira
3. Classificação, KNN, árvores e florestas - Prof. Lucas Garcia Pedroso
4. Naive Bayes e regressão logística - Henrique Laureano
5. **Clusterização - Prof. Walmes Zeviani**
6. Análise descritiva e tratamento de dados - Kally Chung

Aprendizado Supervisionado

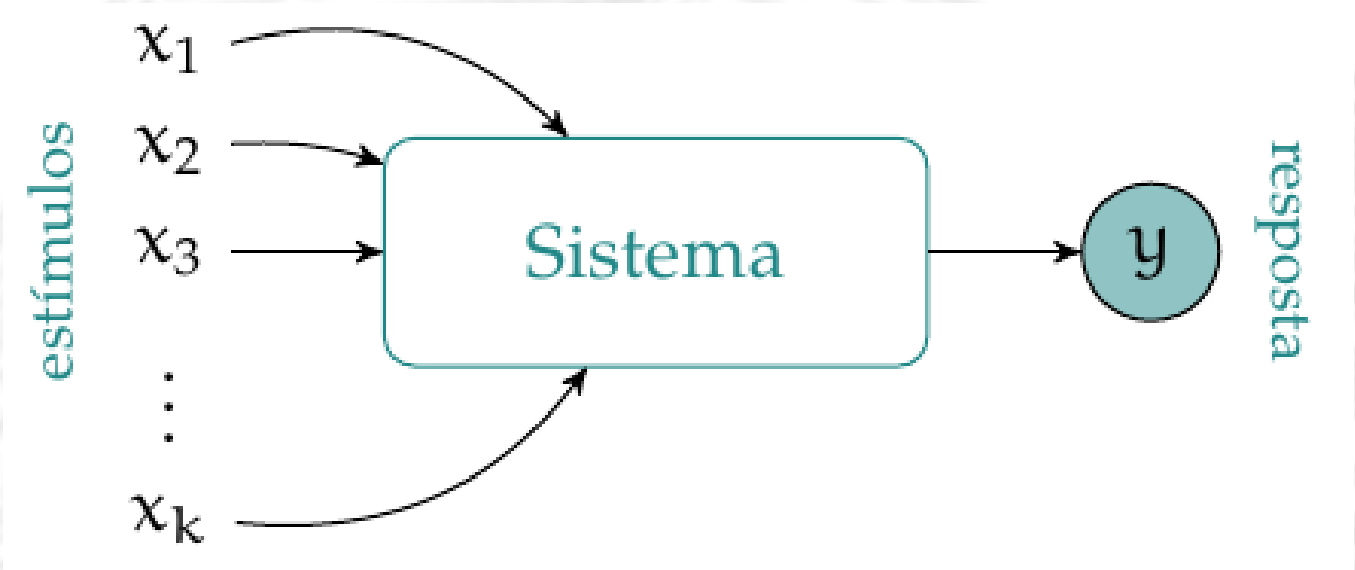


Figura 1. Modelo mental de um algorítmo de aprendizado supervisionado.

- *Aprendizado supervisionado* refere-se ao caso em que um conjunto de variáveis $X_1, X_2, \ldots, X_p$, medidas em n indivíduos, são usadas para explicar (predizer) uma variável resposta (Y).

Figura 2. Variáveis que influenciam o preço de uma habitação.

Aprendizado Não Supervisionado

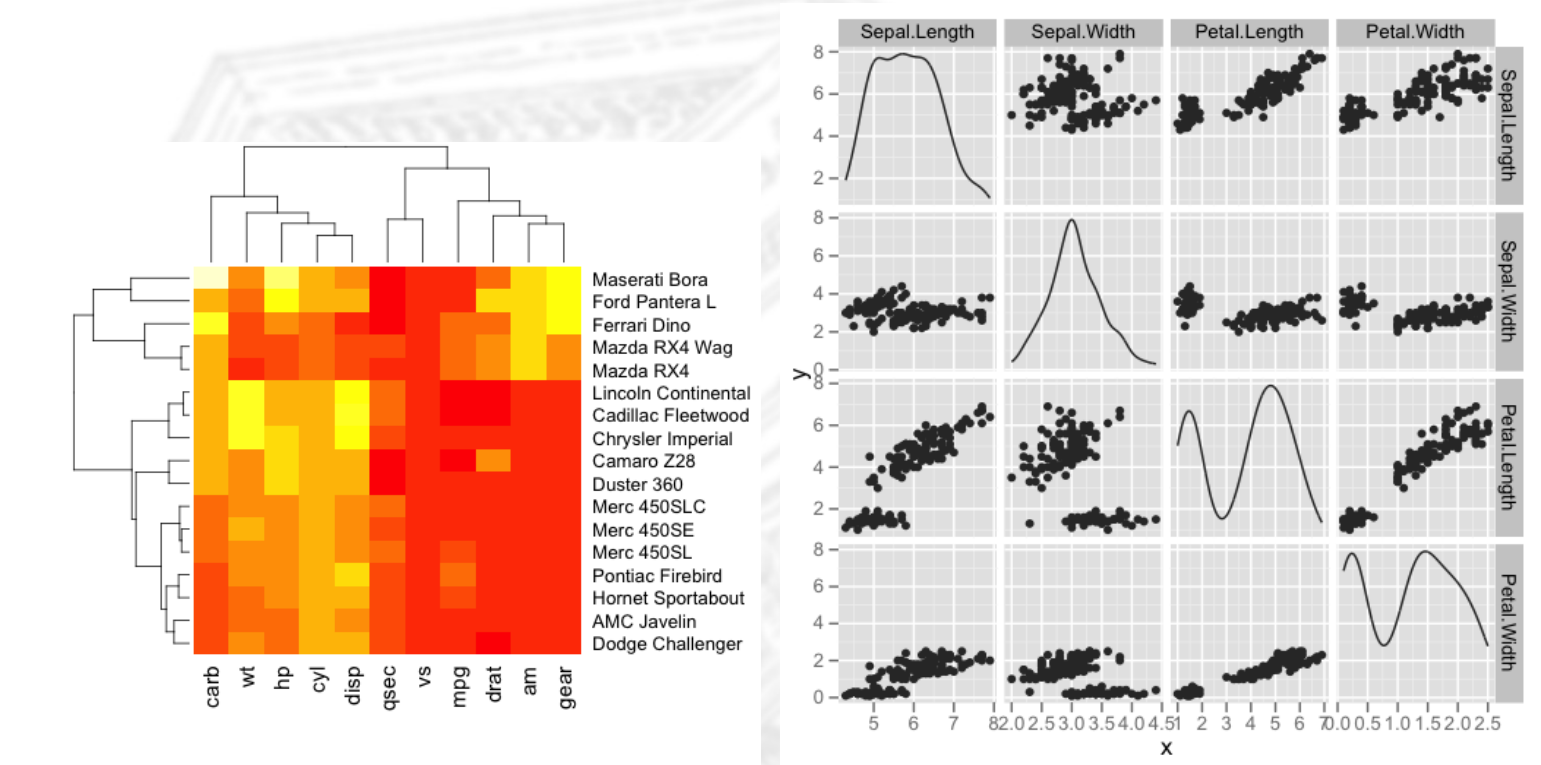


Figura 3. Visualizações gráficas típicas de métodos de aprendizado não supervisionado.

- No caso de *aprendizado não supervisionado*, não temos uma variável resposta, sendo que o interesse é explorar informações do conjunto de variáveis em análise.

Principais tarefas não supervisionadas

- **Análise de agrupamento**.
- Regras de associação: .
- Redução de dimensionalidade.
- Além de outras dentro de Engenharia de Características.

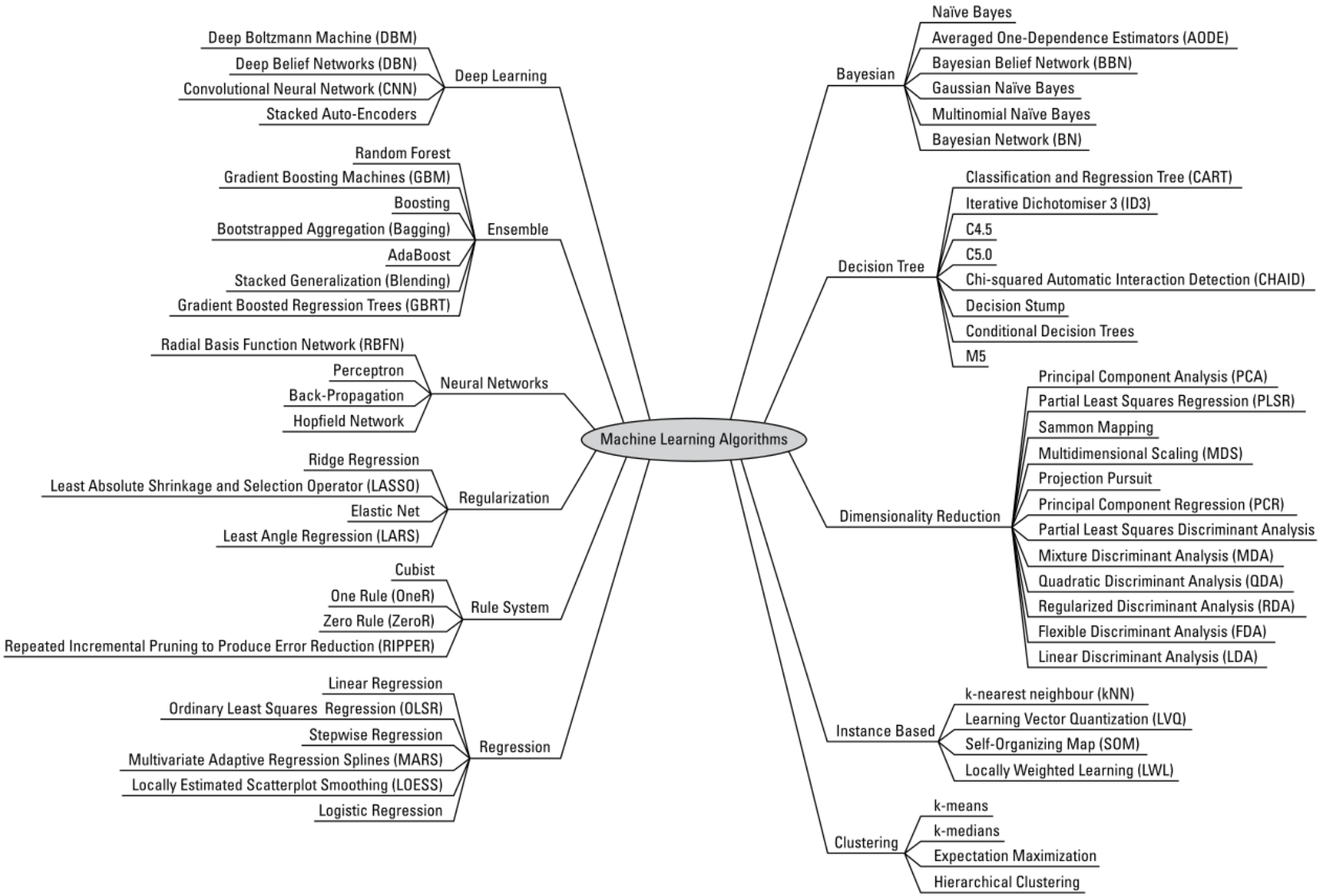


Figura 4. Métodos de machine learning subdivididas em algoritmos. Fonte: Pierson, L. (2017). Data Science for Dummies. John Wiley & Sons.

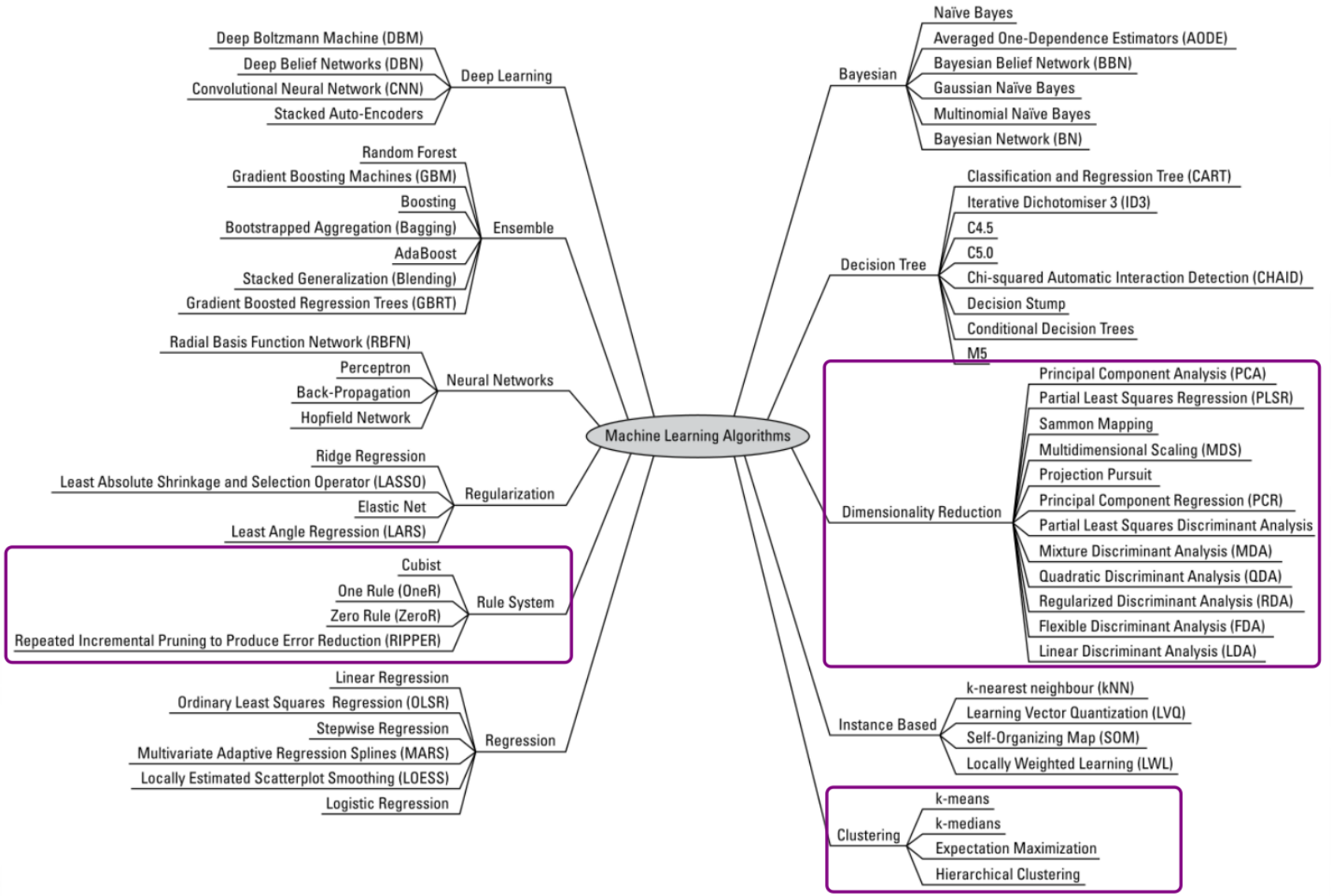


Figura 5. Métodos de machine learning subdivididas em algoritmos com destaque para abordagens não supervisionadas.

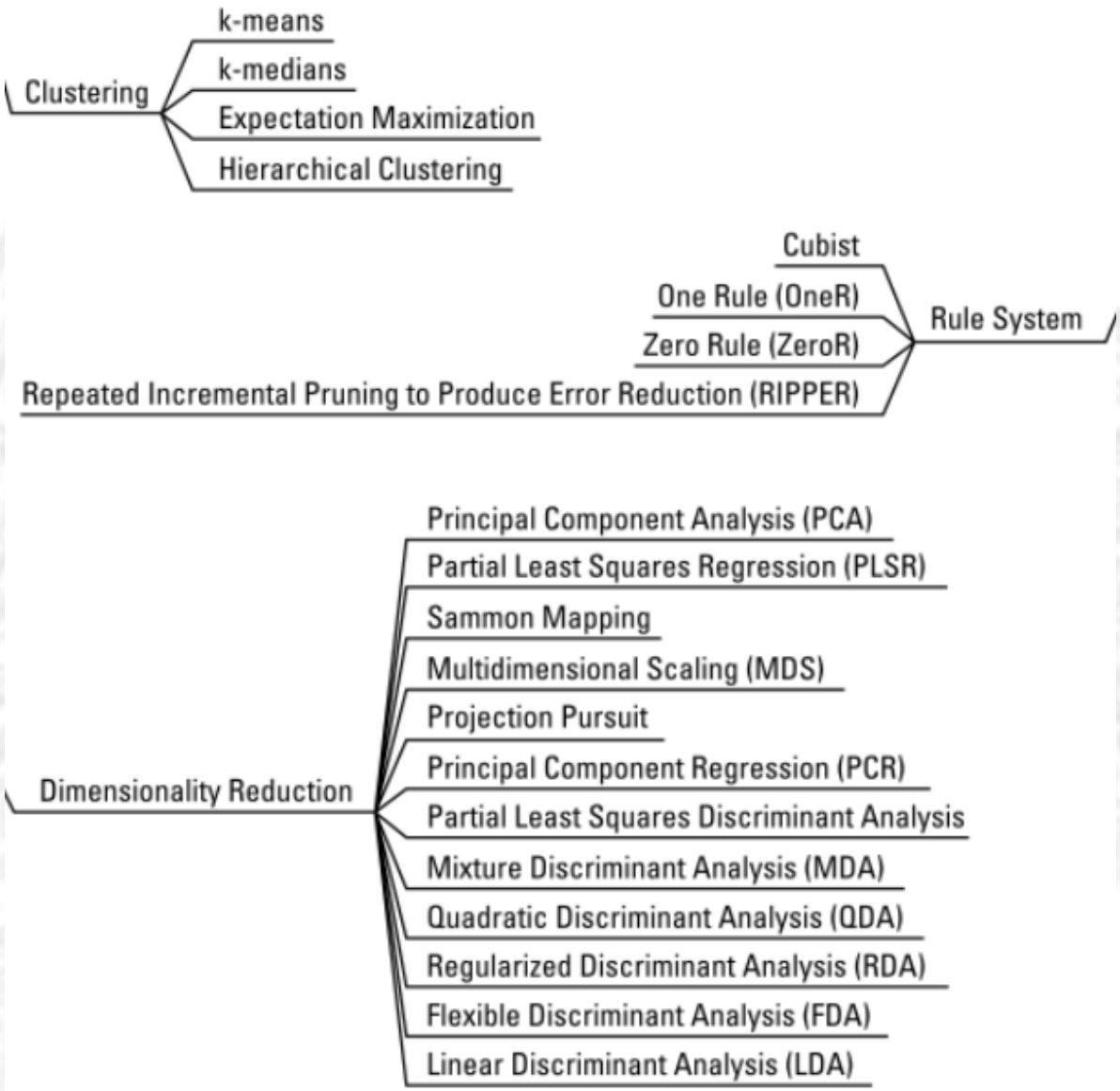


Figura 6. Métodos de machine learning subdivididas em algoritmos com destaque para abordagens não supervisionadas.

Análise de Agrupamento

Objetivo

- A **análise de agrupamento ou clustering** é uma das principais técnicas de aprendizado não supervisionado.
- Seu objetivo principal é agrupar (ou segmentar) indivíduos em *clusters*, de maneira que:
  - Indivíduos de um **mesmo cluster** sejam **semelhantes** (similares) em relação aos valores das das variáveis em análise;
  - Por outro lado, indivíduos de **clusters distintos** sejam **diferentes** (dissimilares).

Exemplos de utilidade

Exemplos do ponto de vista do negócio:

- Dividir para conquistar → segmentação de clientes.
- Marketing direcionado → tipos de desconto para grupos de mesmo perfil de compra.
- Entendimento dos clientes → desenvolvimento de produtos voltados para grupos de necessidades específicas.
- Sistemas de recomendação → recomendação baseada no comportamento dos pares.

Como agrupar esses indivíduos?

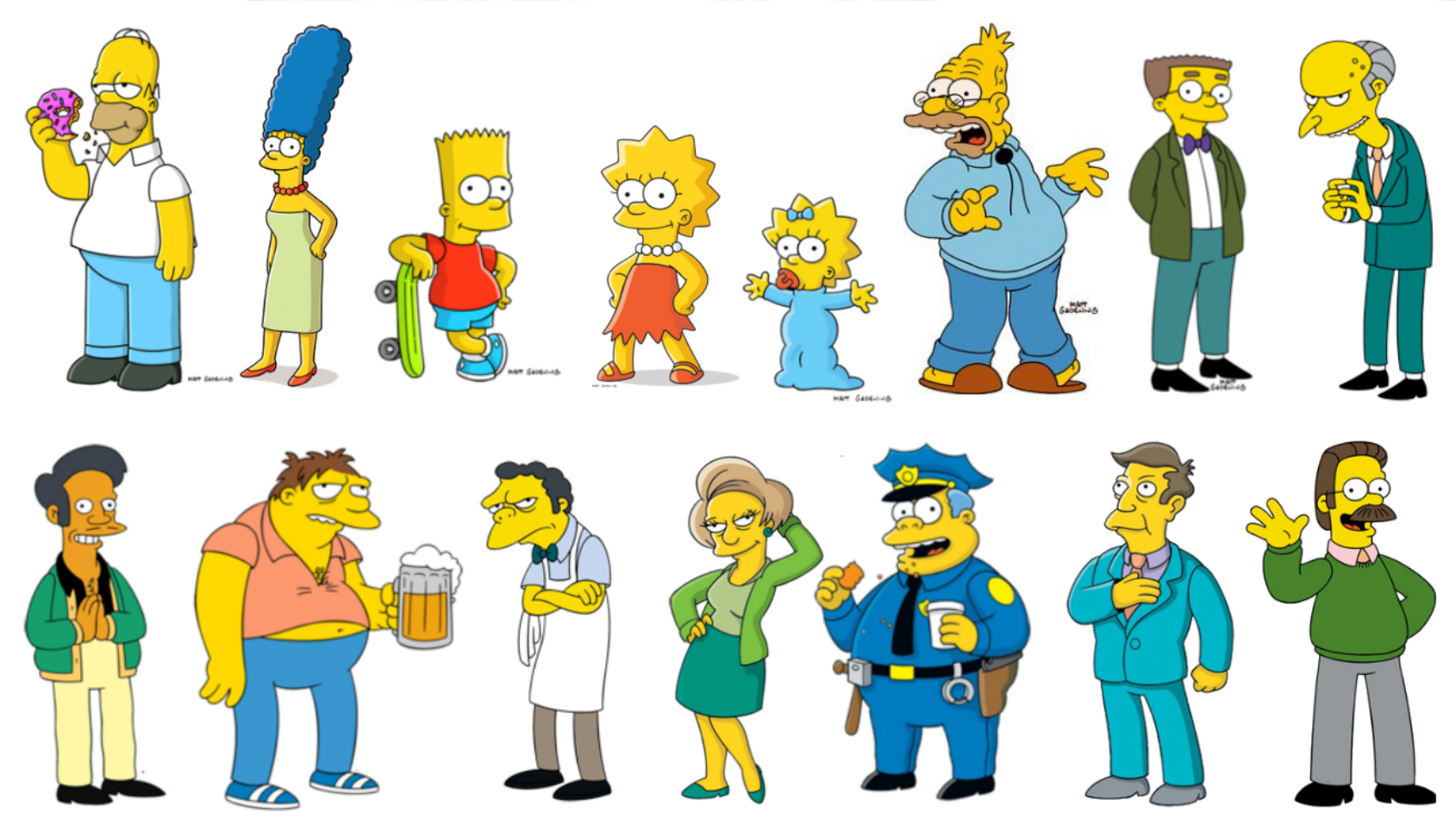


Figura 7. Alguns personagens da série Os Simpons.

Pela altura? Idade? Gênero? Parentesco? Hábitos?

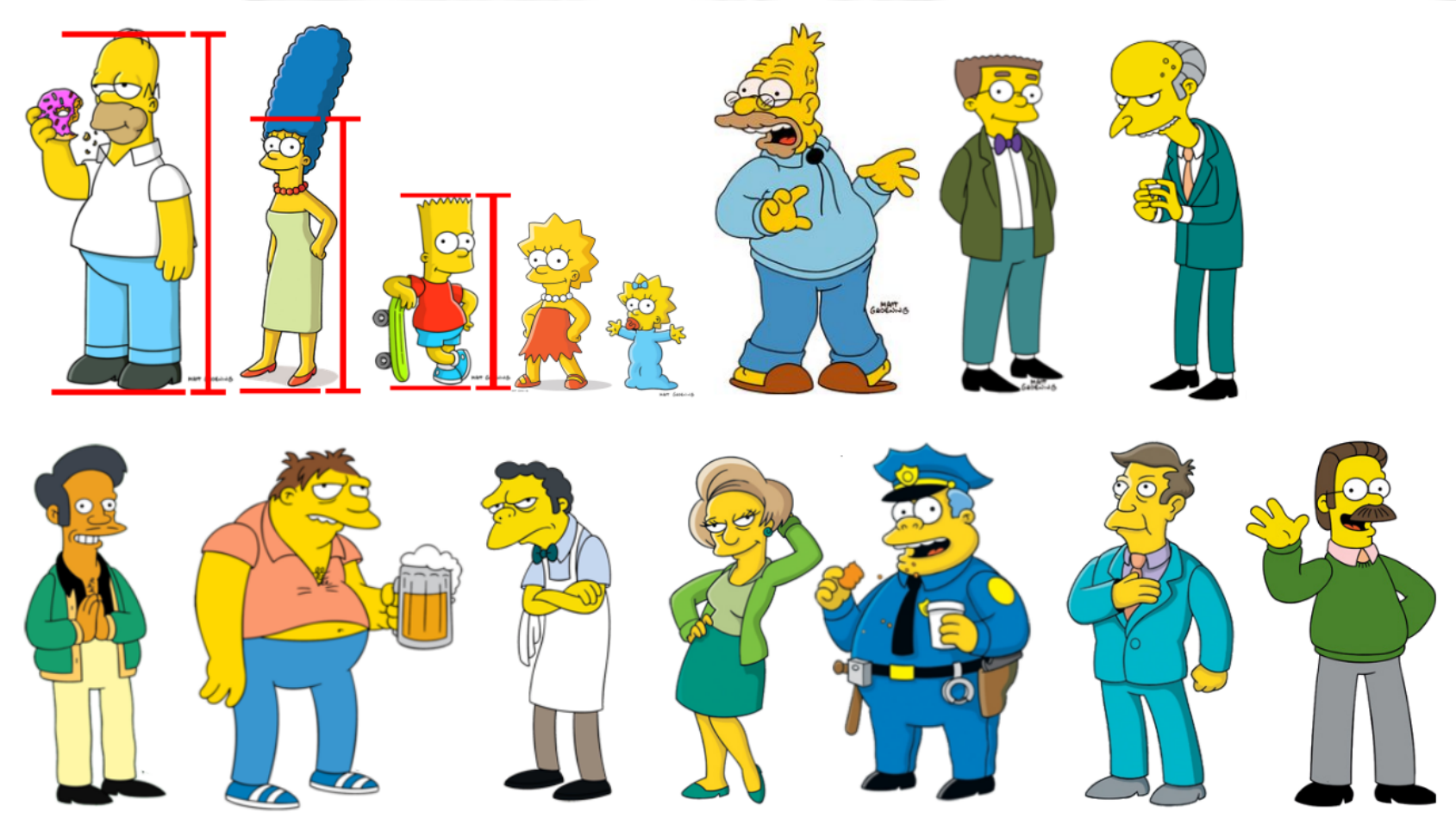


Figura 8. Alguns personagens da série Os Simpons.

O que é preciso para bons agrupamentos?

Sob a perspectiva prática (definido pelo analista)

- Um contexto com propósito bem definido.
- Variáveis relevantes para o agrupamento.
- Uma medida de similaridade e algorítmo apropriado para o contexto.

Sobre a aplicação (característica dos dados)

- Densidade: indivíduos similares nas características relevantes ocupando uma mesma região do espaço.
- Separação: regiões vazias separando os indivíduos dissimilares.

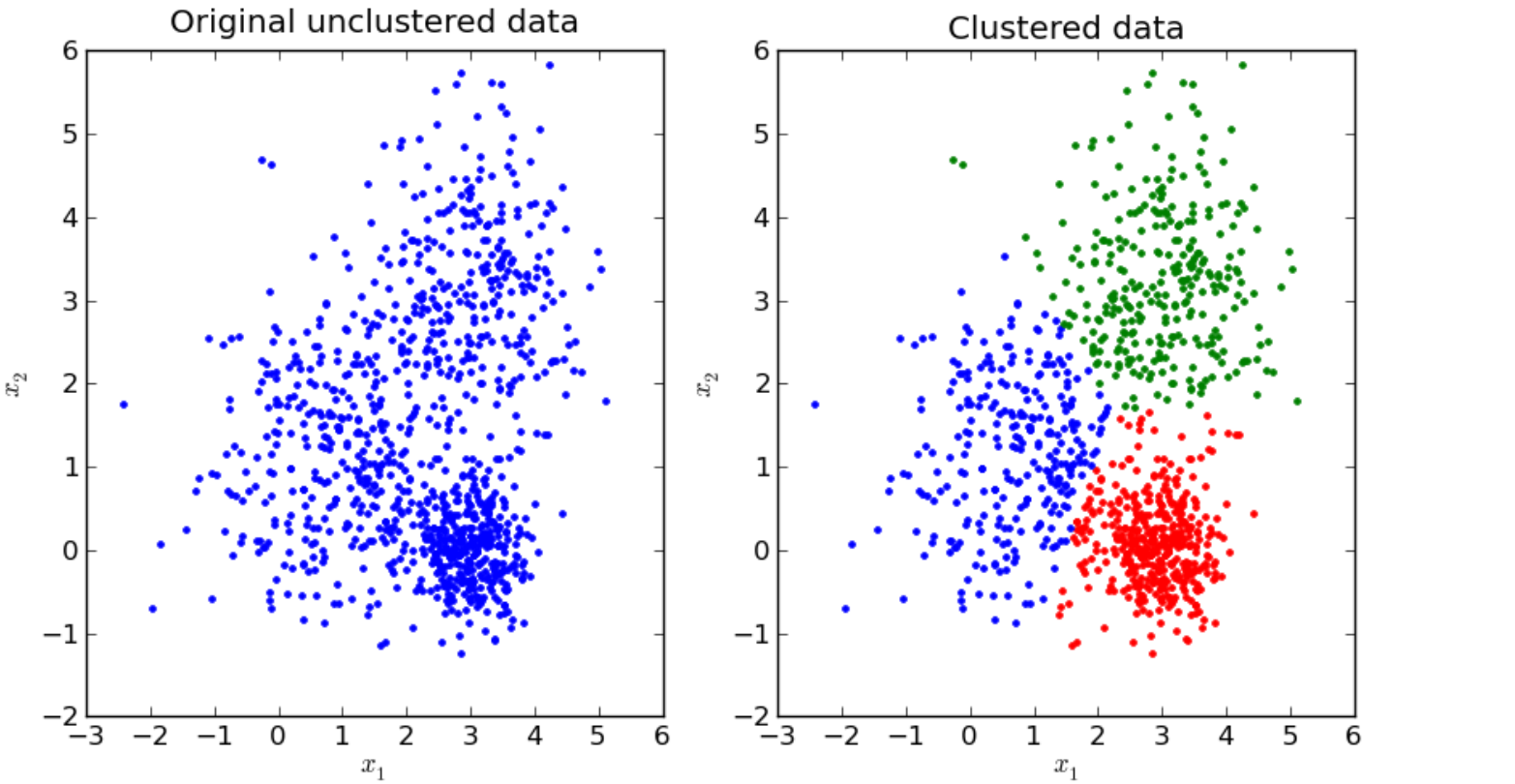


Figura 9. Dados para aplicação de métodos de agrupamento.

Medidas de dissimilaridade

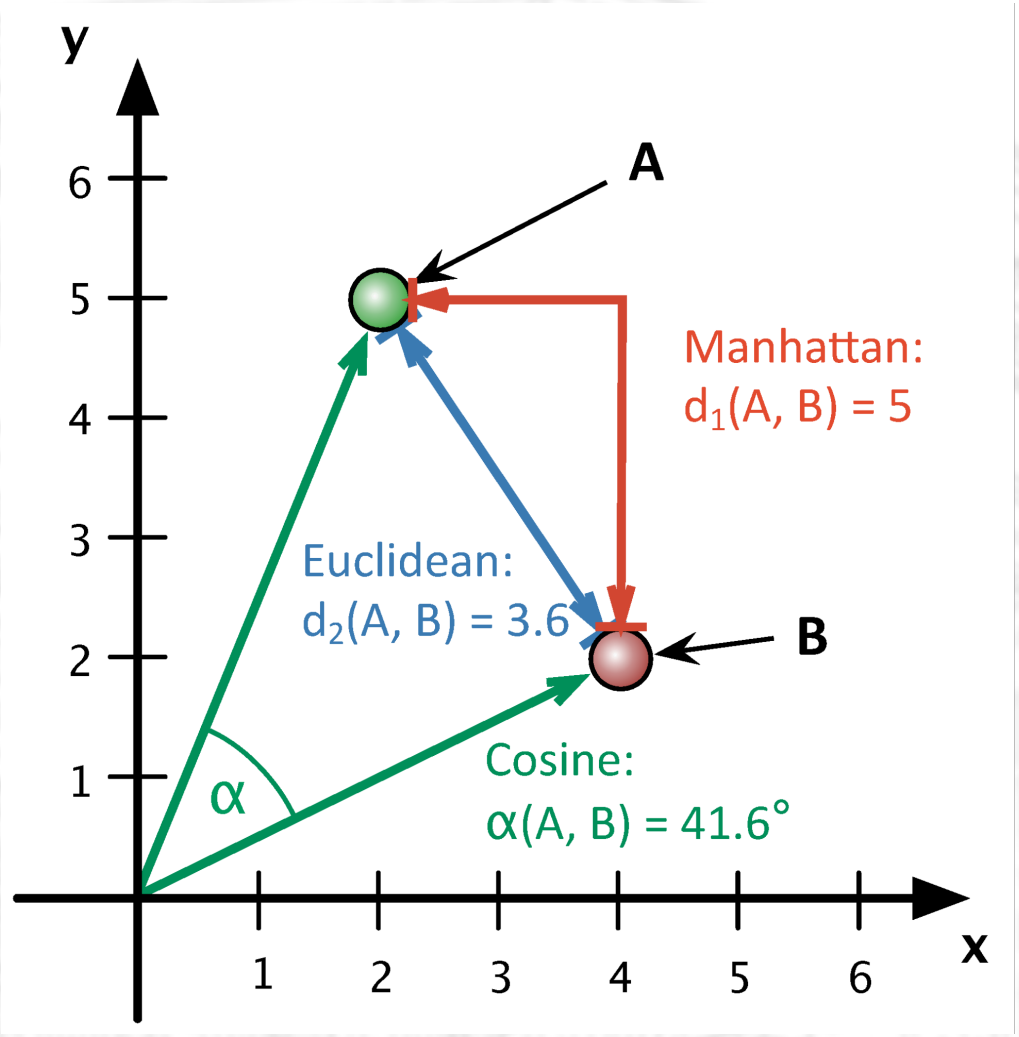


Figura 10. Ilustração em 2D de medidas de dissimilaridade ou distâncias.

Medidas de dissimilaridade

- Pode-se expressar proximidade/distância de forma matemática.
- Algoritmos de análise de clusters baseiam-se em **medidas de dissimilaridade**.
- Elas permitem quantificar a diferença entre indivíduos com base nos valores apresentados para o conjunto de variáveis.
- Medidas de dissimilaridade podem ser aplicadas a cada par de indivíduos entre os n disponíveis.
- Vamos denotar por $d_{ii'}$, com i e $i' \in \{1, 2, \ldots, n\}$, a dissimilaridade avaliada para um par de indivíduos i e i'.

Medidas de dissimilaridade

- O conjunto de medidas de dissimilaridade, calculadas para cada par de indivíduos, é usualmente representado numa matriz de dimensão $n \times n$, denominada **matriz de dissimilaridades**, dada por:

$$\boldsymbol{D} = \begin{bmatrix} d_{11} & d_{12} & \dots & d_{1n} \\ d_{21} & d_{22} & \dots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & d_{nn} \end{bmatrix}.$$

- Há uma grande variedade de formas de se definir (e quantificar) dissimilaridades.

Medidas de dissimilaridade

- Dissimilaridades podem ser estabelecidas por um processo *informal*, em que especialistas (juízes) atribuem valores (escores) de dissimilaridade para cada par de indivíduos.
- Em geral, no entanto, os algoritmos de análise de clusters baseiam-se em medidas de dissimilaridade que atendem às seguintes propriedades:
  1. $d_{ii'} \geq 0$, com $d_{ii'} = 0$ se $i = i'$;
  2. $d_{ii'} = d_{i'i}$ para todo $i, i' \in 1, 2, \ldots, n$ (simetria);
  3. $d_{ii'} \leq d_{ik} + d_{i'k}$, para todo $k \in 1, 2, \ldots, n$ (desigualdade triangular).

Cálculo para variáveis contínuas

- Vamos assumir, num primeiro momento, uma variável x_j contínua.
- Algumas medidas usuais de dissimilaridade, neste caso, são:

1. Distância quadrática:

$$d_j(x_{ij}, x_{i'j}) = \left(x_{ij} - x_{i'j}\right)^2.$$

2. Diferença absoluta:

$$d_j(x_{ij}, x_{i'j}) = \left|x_{ij} - x_{i'j}\right|.$$

Cálculo para variáveis de escala ordinal

- Em algumas aplicações, determinadas variáveis apresentam escala ordinal.
- Como exemplo, podemos citar:
  - Nível de satisfação com um serviço (nada, pouco, muito, totalmente satisfeito).
  - Formação escolar (sem escolaridade, ensino primário, ensino médio, etc.).
  - Estágio de uma doença (não manifestada, estágio inicial, estágio intermediário, etc.).
  - Categoria do cliente (bronze, prata, ouro, platinum).
- Uma das formas de proceder em situações desse tipo é ranquear as (digamos M) categorias da escala ordinal em ordem crescente.

Cálculo para variáveis de escala ordinal

- As medidas de dissimilaridade para escalas contínuas podem ser aplicadas substituindo as observações originais por:

$$x_{ij}^{*} = \frac{k - 1/2}{M},$$

em que $k \in \{1, 2, \ldots, M\}$ representa o ranking correspondente ao resultado de x_j em i.

- Se não for razoável atribuir ranks equidistantes às M categorias de x_j, alguma outra configuração mais apropriada de valores pode ser assumida.

Cálculo para variáveis de escala nominal

- São exemplos de variáveis com escala nominal:
  - Marca do modelo de veículo/celular/geladeira, etc.
  - Time para o qual torce.
  - Meio de transporte para o trabalho (à pe, bike, carro, público).
  - Forma de pagamento (dinheiro, cartão de crédito, cartão de débito, cheque).
  - Finalidade de um empréstimo bancário (pagamento de dívida, compra de imóvel, compra de automóvel, abertura de negócio próprio, etc.).
- Nesse caso, geralmente não faz sentido atribuir ranks ou escores às categorias, dada a ausência de qualquer sentido de ordenação natural.

Cálculo para variáveis de escala nominal

- A forma mais simples de medir dissimilaridades consiste em considerar:

$$d_j(x_{ij}, x_{i'j}) = 0, \text{ se } x_{ij} = x_{i'j},$$

e

$$d_j(x_{ij}, x_{i'j}) > 0, \text{ caso contrário.}$$

- O mais comum é definir $d_j(x_{ij}, x_{i'j}) = 1$ sempre que $x_{ij} \neq x_{i'j}$, embora configurações alternativas permitam atribuir dissimilaridades maiores para algumas combinações de valores.

Situação típica em Ciência de Dados

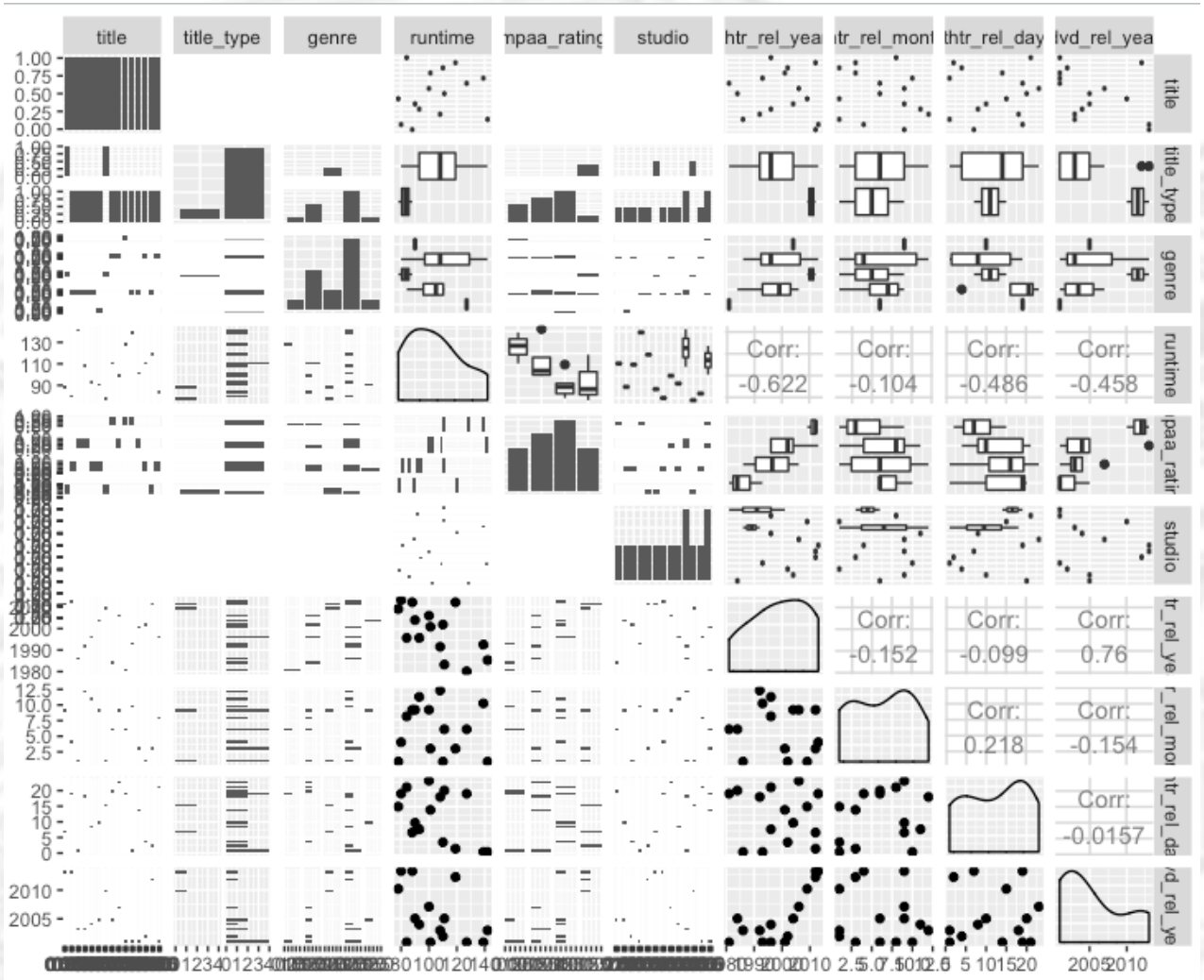


Figura 11. Variáveis de tipos mistos.

Situação típica em Ciência de Dados

- Em contextos aplicados dispõe-se de mais de uma variável relevante para a segmentação.
- Não raramente, as variáveis podem ser de tipos mistos (contínuas, ordinais e nominais).
- Como representar a dissimilaridade diante desse cenário?

Dissimilaridade entre pares de indivíduos

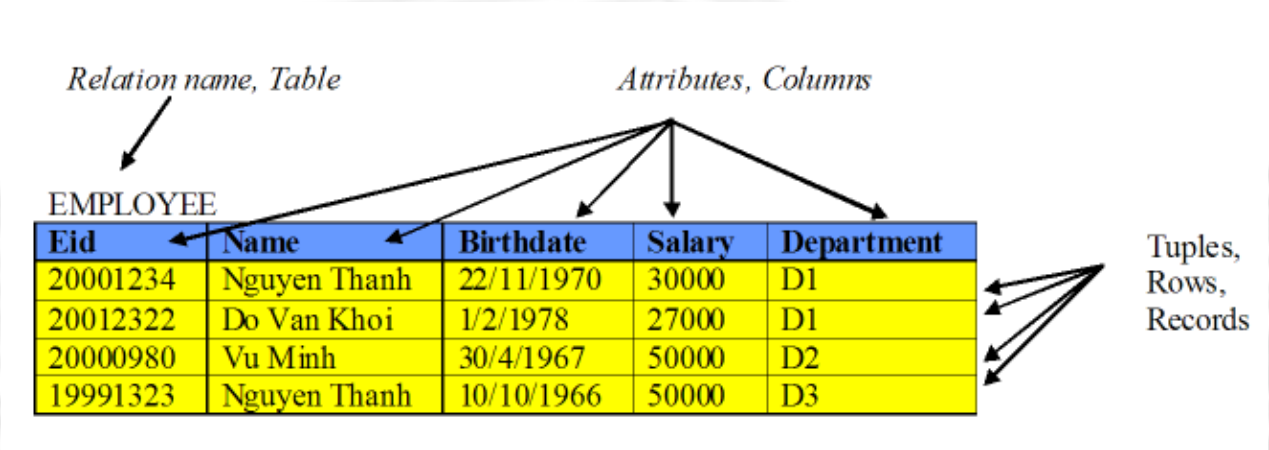


Figura 12. Uma típica tabela de dados de cadastro de pessoa.

- Pesos diferentes para cada variável podem ser estabelecidos, por exemplo, para refletir a **importância** de cada variável na análise, mas pode ser **subjetivo**.
- Um motivo adicional para ponderação é remover o efeito de **escala** das p variáveis.
- Não havendo motivos para diferentes ponderações, podemos assumir $\omega_j = 1$, para $j = 1, 2, \ldots, p$.

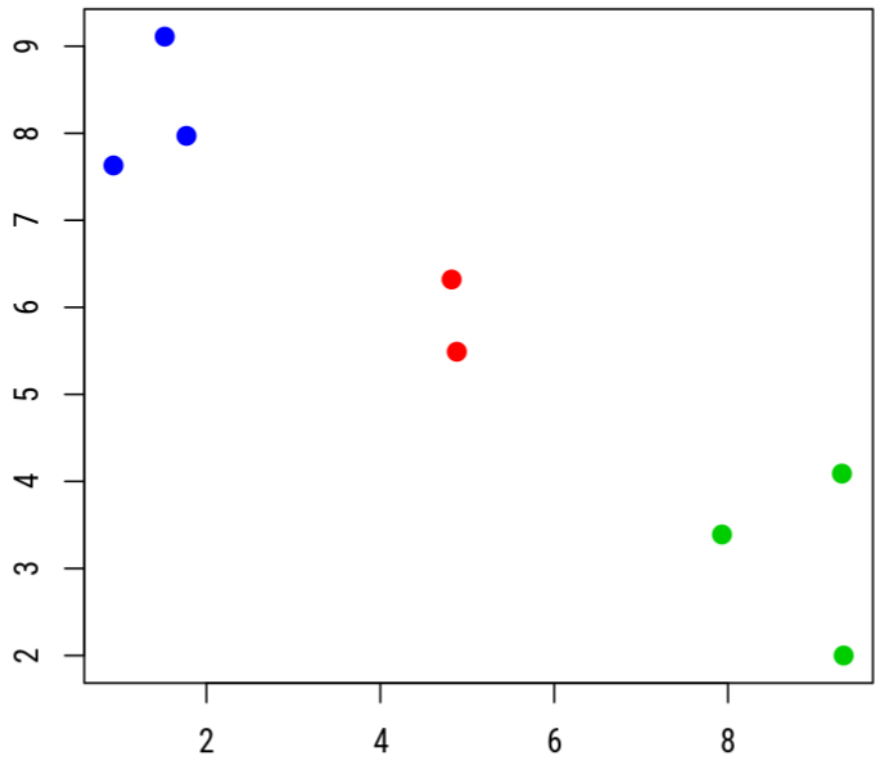
Algoritmos para análise de agrupamento

Alocação dos indivíduos aos grupos

- Como resultado para uma análise de clusters, cada indivíduo (i) é alocado a um cluster k ($k \in \{1, 2, \ldots, K\}$) segundo um *codificador* $k = C(i)$.
- O objetivo é encontrar um codificador "ótimo", que permita constituir, o máximo possível, clusters homogêneos internamente e heterogêneos entre si.
- A performance de um codificador C pode ser avaliada, por exemplo, pela dissimilaridade entre observações alocadas a um mesmo cluster:

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} d_{ii'}.$$

Decomposição da dissimilaridade total



Dissimilaridades dentro de cluster · $W(C)$

	1	2	3	4	5	6	7	8
1	0	1.59	0.91	4.10	4.49	9.10	8.18	10.11
2		0.00	1.17	4.32	4.94	9.27	8.59	10.56
3			0.00	3.47	3.98	8.48	7.68	9.63
4				0.00	0.83	5.01	4.27	6.25
5					0.00	4.65	3.70	5.66
6						0.00	1.55	2.09
7							0.00	1.97
8								0.00

Dissimilaridades entre cluster · $B(C)$

	1	2	3	4	5	6	7	8
1	0	1.59	0.91	4.10	4.49	9.10	8.18	10.11
2		0.00	1.17	4.32	4.94	9.27	8.59	10.56
3			0.00	3.47	3.98	8.48	7.68	9.63
4				0.00	0.83	5.01	4.27	6.25
5					0.00	4.65	3.70	5.66
6						0.00	1.55	2.09
7							0.00	1.97
8								0.00

Figura 13. Decomposição da dissimilaridade total.

Decomposição da dissimilaridade total

- A dissimilaridade total para o conjunto de n observações da amostra pode ser decomposta por:

$$\begin{aligned} T &= W(C) + B(C) \\ &= \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')=k} d_{ii'} + \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')\neq k} d_{ii'}, \end{aligned}$$

  em que $W(C)$ quantifica a dissimilaridade *intra clusters* e $B(C)$ a dissimilaridade *entre clusters*;
- Fixado K, quanto menor $W(C)$ (e maior, consequentemente, $B(C)$), melhor o codificador (composição dos clusters).

O número de agrupamentos

- Fixado o **número de clusters** (K) o número de codificadores distintos e, consequentemente, diferentes soluções para a análise de clusters, é dado por:

$$S(n, K) = \frac{1}{K!} \sum_{k=1}^{K} (-1)^{K-k} \binom{K}{k} k^n.$$

- O **número de soluções** aumenta muito rapidamente conforme aumentam n e k.
- Assim, a avaliação de todas as possíveis soluções torna-se **inviável** mesmo para valores "moderados" de n e k.

Heurísticas dos algoritmos

- Os algoritmos de análise de cluster permitem avaliar uma **fração** das possíveis soluções e identificar, baseado em algum critério, a melhor.
- Ao não avaliar todas as possíveis soluções, a solução encontrada pode ser **sub-ótima**.
- Adicionalmente, diferentes algoritmos (e critérios de avaliação) podem conduzir a soluções bastante diferentes.

Algoritmos de agrupamento não hierárquicos

- Os algoritmos **hierárquicos** baseiam-se em sucessivas aglomerações (ou partições) dos indivíduos com base numa matiz de dissimilaridades.
- Os algoritmos **não hierárquicos**, por sua vez, baseiam-se em sucessivas re-alocações dos indivíduos aos clusters, visando a constituição de clusters internamente mais homogênos.
- Dentre os algoritmos não hierárquicos mais conhecidos destacam-se o *K-means* e o *K-medoids*.

Algoritmo *K-means*

- O algoritmo $K - means$ se aplica quando as variáveis sob análise são quantitativas e a dissimilaridade é baseada na distância Euclideana:

$$d_{ii'} = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = ||\boldsymbol{x}_i - \boldsymbol{x}_{i'}||^2,$$

  que pode, eventualmente, ser ponderada.
- A ponderação pode estar embutida na etapa de padronização das variáveis.
  - Padronização Z-escore: média 0 e variância 1 ou outra.
  - Padronização unitária: mínimo 0 e máximo 1 ou outro.

Algoritmo *K-means*

- A dissimilaridade total intra clusters fica dada por:

$$W(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')=k}||\boldsymbol{x}_i - \boldsymbol{x}_{i'}||^2$$

$$= \sum_{k=1}^{K} N_K \sum_{C(i)=k}||\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k||^2,$$

em que N_k é o número de indivíduos e $\bar{\boldsymbol{x}}'_k = (\bar{x}_{1k}, \bar{x}_{2k}, \ldots, \bar{x}_{pk})$ é o vetor de médias no cluster k.

- O algoritmo $k-means$ busca identificar uma codificação (C^*) em K clusters (K fixado) em que a distância das observações à média do cluster seja mínima,

$$C^* = \min_{C}\sum_{k=1}^{K}\left[N_k \sum_{C(i)=k}||\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k||^2\right].$$

Algoritmo *K-means*

- Dado que, para qualquer conjunto de observações S:

$$\bar{\boldsymbol{x}}_S = \underset{\mathrm{m}}{\operatorname{argmin}} \sum_{i \in S} ||\boldsymbol{x}_i - m||^2,$$

então a solução do método $k - means$ corresponde à solução do seguinte problema de otimização:

$$\min_{C, m_k} \sum_{k=1}^{K} \left[N_k \sum_{C(i)=k} ||\boldsymbol{x}_i - m_k||^2 \right].$$

- O algoritmo $k - means$ é apresentado na sequência.

Descrição do algoritmo *K-means*

- **Passo 1:** Para um dado codificador C, a variância total intra cluster é minimizada com relação a $m_1, m_2, \ldots, m_K$, produzindo as médias da alocação atual;
- **Passo 2:** Dadas as médias atuais, a função objetivo é minimizada re-alocando cada observação ao cluster com média mais próxima, ou seja:

$$C(i) = \underset{1 \leq k \leq K}{\arg\min} ||\boldsymbol{x}_i - m_k||^2;$$

- **Passo 3:** Repetir os passos 1 e 2 até que não haja novas re-alocações.

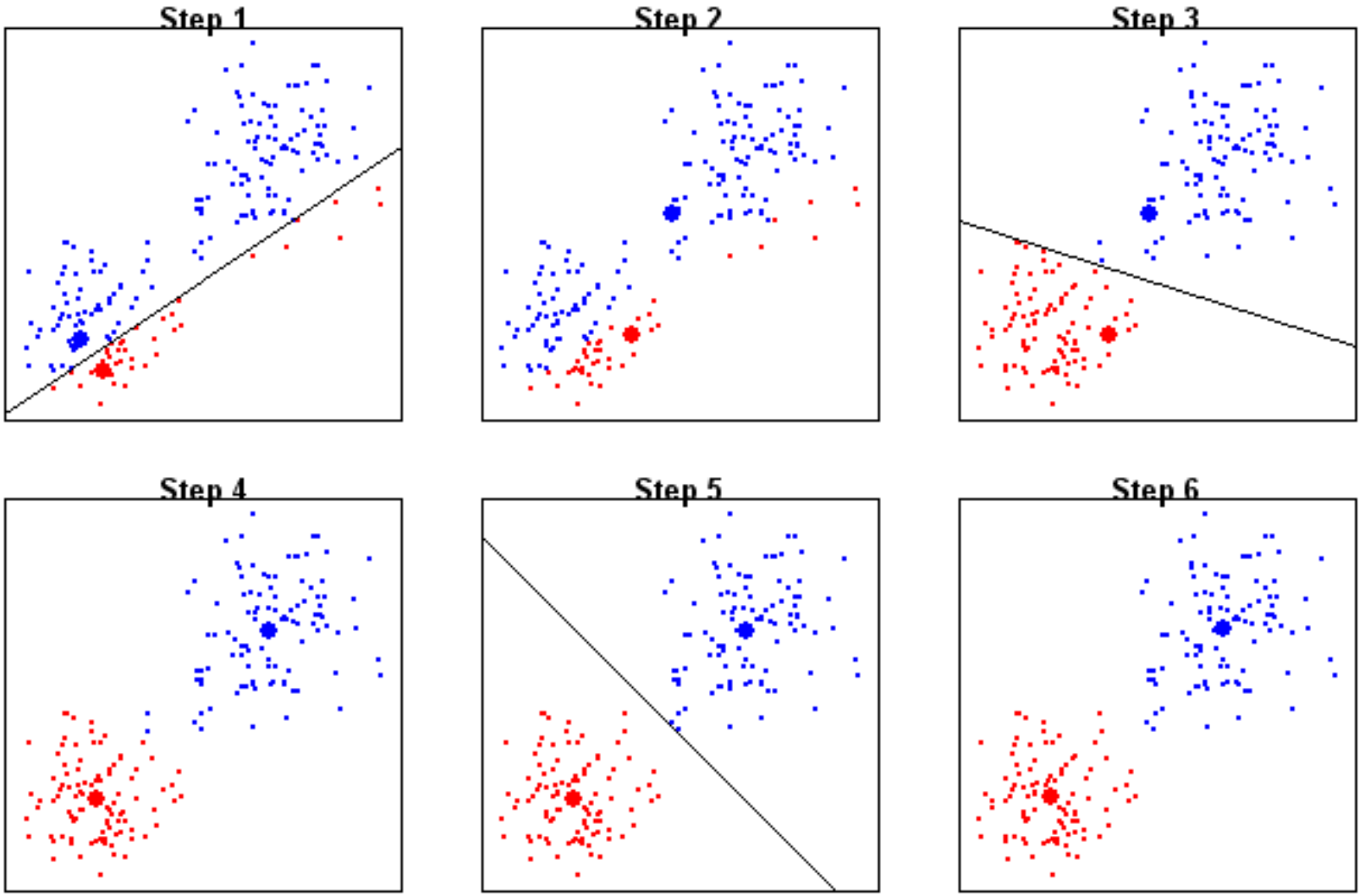


Figura 14. Algortimo *k-means* com 2 grupos.
https://animoidin.files.wordpress.com/2018/07/0-rrzg3lyonavoepbj.png.

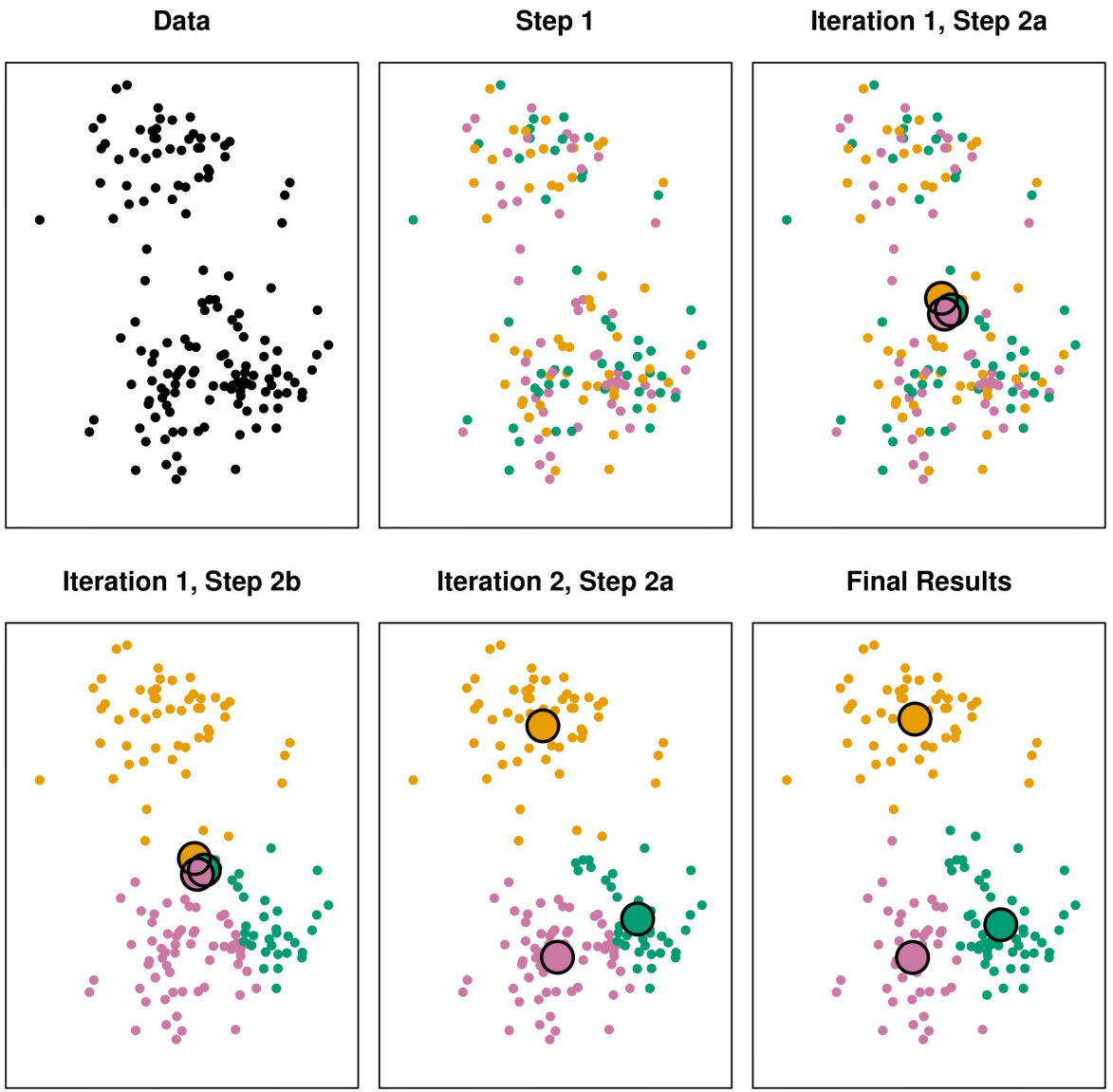


Figura 15. Algortimo *K-means* com 3 grupos.
https://i.stack.imgur.com/FQhxk.jpg.

Considerações sobre o algoritmo *K-means*

- O algoritmo *K-means* é sensível à configuração inicial dos clusters no passo 1, podendo produzir resultados diferentes mediante diferentes partições iniciais.
- O usual é considerar, inicialmente, $t > 1$ "sementes", que seriam t pontos definidos em $\mathbb{R}^p$.
- A solução que produzir menor distância média das observações às respectivas médias dos nós é escolhida.

Algoritmos de agrupamento hierárquicos

- Nos **métodos hierárquicos aglomerativos**, cada indivíduo, originalmente, é um cluster, iniciando-se o processo com n clusters.
- Na sequência, indivíduos similares são sucessivamente agrupados, até a formação de um único grupo contendo toda a amostra.
- Nos **métodos divisivos**, partimos de um único cluster que contém toda a amostra, que é sucessivamente subdividido.

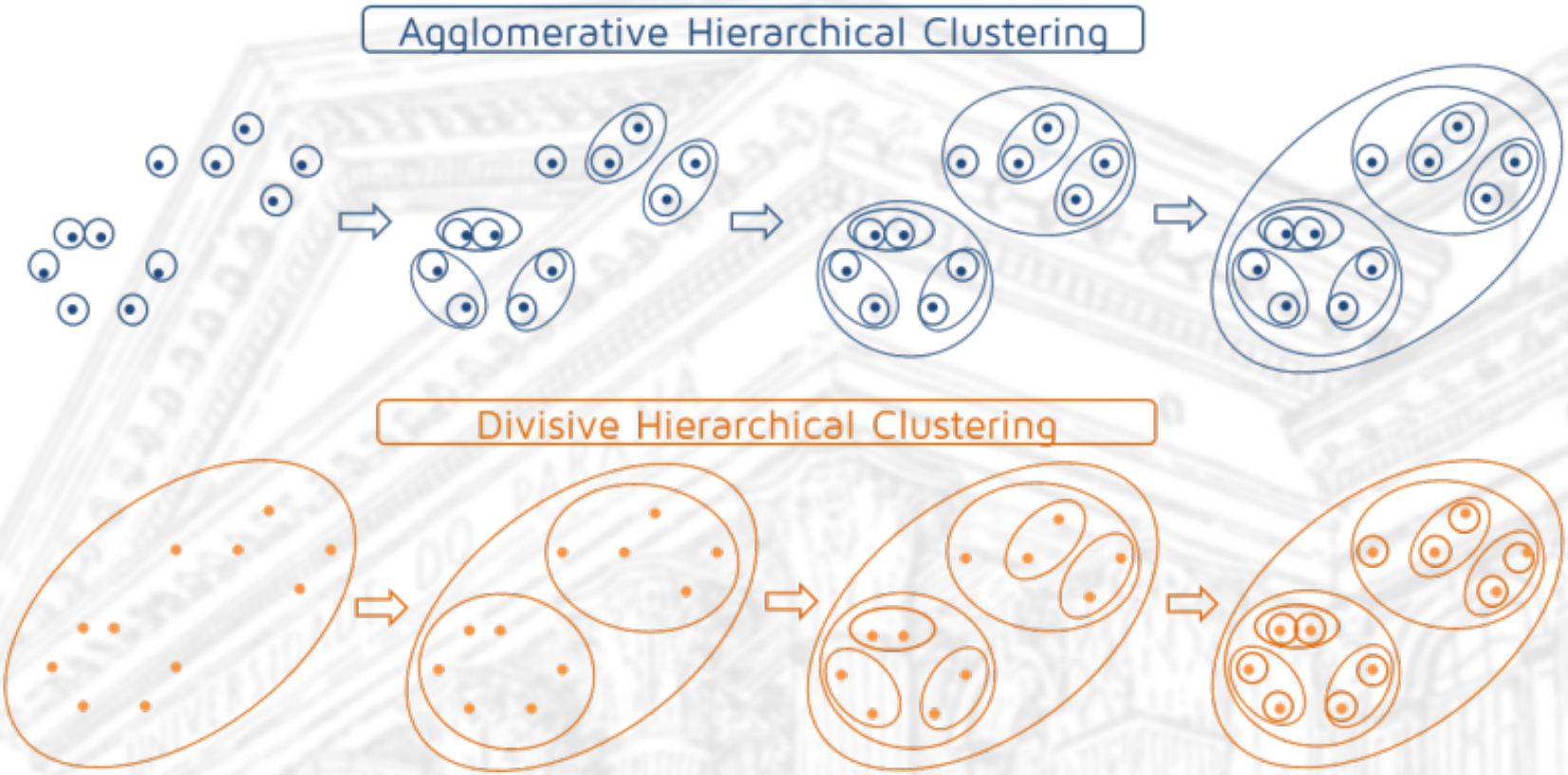


Figura 16. Funcionamento de métodos aglomerativos e divisivos. http://quantdare.com/wp-content/uploads/2016/06/AggloDivHierarClustering-800x389.png.

Algoritmos de agrupamento hierárquicos aglomerativos

- **Passo 1** - Calcule a matriz de distâncias para os n indivíduos. Nesta etapa, cada indivíduo é um cluster.
- **Passo 2** - Identifique, na matriz de distâncias, os dois clusters mais similares (menos distantes).
- **Passo 3** - Agrupe os dois clusters identificados no passo anterior em um único cluster.

Algoritmos de agrupamento hierárquicos aglomerativos

- **Passo 4** - Atualize a matriz de distâncias, considerando os clusters remanescentes.
- **Passo 5** - Repita os passos 2, 3 e 4 sucessivamente, até formar um único cluster.
- **Passo 6** - Represente os resultados da análise em um gráfico apropriado (dendrograma).

Algoritmos de agrupamento hierárquicos aglomerativos

- Ao longo das etapas de algoritmos hierárquicos (aglomerativos ou divisivos), precisamos atribuir dissimilaridades entre pares de indivíduos, indivíduos e cluster e entre pares de clusters.
- Há diferentes métodos disponíveis para medir dissimilaridades envlvendo clusters, dentre as quais algumas são descritas na sequência.
- Em todos os casos vamos considerar dois clusters, denotados por A e B.
- Procedimentos similares podem ser aplicados ao medir dissimilaridades entre observações e clusters.

Métodos aglomerativos

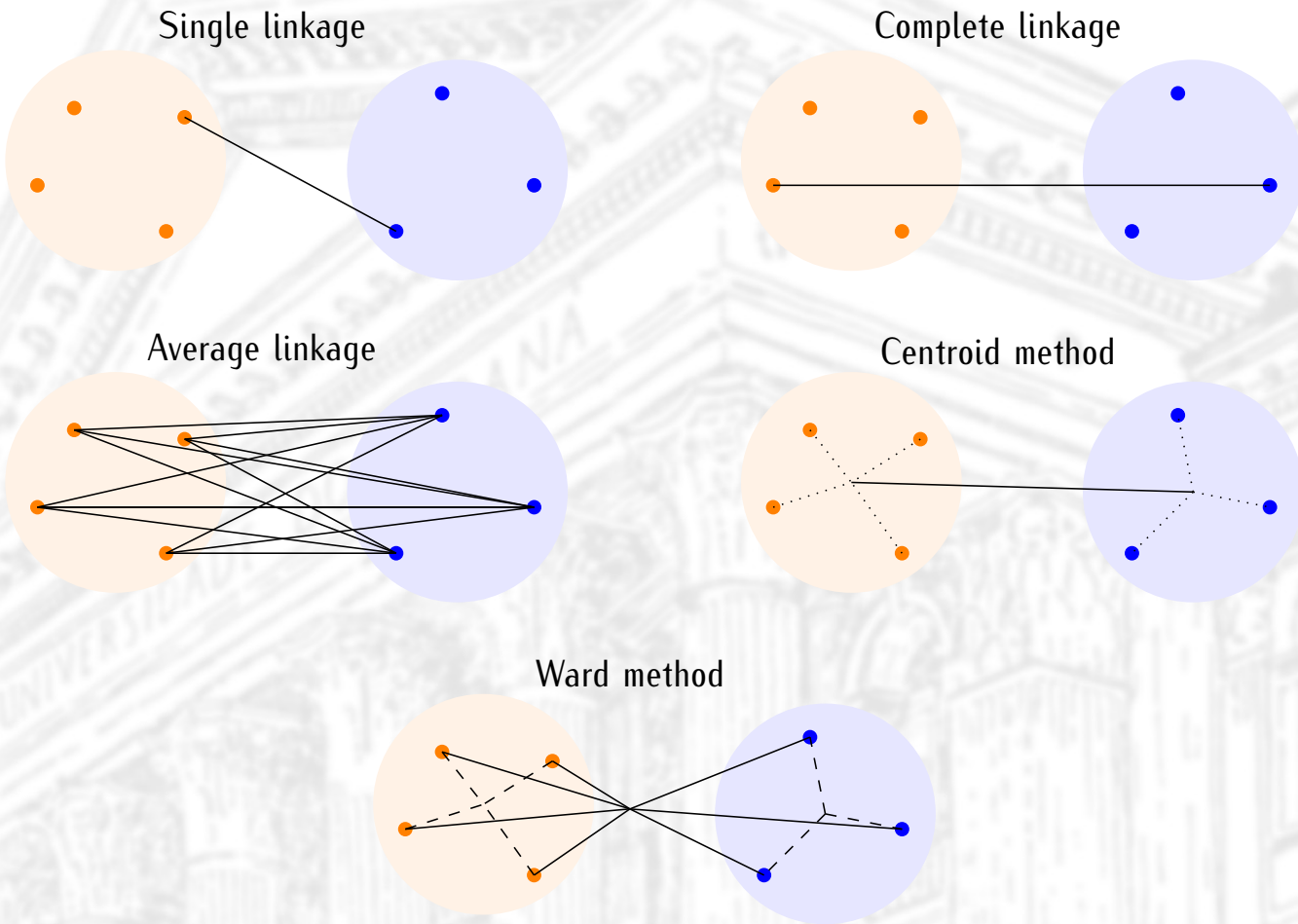


Figura 17. Métodos aglomerativos. Fonte: os autores.

Métodos aglomerativos

1. **Single linkage** - É o método do vizinho mais próximo, em que a distância entre A e B é definida como a menor distância entre uma observação de A e uma observação de B.

$$d(A, B) = min\{d(\boldsymbol{x}_i, \boldsymbol{x}_{i'})\}, \text{ para } \boldsymbol{x}_i \in A, \boldsymbol{x}_{i'} \in B.$$

2. **Complete linkage** - É o método do vizinho mais distante, em que a distância entre A e B é a distância entre o elemento de A mais distante de algum elemento de B.

$$d(A, B) = max\{d(\boldsymbol{x}_i, \boldsymbol{x}_i)\}, \text{ para } \boldsymbol{x}_i \in A, \boldsymbol{x}_{i'} \in B.$$

Métodos aglomerativos

3. **Average linkage** – Neste caso, a distância entre A e B é a média das $n_A \times n_B$ distâncias entre os n_A pontos de A e os n_B pontos de B.

$$d(A, B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{i'=1}^{n_B} d(\boldsymbol{x}_i, \boldsymbol{x}_{i'}).$$

4. **Centroide** – A distância entre A e B é definida como a distância euclideana entre os centroides (vetores de médias) dos dois clusters:

$$d(A, B) = d(\bar{\boldsymbol{x}}_A, \bar{\boldsymbol{x}}_B).$$

Métodos aglomerativos

- No método do centroide, após a junção de dois clusters A e B, o centroide do novo cluster AB fica dado pela média ponderada:

$$\bar{\boldsymbol{x}}_{AB} = \frac{n_A \bar{\boldsymbol{x}}_A + n_B \bar{\boldsymbol{x}}_B}{n_A + n_B}.$$

5. **Median** - Similar ao método do centroide mas, ao fundir dois clusters A e B, define-se o ponto mediano entre $\bar{\boldsymbol{x}}_A$ e $\bar{\boldsymbol{x}}_B$ como referência para calcular distâncias para outros clusters:

$$\boldsymbol{m}_{AB} = \frac{1}{2}(\bar{\boldsymbol{x}}_A + \bar{\boldsymbol{x}}_B).$$

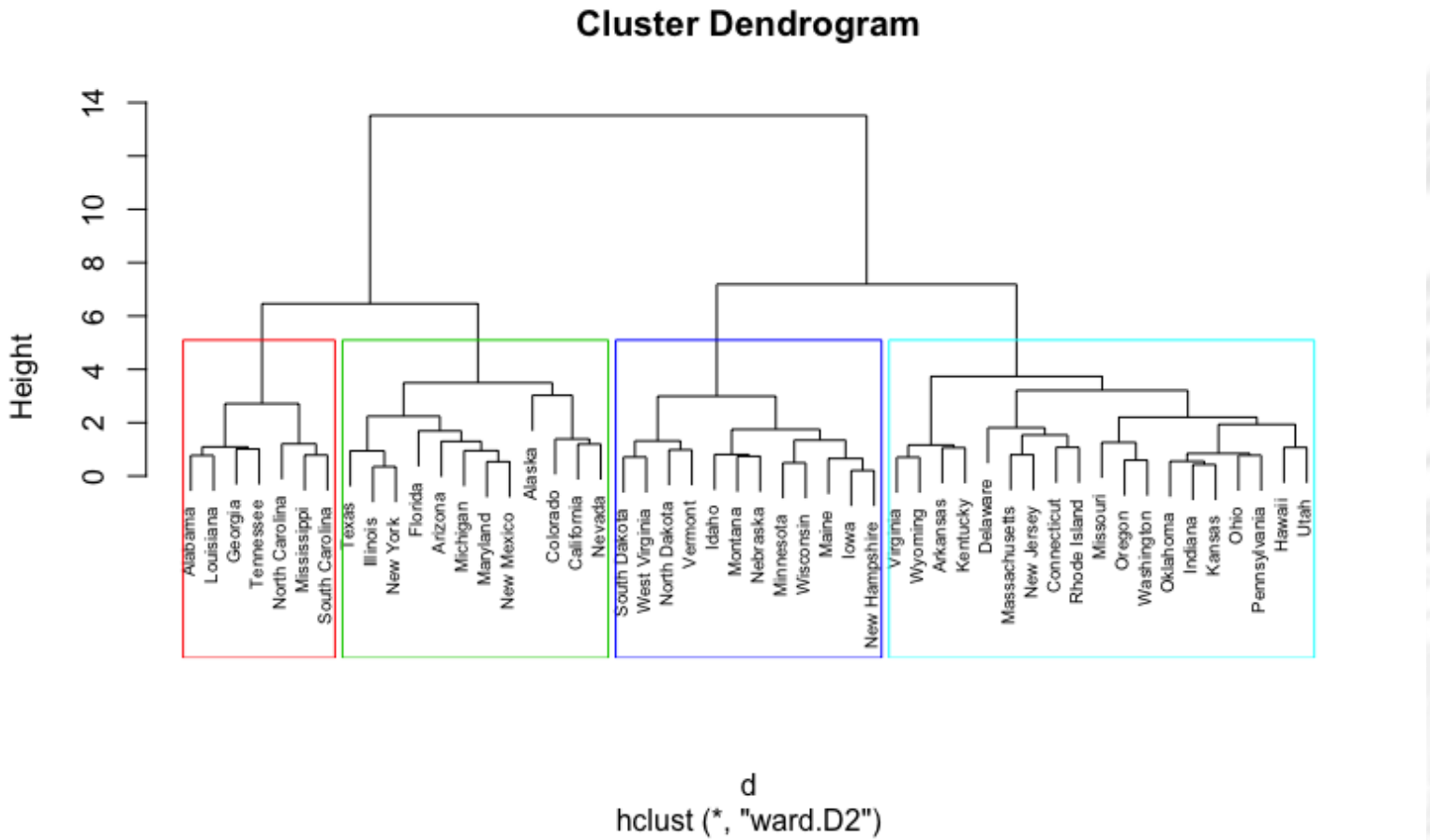


Figura 18. Dedrograma de agrupamento hierárquico. https://uc-r.github.io/public/images/analytics/clustering/hierarchical/unnamed-chunk-13-1.png.

Métodos aglomerativos

- Considere a soma de quadrados intra-cluster de A:

$$SQE_A = \sum_{i=1}^{n_A} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_A)'(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_A).$$

- Definimos o acréscimo na soma de quadrados resultante da junção de dois clusters A e B em um cluster AB por:

$$I_{AB} = SQE_{AB} - (SQE_A + SQE_B).$$

- Os clusters A e B que proporcionarem menor acréscimo na SQE é executada.

Determinação do número de clusters

- Uma das principais definições a se fazer, numa análise de clusters, é quanto ao **número de clusters** (K) que devem ser formados.
- Diferentes critérios podem ser adotados na determinação do **número ótimo** de clusters.
- Boa parte dos critérios baseiam-se na soma de quadrados intra-cluster total.

Determinação do número de clusters

- Num gráfico da soma de quadrados intra-cluster total *vs* número de clusters pode ajudar na escolha do número de clusters;
- O número de clusters a partir do qual a soma de quadrados intra-cluster total pouco reduzir, a cada novo cluster formado, é o número de clusters a ser escolhido.
- Na Figura ??, por exemplo, os resultados apontam a solução com $K = 3$ clusters, ou, eventualmente, $K = 4$.

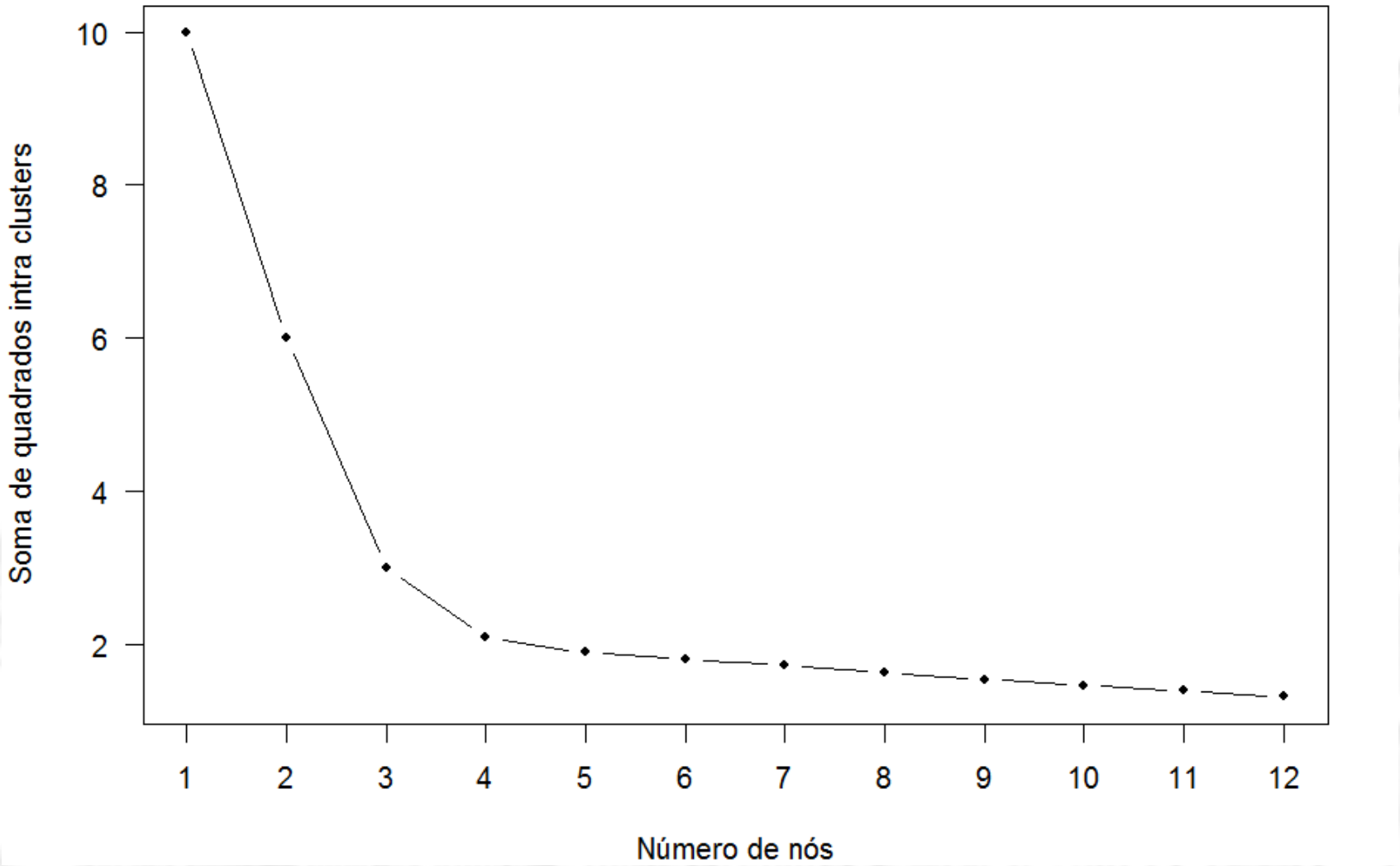


Figura 19. Escolha do número de cluster pela análise da soma de quadrados intra cluster.

Gráfico da silhueta

- A análise (gráfico) da silhueta é um método utilizado para interpretação e validação de uma análise de clusters.
- Consiste no cálculo e representação gráfica de uma medida de (boa) alocação de cada indivíduio ao respectivo cluster.
- Tomando a média dessas medidas em um particular cluster, tem-se uma medida de coesão do cluster.
- Tomando-se a média dessas medidas em toda a amostra, tem-se uma medida de consistência dos agrupamentos formados.

Gráfico da silhueta

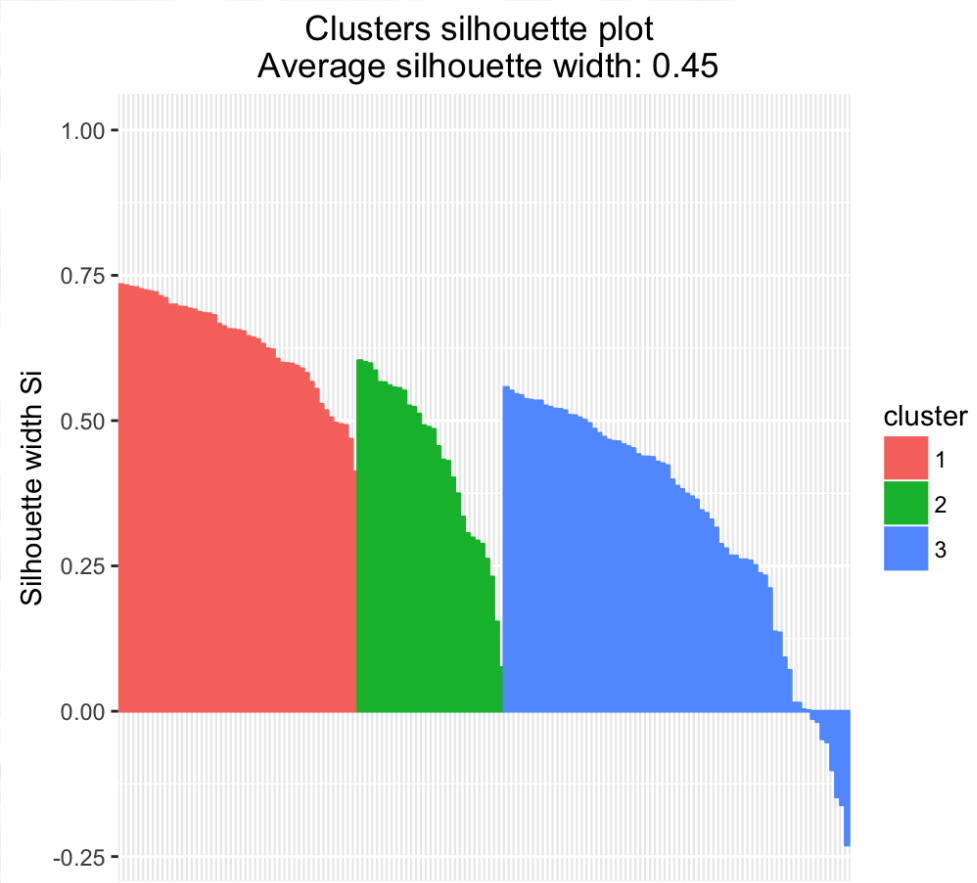


Figura 20. Gráfico da silhueta. http://www.sthda.com/sthda/RDoc/figure/clustering/cluster-analysis-in-r-silhouette-plot-1.png.

Gráfico da silhueta – Medida da silhueta

- Seja $a(i)$ a distância média de um elemento i em relação a todos os elementos do mesmo cluster ao qual ele foi alocado;
- Seja $d(i, B)$ a distância média do elemento i aos elementos de um cluster B, diferente daquele ao qual o elemento i foi alocado;
- Seja $b(i)$ o menor valor dos $d(i, B)'s$, calculados para todos os clusters exceto aquele que contém i.

Gráfico da silhueta - Medida da silhueta

- Define-se a medida da silhueta por:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}, \text{ para } i = 1, 2, \ldots, n.$$

- Repare, pela definição, que $-1 < s(i) < 1$.

Gráfico da silhueta - Medida da silhueta

- Se $a(i) <<< b(i)$, $s(i) \approx 1$, indicando que i é muito menos dissimilar dos elementos de seu grupo do que dos elementos dos outros grupos (ou seja, i está bem alocado);
- Se $a(i) >>> b(i)$, $s(i) \approx -1$, indicando que i é muito mais dissimilar dos elementos de seu grupo do que dos elementos do grupo vizinho (ou seja, i está mal alocado);
- Se $a(i) \approx b(i)$, $s(i) \approx 0$, indicando que i está na fronteira de seu grupo e de um grupo vizinho.

Outros métodos de agrupamento

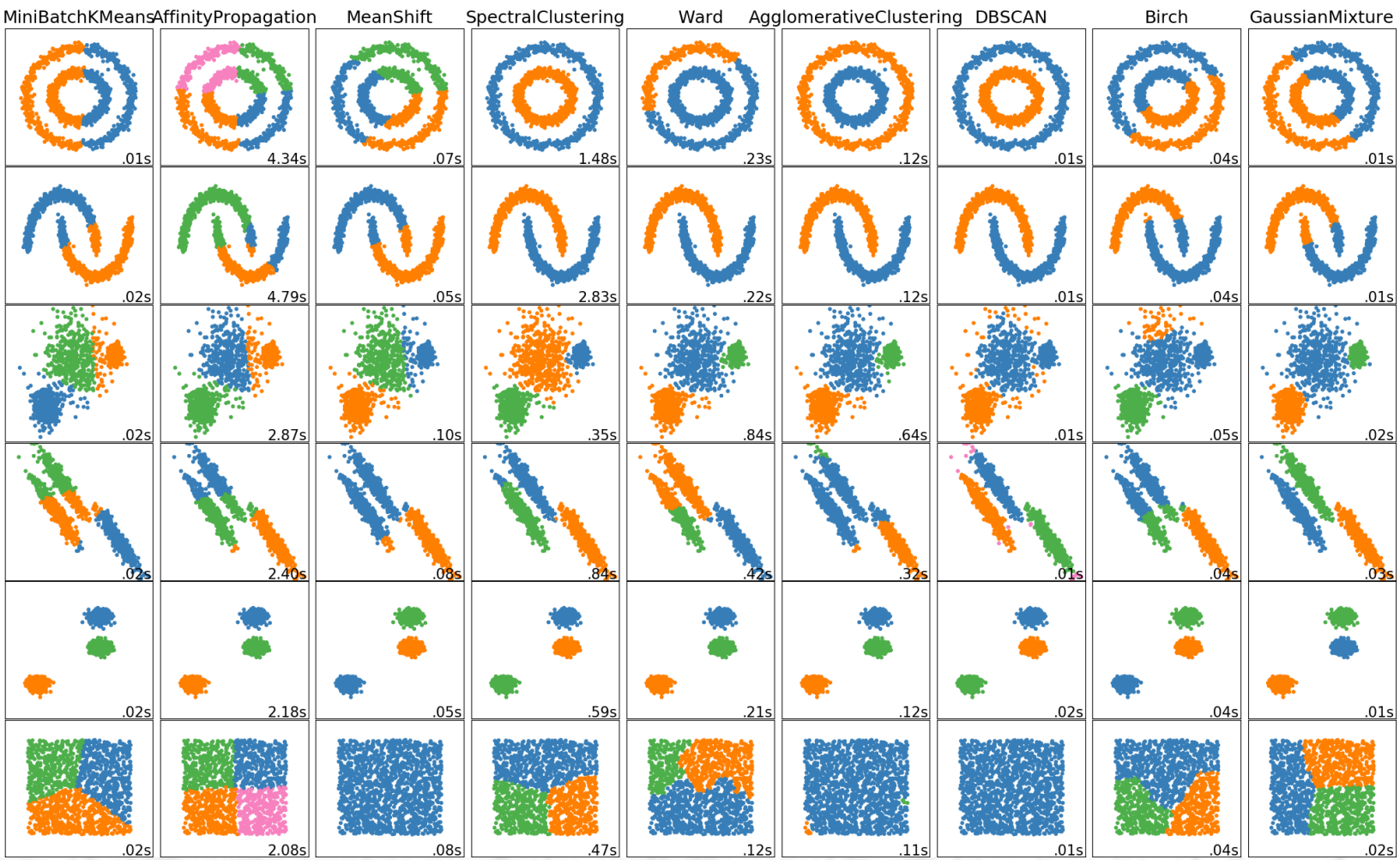
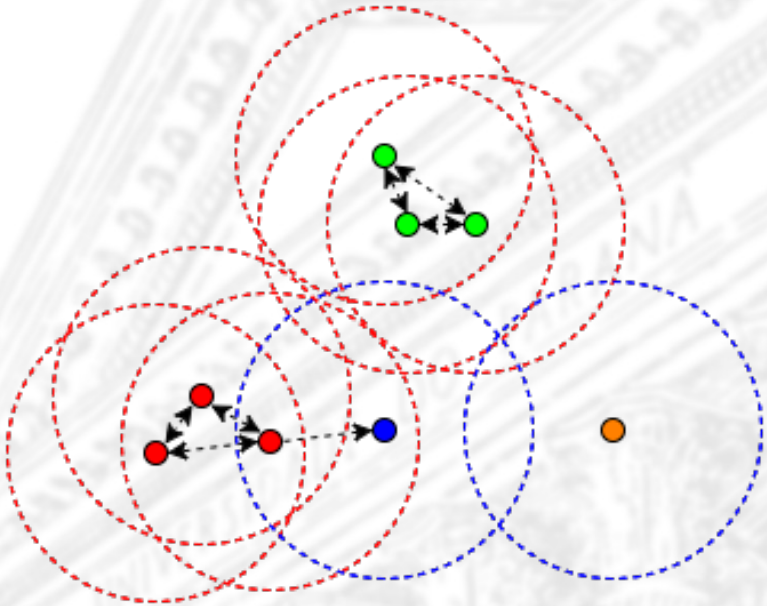


Figura 21. Métodos de agrupamento.

DBSCAN



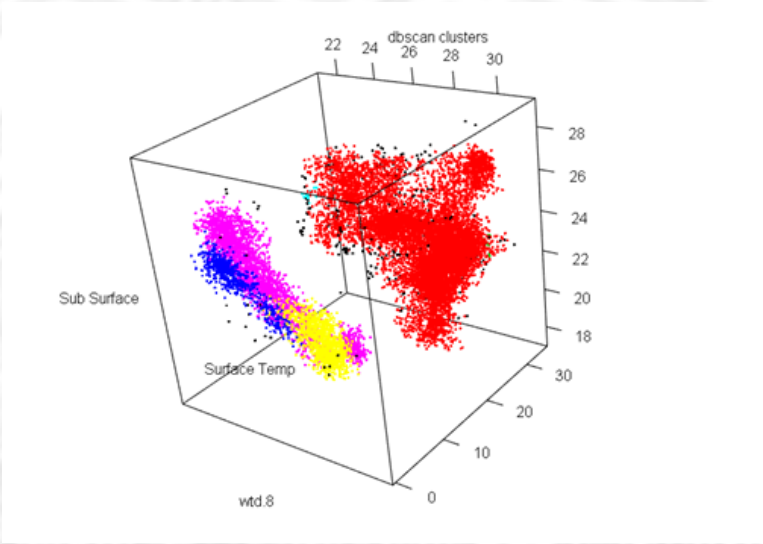
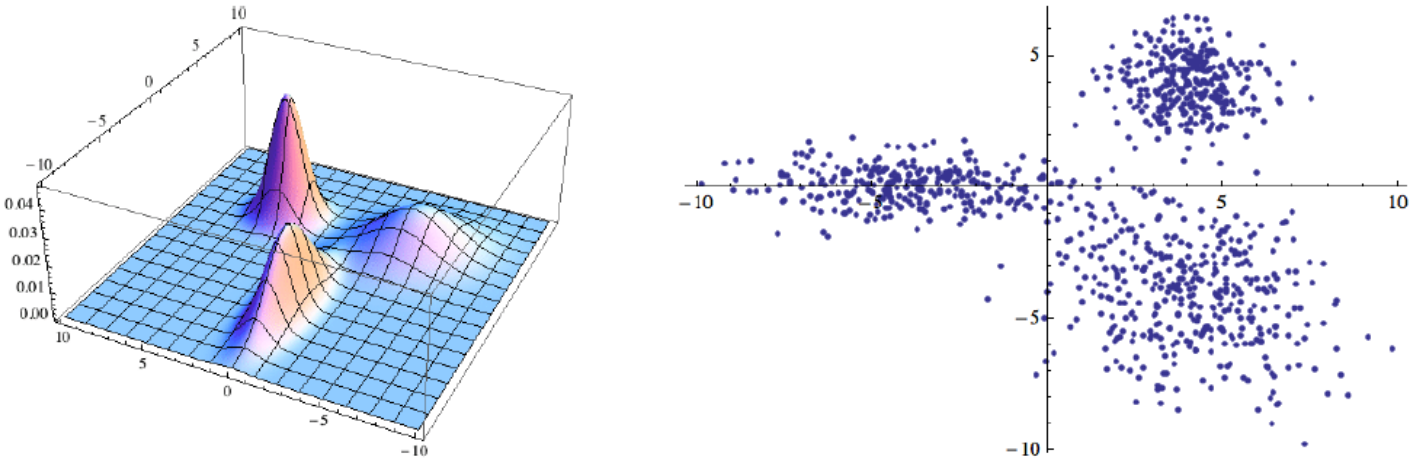


Figura 22. Ilustração do funcionamento do DBSCAN.

Modelos de mistura Gaussiana



(a) A probability distribution on $\mathbb{R}^2$.

(b) Data sampled from this distribution.

Figura 23. Modelos de mistura Gaussiana.

Agrupamento com restrição espacial

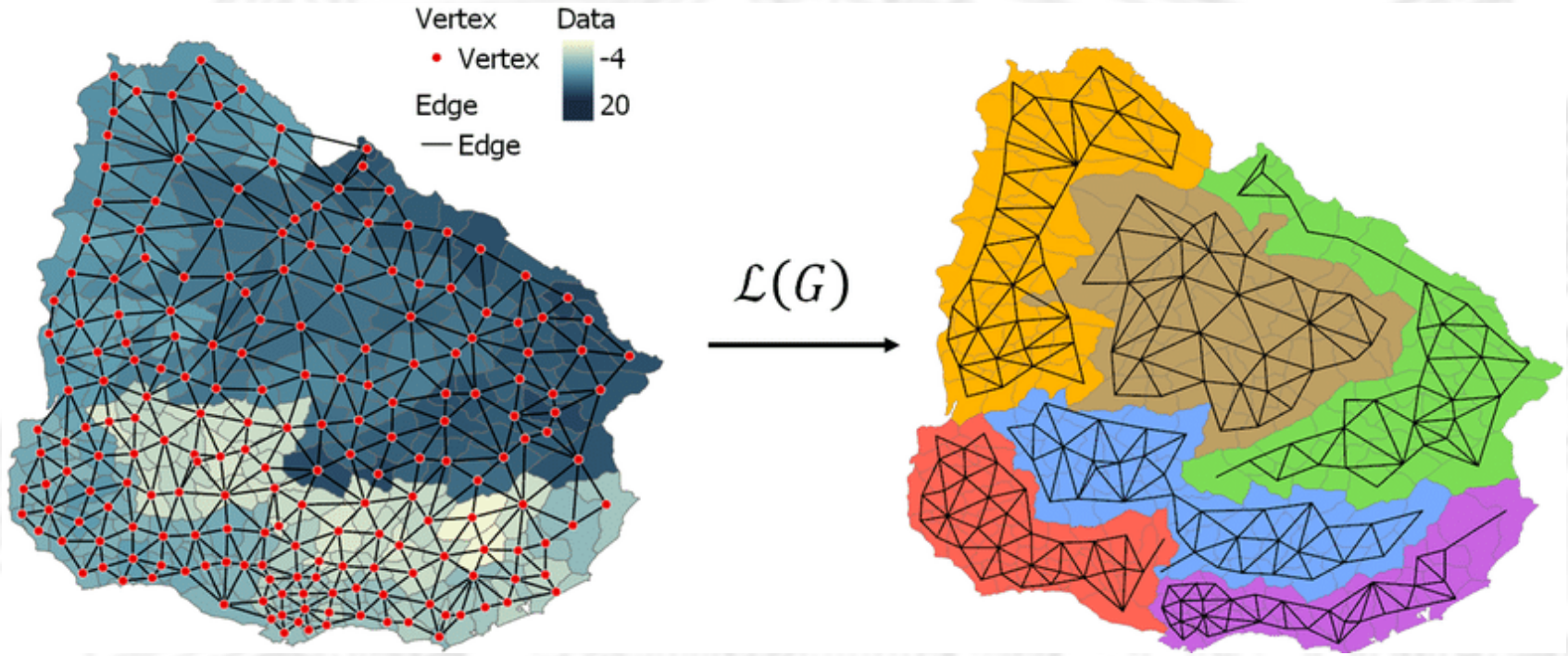


Figura 24. Agrupamento para dados com restrição espacial.

Considerações finais

Principais pontos

- Utilidade prática dos métodos de agrupamento.
- Tipos de agrupamento.
  - Não hierárquico.
  - Hierárquico.
- Número ótimo de clusters.
- Medidas de qualidade do agrupamento.
  - Gráfico da silhueta.
  - Estatística GAP.