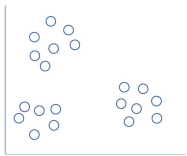
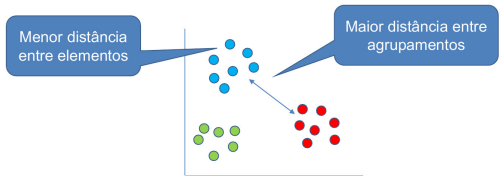
Análise de Agrupamentos

O que é clustering?

É uma **técnica estatística multivariada** para identificar agrupamentos dos dados de acordo com o grau de semelhança. Queremos achar um grupo de objetos, que são similares entre si e diferentes de outros grupos.

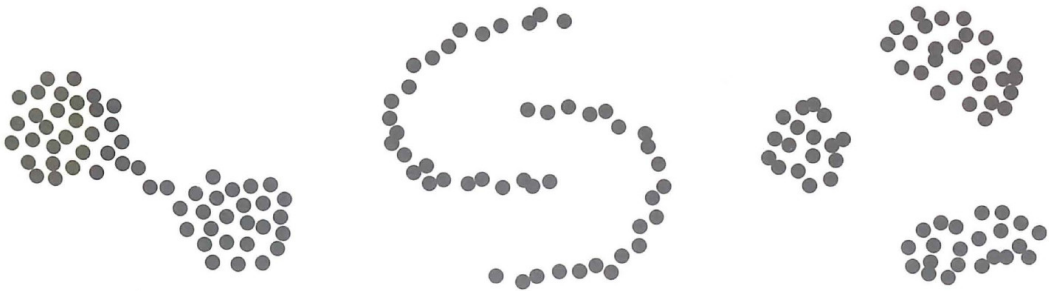




- O **objetivo** típico em clustering é descobrir os "agrupamentos naturais" presente nos dados.

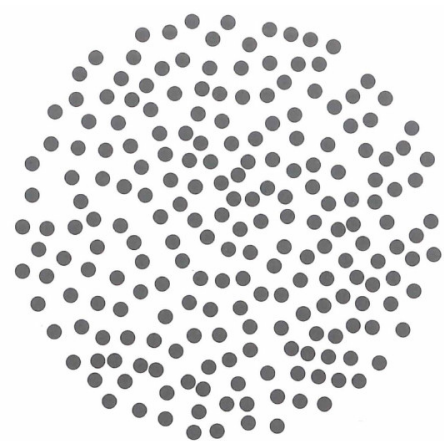
O que é clustering?

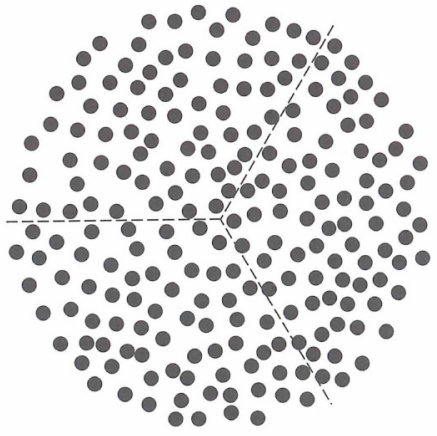
- Conjunto de dados **contendo** agrupamentos "naturais":



O que é clustering?

- Conjunto de dados **sem** agrupamentos "naturais":





Tendência de agrupamentos: “Faz sentido aplicar um algoritmo de clusterização nesses dados?”

Método estatístico: Estatística de Hopkins

A **Estatística de Hopkins** é uma medida pré-clusterização muito usada para avaliar se um conjunto de dados apresenta tendência natural à formação de agrupamentos (clusters) ou se os pontos estão distribuídos de forma aproximadamente aleatória no espaço das variáveis.

Tendência de agrupamentos: "Faz sentido aplicar um algoritmo de clusterização nesses dados?"

Método estatístico: Estatística de Hopkins

- A ideia central é comparar distâncias:
  - Distâncias entre pontos reais do conjunto de dados
  - Distâncias entre pontos artificiais gerados aleatoriamente no mesmo espaço amostral
- Se os dados tiverem estrutura de clusters, os pontos reais tendem a estar mais próximos entre si do que pontos gerados aleatoriamente.

Tendência de agrupamentos: "Faz sentido aplicar um algoritmo de clusterização nesses dados?"

Método estatístico: Estatística de Hopkins

- Considerando um inteiro $m \ll n$, de pontos amostrados (normalizados), calcula-se a distância de cada ponto até o vizinho mais próximo, excluindo o próprio ponto

$$w_i = \min_{x_j \in X \setminus \{x_i\}} d(x_i, x_j), \qquad i = 1, \ldots, m$$

Tendência de agrupamentos: "Faz sentido aplicar um algoritmo de clusterização nesses dados?"

Método estatístico: Estatística de Hopkins

- Geram-se m pontos artificiais $y_1, \ldots, y_m$, assumindo distribuição uniforme no hipercubo definido pelos limites dos dados e independência entre as dimensões. Para cada ponto artificial j, calcula-se a distância até o vizinho mais próximo no conjunto real.

$$u_i = \min_{x \in X} d(y_i, x), \qquad i = 1, \ldots, m$$

Tendência de agrupamentos: "Faz sentido aplicar um algoritmo de clusterização nesses dados?"

Método estatístico: Estatística de Hopkins

$$H = \frac{\sum_{i=1}^{m} u_i}{\sum_{i=1}^{m} u_i + \sum_{i=1}^{m} w_i}$$

- w_i: distância do ponto real ao vizinho mais próximo
- u_i: distância do ponto artificial ao ponto real mais próximo

Tendência de agrupamentos: "Faz sentido aplicar um algoritmo de clusterização nesses dados?"

Interpretação da Estatística de Hopkins

Sob a hipótese de ausência de clusters (padrão de referência), espera-se que as distâncias u_i e w_i sejam, em média, comparáveis, o que leva a:

$$H \approx 0.5$$

Em termos práticos:

- $H \approx 0.5$: dados compatíveis com **padrão aleatório** (baixa tendência à clusterização);
- $H > 0.5$: evidência de **tendência à formação de clusters** (quanto mais próximo de 1, mais agregado);
- $H < 0.5$: padrão mais **regular ou espalhado** do que o aleatório.

Tendência de agrupamentos: "Faz sentido aplicar um algoritmo de clusterização nesses dados?"

Interpretação da Estatística de Hopkins

Observação: *como H depende de amostragem aleatória (seleção dos m pontos e geração dos pontos artificiais), é recomendável repetir o cálculo várias vezes e resumir o resultado por média ou mediana.*

Tendência de agrupamentos

Método visual: Avaliação visual da tendência de agrupamento (VAT)

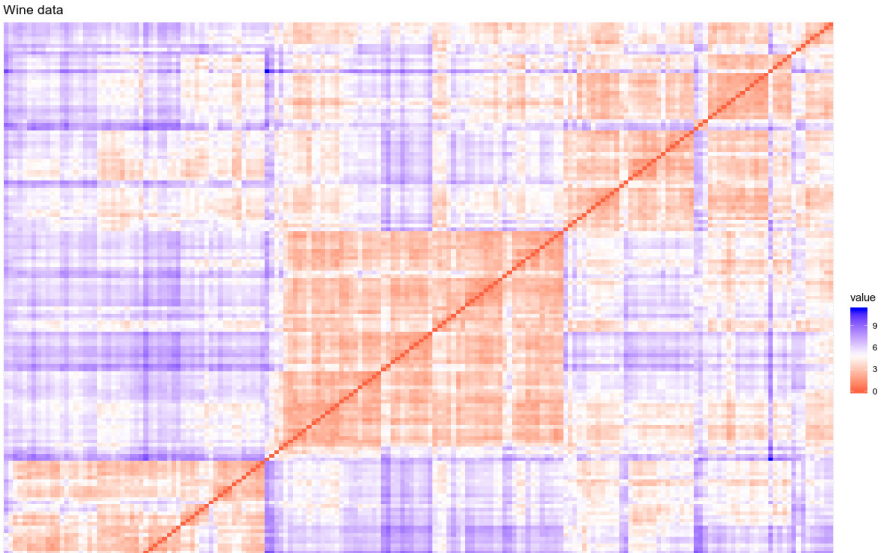
- A Avaliação Visual da Tendência de Agrupamento (VAT) é um método exploratório que avalia, de forma gráfica, se um conjunto de dados apresenta evidência de agrupamentos naturais.
- O procedimento baseia-se na matriz de distâncias entre as observações. Após o cálculo da matriz

$$D = [d(x_i, x_j)]_{n \times n},$$

as observações são reordenadas de modo que pontos similares fiquem próximos, e a matriz resultante é exibida como um mapa de calor.

Tendência de agrupamentos

Método visual: Avaliação visual da tendência de agrupamento (VAT)



Tendência de agrupamentos

Método visual: Avaliação visual da tendência de agrupamento (VAT)

- A interpretação é feita visualmente:
  - blocos escuros ao longo da diagonal indicam possível presença de clusters;
  - ausência de blocos bem definidos sugere padrão aleatório ou ausência de estrutura de agrupamento.
- O VAT é uma ferramenta pré-clusterização e deve ser utilizado de forma complementar a medidas quantitativas, como a Estatística de Hopkins.

Que perguntas a clusterização pode responder?

A clusterização é uma técnica de aprendizado não supervisionado cujo objetivo é identificar grupos de observações semelhantes, sem rótulos prévios. Ela é usada tanto como **objetivo final** quanto como **ferramenta de apoio à decisão**.

Que perguntas a clusterização pode responder?

A clusterização é uma técnica de aprendizado não supervisionado cujo objetivo é identificar grupos de observações semelhantes, sem rótulos prévios. Ela é usada tanto como **objetivo final** quanto como **ferramenta de apoio à decisão**.

- **Como agrupo os clientes que compram no site?**
  A clusterização permite identificar perfis de clientes com comportamentos de compra semelhantes, apoiando estratégias de segmentação e marketing.

Que perguntas a clusterização pode responder?

A clusterização é uma técnica de aprendizado não supervisionado cujo objetivo é identificar grupos de observações semelhantes, sem rótulos prévios. Ela é usada tanto como **objetivo final** quanto como **ferramenta de apoio à decisão**.

- **Como agrupo os clientes que compram no site?**
  A clusterização permite identificar perfis de clientes com comportamentos de compra semelhantes, apoiando estratégias de segmentação e marketing.

- **Onde coloco uma antena de operadora de celular?**
  A partir da clusterização espacial de usuários ou regiões, é possível identificar áreas de alta concentração de demanda, sugerindo locais estratégicos para instalação.

Que perguntas a clusterização pode responder?

- **Onde deixo uma ambulância estacionada para atender uma emergência?**
  A clusterização de ocorrências passadas ajuda a identificar zonas críticas, auxiliando na alocação eficiente de recursos de emergência.

Que perguntas a clusterização pode responder?

- **Onde deixo uma ambulância estacionada para atender uma emergência?**
  A clusterização de ocorrências passadas ajuda a identificar zonas críticas, auxiliando na alocação eficiente de recursos de emergência.

- **Como classifico cervejas em categorias distintas?**
  A clusterização explora similaridades entre produtos e pode revelar categorias naturais ou estilos semelhantes, apoiando a construção de classificações.

Algumas questões...

Como medir a homogeneidade entre indivíduos?

E entre grupos de indivíduos?

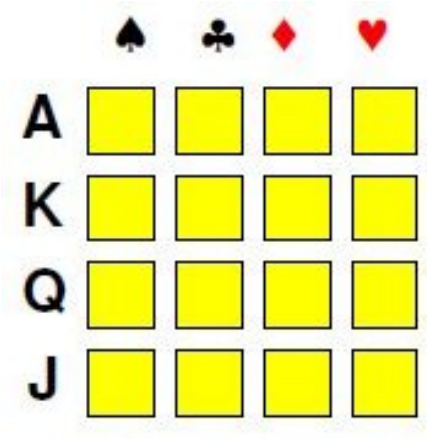
Dado um conjunto de indivíduos, quantos grupos posso formar?

Algumas questões...

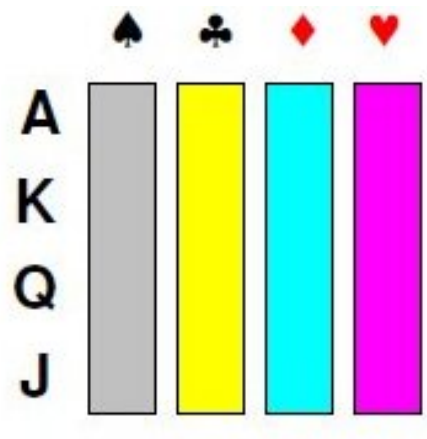
Imagine que você tenha 16 cartas figuradas (A, K, Q, J) e que queira formar grupos de cartas semelhantes...

Como você formaria esses grupos?

Algumas questões...

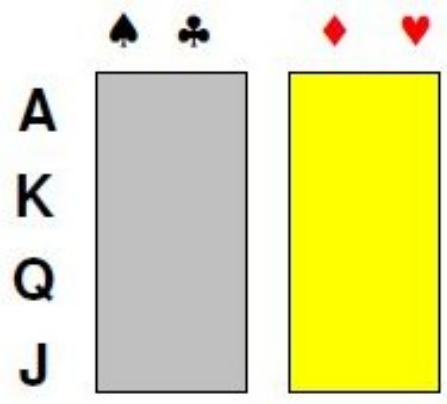


(a) Cartas individuais?

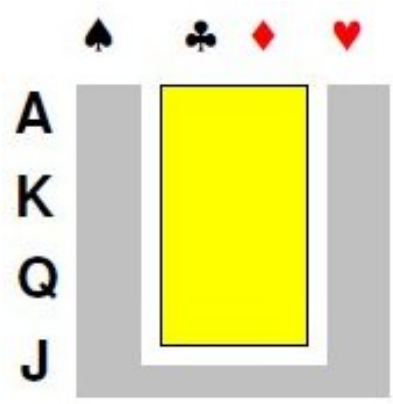


(a) Agrupar por naipes?

Algumas questões...

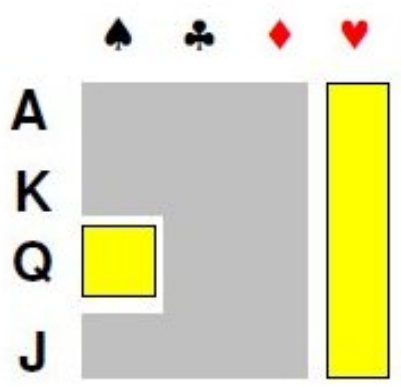


(a) Agrupar por cor do naipe?



(a) Agrupar por naipes maiores e menores?

Algumas questões...



(a) Copas mais a Rainha de espadas e outros naipes?

(a) Agrupar por face da carta?

Algumas questões...

Resumindo...

Necessidade da definição de medidas de **similaridade** (ou dissimilaridade)

Similaridade e dissimilaridade

Medidas de Similaridade: quanto **maior** o valor, **maior a semelhança** entre os objetos

Similaridade e dissimilaridade

Medidas de Similaridade: quanto **maior** o valor, **maior a semelhança** entre os objetos

Medidas de Dissimilaridade (Distância): quanto **maior** o valor, **mais diferentes** são os objetos

Similaridade e dissimilaridade (exemplo)

Pesquisa com clientes de uma loja de equipamentos automotivos

Similaridade e dissimilaridade (exemplo)

Pesquisa com clientes de uma loja de equipamentos automotivos

Variáveis mensuradas

- **Idade** (em anos completos) - Variável quantitativa discreta
- **Número de carros** - Variável quantitativa discreta
- **Classe social**: A, B, C ou D - Variável qualitativa ordinal
- **Potência do motor**: Baixa, Média ou Alta - Variável qualitativa ordinal
- **Combustível**: Gasolina ou Álcool - Variável qualitativa nominal
- **Modelo**: Esporte, Luxo ou Standard - Variável qualitativa nominal

Base de dados (exemplo)

Cliente	**Idade**	**N.º de carros**	**Classe social**	**Potência do motor**	**Combustíve**	**Modelo**
1	20	1	A	Baixa	Gasolina	Esporte
2	37	2	A	Alta	Gasolina	Luxo
3	51	1	C	Média	Gasolina	Esporte
4	32	1	D	Alta	Álcool	Standard
5	30	2	B	Média	Álcool	Standard
6	55	3	A	Alta	Gasolina	Luxo

Base de dados (exemplo)

Cliente	**Idade**	**N.º de carros**	**Classe social**	**Potência do motor**	**Combustíve**	**Modelo**
1	20	1	A	Baixa	Gasolina	Esporte
2	37	2	A	Alta	Gasolina	Luxo
3	51	1	C	Média	Gasolina	Esporte
4	32	1	D	Alta	Álcool	Standard
5	30	2	B	Média	Álcool	Standard
6	55	3	A	Alta	Gasolina	Luxo

Como medir a similaridade ou dissimilaridade entre os indivíduos?

Variáveis quantitativas

Dissimilaridade

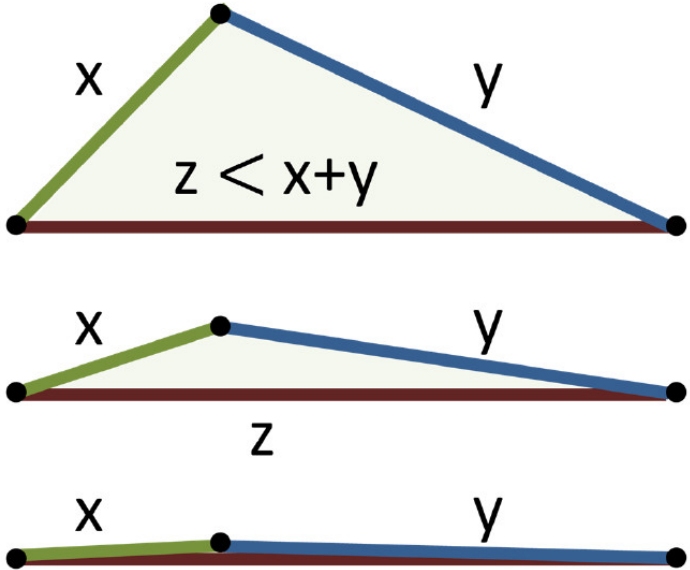
As **distâncias** são as medidas de **dissimilaridade** mais utilizadas no estudo de bancos de dados com variáveis numéricas

Dissimilaridade

As **distâncias** são as medidas de **dissimilaridade** mais utilizadas no estudo de bancos de dados com variáveis numéricas

- Uma medida d_{ij} representa uma **medida de distância** entre os indivíduos i e j se, e somente se,
  a) $d_{ij} \geqslant 0$ para todo i e j;
  b) $d_{ij} = 0$ se, e somente se, $i = j$;
  c) $d_{ij} = d_{ji}$
  d) $d_{ij} \leqslant d_{ik} + d_{kj}$ para qualquer indivíduo k.

Dissimilaridade



Dissimilaridade

Principais medidas de distância

Distância Euclidiana

$$d_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^t(\mathbf{x}_i - \mathbf{x}_j)} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$$

Distância geométrica entre dois pontos

Dissimilaridade

Principais medidas de distância

Distância Euclidiana Generalizada

$$d_{ij} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^t W (\mathbf{x}_i - \mathbf{x}_j)}$$

- Se $W = \text{diag}\left(\frac{1}{p}\right)$: **distância euclidiana média**
- Se $W = \Sigma^{-1}$: **distância de Mahalanobis**

Dissimilaridade

Principais medidas de distância

Distância de Minkowski

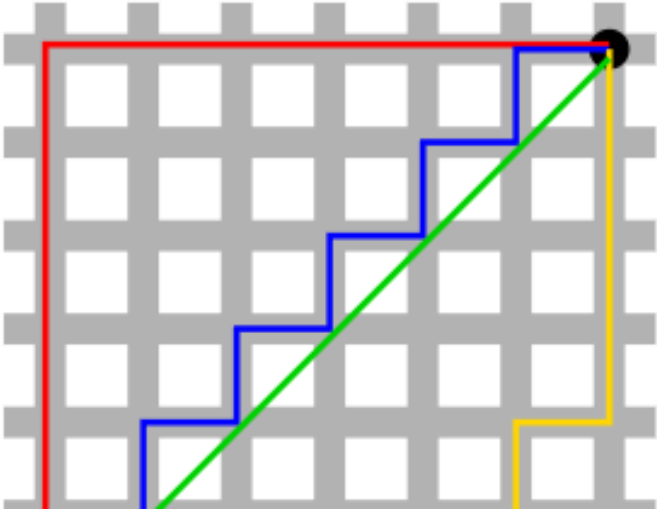
$$d_{ij} = \left(\sum_{k=1}^{P} |X_{ik} - X_{jk}|^{\lambda} \right)^{\frac{1}{\lambda}}$$

- Se $\lambda = 1$, temos a chamada **métrica de Manhattan**. É também conhecida como *city block*.
- Se $\lambda = 2$, temos a distância euclidiana.
- A métrica de Minkowski é menos afetada pela presença de valores discrepantes na amostra do que a distância Euclidiana.

Dissimilaridade

Principais medidas de distância

Distância de Manhattan (city block)



Dissimilaridade

Dissimilaridade $\Longrightarrow$ Similaridade

$$s_{ij} = 1 - d^0_{ij}$$

em que

$$d^0_{ij} = \frac{d_{ij} - \min(D)}{\max(D) - \min(D)}$$

Sendo $\min(D)$ e $\max(D)$ o menor e o maior valor de distância observados na matriz de distâncias $D_{n \times n}$, sem levar em consideração os elementos da diagonal principal dessa matriz.

Variáveis qualitativas

Similaridade

Variáveis qualitativas nominais

Neste caso, utilizamos variáveis fictícias (variáveis *dummy*) para **codificar** as variáveis:

Combustível	N_1
Gasolina	1
Álcool	0

Modelo	N_2	N_3
Esporte	1	0
Luxo	0	1
Standard	0	0

Similaridade

Variáveis qualitativas nominais

No exemplo:

Cliente	Combustível	N_1
1	Gasolina	1
2	Gasolina	1
3	Gasolina	1
4	Álcool	0
5	Álcool	0
6	Gasolina	1

Cliente	Modelo	N_2	N_3
1	Esporte	1	0
2	Luxo	0	1
3	Esporte	1	0
4	Standard	0	0
5	Standard	0	0
6	Luxo	0	1

Similaridade

Variáveis qualitativas nominais

De forma que,

Cliente	N_1	N_2	N_3
1	1	1	0
2	1	0	1
3	1	1	0
4	0	0	0
5	0	0	0
6	1	0	1

Similaridade

Variáveis qualitativas ordinais

Utilizamos variáveis fictícias (variáveis *dummy*) para **codificar** as variáveis, levando em consideração a ordinalidade das variáveis:

Classe social	O_1	O_2	O_3
A	1	1	1
B	0	1	1
C	0	0	1
D	0	0	0

Potência	O_4	O_5
Alta	1	1
Média	0	1
Baixa	0	0

Similaridade

Observação

As variáveis **Classe social** e **Potência do motor** possuem **ordem natural**. Por isso, foram codificadas por meio de **variáveis ordinais cumulativas**:

- valores mais altos acumulam mais indicadores iguais a 1;
- valores mais baixos recebem apenas zeros.

Essa codificação preserva a informação de ordem, mas impõe uma estrutura geométrica específica ao espaço dos dados.

Similaridade

Variáveis qualitativas ordinais

No exemplo:

Cliente	Classe social	O_1	O_2	O_3
1	A	1	1	1
2	A	1	1	1
3	C	0	0	1
4	D	0	0	0
5	B	0	1	1
6	A	1	1	1

Cliente	Potência	O_4	O_5
1	Baixa	0	0
2	Alta	1	1
3	Média	0	1
4	Alta	1	1
5	Média	0	1
6	Alta	1	1

Similaridade

Variáveis qualitativas ordinais

De forma que,

Cliente	O_1	O_2	O_3	O_4	O_5
1	1	1	1	0	0
2	1	1	1	1	1
3	0	0	1	0	1
4	0	0	0	1	1
5	0	1	1	0	1
6	1	1	1	1	1

Similaridade

Variáveis qualitativas nominais e ordinais

Juntando tudo...

Cliente	O_1	O_2	O_3	O_4	O_5	N_1	N_2	N_3
1	1	1	1	0	0	1	1	0
2	1	1	1	1	1	1	0	1
3	0	0	1	0	1	1	1	0
4	0	0	0	1	1	0	0	0
5	0	1	1	0	1	0	0	0
6	1	1	1	1	1	1	0	1

Similaridade

Variáveis qualitativas nominais e ordinais

Considere os clientes 1 e 3. Vamos calcular a medida de similaridade entre eles:

Cliente	O_1	O_2	O_3	O_4	O_5	N_1	N_2	N_3
1	1	1	1	0	0	1	1	0
3	0	0	1	0	1	1	1	0

Similaridade

Tabela de concordância entre dois indivíduos

Indivíduo i \ Indivíduo j	1	0	Total
1	a	b	$a+b$
0	c	d	$c+d$
Total	$a+c$	$b+d$	p

- a: número de variáveis em que ambos têm valor 1
- b: número de variáveis em que $i = 1$ e $j = 0$
- c: número de variáveis em que $i = 0$ e $j = 1$
- d: número de variáveis em que ambos têm valor 0

$$p = a + b + c + d$$

Similaridade

Tabela de concordância entre dois indivíduos

No exemplo:

Indivíduo 1 \ Indivíduo 3	1	0	Total
1	3	2	5
0	1	2	3
Total	4	4	8

Como extrair as **medidas de similaridade** baseadas na **tabela de concordância** dos elementos?

Critérios de similaridade

Coeficiente de Concordância Simples

$$s_{ij} = \frac{a+d}{p}$$

Indivíduo 1 \ Indivíduo 3	1	0	Total
1	3	2	5
0	1	2	3
Total	4	4	8

$$s_{13} = \frac{a+d}{p} = \frac{3+2}{8} = \frac{5}{8} = 0,625$$

Critérios de similaridade

Coeficiente de Concordância Simples

Interpretação:

- O valor 0,625 indica que os indivíduos 1 e 3 concordam em 62,5% das variáveis binárias.
- Tanto concordâncias em 1 quanto em 0 contribuem igualmente para a similaridade.

Critérios de similaridade

Coeficiente de Concordância Positiva

$$s_{ij} = \frac{a}{p}$$

Indivíduo 1 \ Indivíduo 3	1	0	Total
1	3	2	5
0	1	2	3
Total	4	4	8

$$s_{13} = \frac{a}{p} = \frac{3}{8} = 0,375$$

Critérios de similaridade

Coeficiente de Concordância Positiva

Interpretação:

- Os indivíduos 1 e 3 compartilham simultaneamente o valor 1 em **37,5%** das variáveis.
- Ausências simultâneas (valor 0) **não contribuem** para a similaridade.
- O critério enfatiza **presença conjunta**, não coincidência geral.

Critérios de similaridade

Coeficiente de Concordância de Jaccard

$$s_{ij} = \frac{a}{a+b+c}$$

Indivíduo 1 \ Indivíduo 3	1	0	Total
1	3	2	5
0	1	2	3
Total	4	4	8

$$s_{13} = \frac{3}{3+2+1} = \frac{3}{6} = 0,5$$

Critérios de similaridade

Coeficiente de Concordância de Jaccard

Interpretação:

- Os indivíduos 1 e 3 compartilham **50% das presenças observadas**.
- Ausências simultâneas (valor 0) **não contribuem** para a similaridade.
- O coeficiente foca exclusivamente na **interseção de atributos presentes**.

Critérios de similaridade

Coeficiente de Concordância de Gower e Legendre

O **Coeficiente de Concordância de Gower–Legendre** é uma medida de similaridade projetada para **dados mistos**, isto é, conjuntos de dados que contêm, ao mesmo tempo:

- variáveis quantitativas,
- variáveis qualitativas nominais,
- variáveis qualitativas ordinais.

Ele permite combinar diferentes tipos de variáveis em **uma única medida de similaridade**, respeitando a natureza de cada uma.

Critérios de similaridade

Coeficiente de Concordância de Gower e Legendre

Considere dois indivíduos i e j descritos por p variáveis. A similaridade de Gower–Legendre é definida como:

$$S_{ij} = \frac{\sum_{k=1}^{p} w_{ijk}\, s_{ijk}}{\sum_{k=1}^{p} w_{ijk}},$$

onde:

- s_{ijk} é a similaridade parcial da variável k entre os indivíduos i e j;
- w_{ijk} é um peso (geralmente 0 ou 1), usado para tratar dados ausentes.

Critérios de similaridade

Coeficiente de Concordância de Gower e Legendre

O coeficiente assume valores em $[0, 1]$:

- $S_{ij} = 1$ indica indivíduos idênticos;
- $S_{ij} = 0$ indica máxima dissimilaridade.

Critérios de similaridade

Coeficiente de Concordância de Gower e Legendre

Similaridade parcial por tipo de variável

Variáveis quantitativas

- Para uma variável quantitativa k:

$$s_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{R_k},$$

onde R_k é o intervalo da variável k (máximo $-$ mínimo).

Critérios de similaridade

Coeficiente de Concordância de Gower e Legendre

Similaridade parcial por tipo de variável

Variáveis qualitativas nominais

- Para uma variável nominal:

$$s_{ijk} = \begin{cases} 1, & \text{se } x_{ik} = x_{jk}, \\ 0, & \text{caso contrário.} \end{cases}$$

Critérios de similaridade

Coeficiente de Concordância de Gower e Legendre

Similaridade parcial por tipo de variável

Variáveis qualitativas ordinais

Para variáveis ordinais, os níveis são primeiro convertidos em **postos normalizados** no intervalo $[0, 1]$. Em seguida, aplica-se a mesma fórmula das variáveis quantitativas:

$$s_{ijk} = 1 - |r_{ik} - r_{jk}|,$$

onde r_{ik} e r_{jk} são os postos normalizados.

Critérios de similaridade

Coeficiente de Concordância de Gower e Legendre

Interpretação:

- Cada variável contribui **de forma controlada** para a similaridade total.
- Variáveis com escalas diferentes **não dominam** o cálculo.
- A ordem das categorias é respeitada para variáveis ordinais, **sem impor distâncias artificiais**.

Critérios de similaridade

Similaridade $\Longrightarrow$ Dissimilaridade

$$d_{ij}^{*} = 1 - s_{ij}$$

Critérios de similaridade

Coeficiente de Concordância de Gower e Legendre

No exemplo:

Cliente	**Idade**	**N.º de carros**	**Classe social**	**Potência do motor**	**Combustíve**	**Modelo**
1	20	1	A	Baixa	Gasolina	Esporte
2	37	2	A	Alta	Gasolina	Luxo
3	51	1	C	Média	Gasolina	Esporte
4	32	1	D	Alta	Álcool	Standard
5	30	2	B	Média	Álcool	Standard
6	55	3	A	Alta	Gasolina	Luxo

Critérios de similaridade

Coeficiente de Concordância de Gower e Legendre

Temos que:

- **Cliente 1:** Idade = 20, Carros = 1, Classe = A, Potência = Baixa, Combustível = Gasolina, Modelo = Esporte
- **Cliente 3:** Idade = 51, Carros = 1, Classe = C, Potência = Média, Combustível = Gasolina, Modelo = Esporte

Critérios de similaridade

Coeficiente de Concordância de Gower e Legendre

Similaridades parciais

- **Idade**

$$s = 1 - \frac{|20 - 51|}{35} = 1 - \frac{31}{35} = \frac{4}{35} \approx 0,1143$$

- **Nº de carros**

$$s = 1 - \frac{|1 - 1|}{2} = 1$$

Critérios de similaridade

Coeficiente de Concordância de Gower e Legendre

Similaridades parciais

- **Classe social (ordinal)**

$r(A) = 1$ e $r(C) = \frac{1}{3}$:

$$s = 1 - \left|1 - \tfrac{1}{3}\right| = 1 - \tfrac{2}{3} = \tfrac{1}{3} \approx 0,3333$$

- **Potência (ordinal)**

$r(\text{Baixa}) = 0$ e $r(\text{Média}) = \frac{1}{2}$:

$$s = 1 - \left|0 - \tfrac{1}{2}\right| = \tfrac{1}{2} = 0,5$$

Critérios de similaridade

Coeficiente de Concordância de Gower e Legendre

Similaridades parciais

- **Combustível (nominal)**

Gasolina $=$ Gasolina $\Rightarrow s = 1$

- **Modelo (nominal)**

Esporte $=$ Esporte $\Rightarrow s = 1$

Critérios de similaridade

Coeficiente de Concordância de Gower e Legendre

Similaridade total

São $p = 6$ variáveis:

$$S_{1,3} = \frac{0,1143 + 1 + 0,3333 + 0,5 + 1 + 1}{6} \approx 0,6579$$

Métodos de Agrupamento

Métodos de Agrupamento

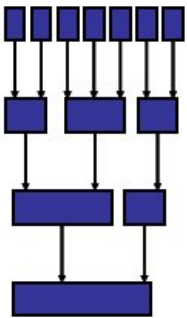
- **Métodos hierárquicos:** Baseiam-se na realização de sucessivas aglomerações ou de sucessivas divisões dos dados. Envolvem a construção de uma hierarquia através de uma estrutura do tipo árvore.

Métodos de Agrupamento

- **Métodos hierárquicos:** Baseiam-se na realização de sucessivas aglomerações ou de sucessivas divisões dos dados. Envolvem a construção de uma hierarquia através de uma estrutura do tipo árvore.
- **Métodos não-hierárquicos:** Produzem uma partição em um número fixo de classes, com sucessivas alocações e re-alocações dos indivíduos, conforme as distâncias aos demais indivíduos. Necessário definir o número de grupos à $\{priori\}$.

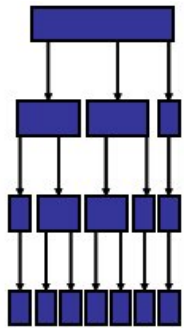
Métodos hierárquicos

- **Aglomerativos:** Os agrupamentos são formados a partir de uma matriz de parecença, que é atualizada a cada união de um par de objetos. Neste procedimento, cada indivíduo, originalmente, é um cluster, configurando n clusters. Indivíduos similares são sucessivamente agrupados, até a formação de um único grupo, contendo toda a amostra.



Métodos hierárquicos

- **Divisivos (ou de Partição):** Fazem o caminho oposto aos métodos hierárquicos aglomerativos. Neste método, um único grupo de objetos é subdividido em dois com a maior distância. Estes subgrupos são então particionados sucessivamente até se obter os objetos individuais.



Métodos de Agrupamento Hierárquicos

Agrupamentos hierárquicos aglomerativos

a) Considerar inicialmente n grupos, sendo n o número de indivíduos. A matriz de distâncias $D_{n\times n}$ é a matriz de distâncias entre os elementos originais;

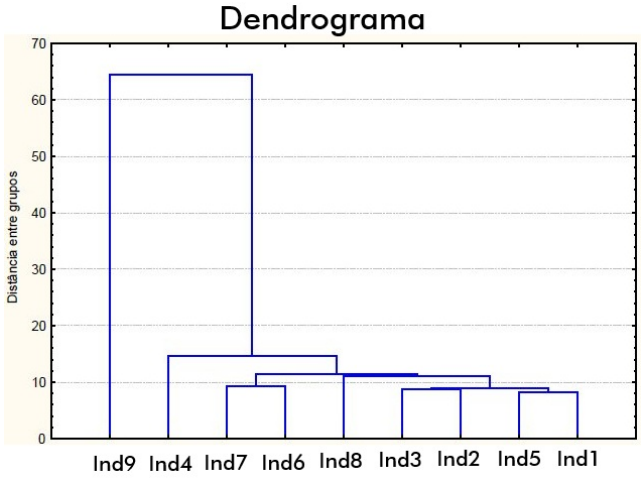
b) Selecionar os dois indivíduos mais próximos na matriz $D_{n\times n}$ e formar com eles um grupo;

c) Substituir os indivíduos utilizados no passo b) para definir o grupo por um novo elemento que represente o grupo construído. A distância entre esse novo elemento e os indivíduos restantes são calculadas utilizando um dos critérios que serão definidos a seguir;

d) Voltar ao passo b) e repetir os passos b) e c) até que tenhamos todos os elementos agrupados em um único grupo.

Métodos de Agrupamento Hierárquicos

Resultado da aplicação do método



- Representa uma síntese gráfica do método de agrupamento
- Esse gráfico é de grande utilidade para a classificação, comparação e discussão de agrupamentos.

Métodos de Agrupamento Hierárquicos

Como definir distância entre grupos?

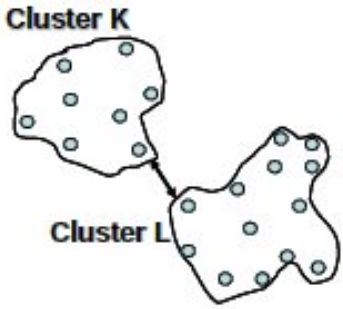
- Suponha que temos um grupo K com n_k indivíduos e um grupo L com n_l indivíduos.
- A distância entre os grupos K e L pode ser calculada com base em um dos cinco métodos seguintes:
  - Método do vizinho mais próximo (*single linkage*)
  - Método do vizinho mais distante (*complete linkage*)
  - Método da distância média (*average linkage*)
  - Método do centroide (*Centroid*)
  - Método de Ward

Métodos de Agrupamento Hierárquicos

Método do vizinho mais próximo

Consiste em considerar que a distância entre os dois grupos é a **menor distância entre as possíveis combinações de indivíduos** tomados dos dois grupos considerados, isto é,

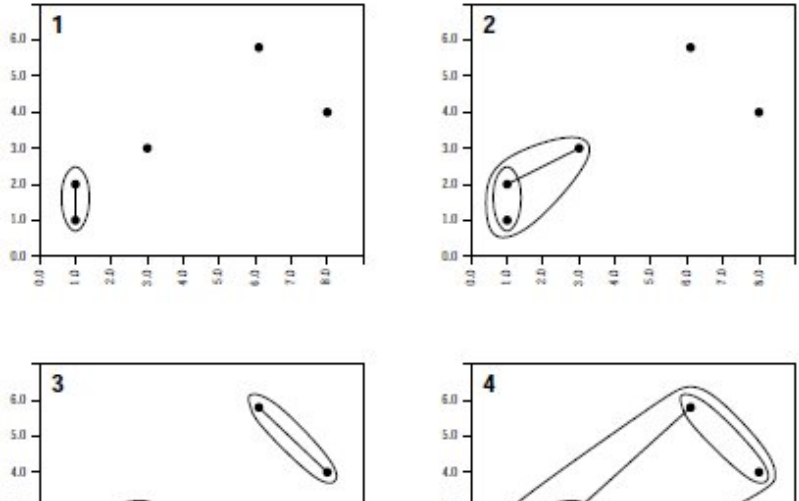
$$d_{(K,L)} = \underbrace{\min(d_{ij})}_{i \in K, j \in L}$$



Métodos de Agrupamento Hierárquicos

Método do vizinho mais próximo

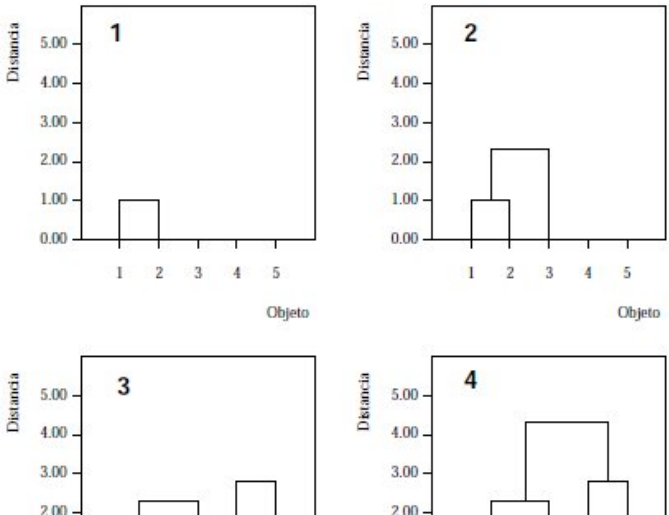
Esquematicamente:



Métodos de Agrupamento Hierárquicos

Método do vizinho mais próximo

Esquematicamente:

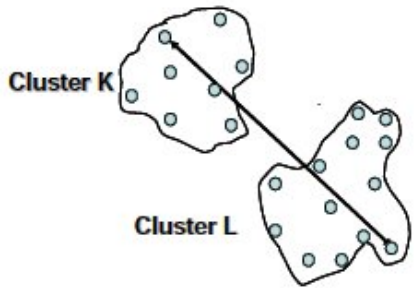


Métodos de Agrupamento Hierárquicos

Método do vizinho mais distante

Consiste em considerar que a distância entre os dois grupos é a **maior distância entre as possíveis combinações de indivíduos** tomados dos dois grupos considerados, isto é,

$$d_{(K,L)} = \underbrace{\max(d_{ij})}_{i \in K, j \in L}$$

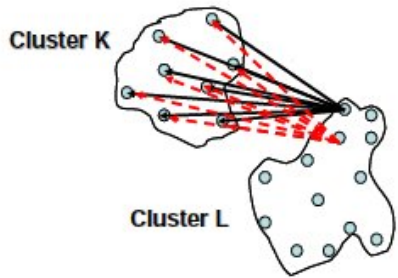


Métodos de Agrupamento Hierárquicos

Método da distância média

Consiste em considerar que a distância entre os dois grupos é a **média aritmética das distâncias entre as possíveis combinações de indivíduos** tomados dos dois grupos considerados, isto é,

$$d_{(K,L)} = \sum_{i \in K} \sum_{j \in L} \frac{d_{ij}}{n_k n_l}$$



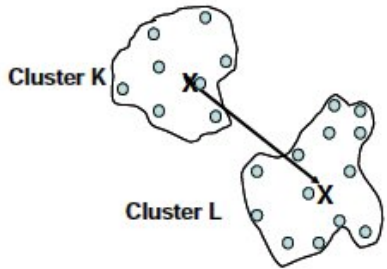
Métodos de Agrupamento Hierárquicos

Método do centroide

Consiste em considerar que a distância entre os dois grupos é a **distância euclidiana ao quadrado entre os centroides** dos dois grupos. O centroide de um grupo é o ponto médio dos objetos contidos no grupo, isto é,

$$d_{(K,L)} = (\bar{K} - \bar{L})^t(\bar{K} - \bar{L})$$

$$\bar{K} = \frac{\sum_{i \in K} i}{n_k} \text{ e } \bar{L} = \frac{\sum_{j \in L} j}{n_l}$$



Métodos de Agrupamento Hierárquicos

Método de Ward

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

$$SQE_A = \sum_{i=1}^{n_A} (x_i - \bar{\mathbf{x}}_A)^t (x_i - \bar{\mathbf{x}}_A)$$

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

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

Métodos de Agrupamento Hierárquicos

Método de Ward

- A união entre clusters A e B que proporcionarem menor acréscimo na SQE é executada.
- Para usar o método de Ward, as variáveis devem ser **quantitativas**.

Métodos de Agrupamento Hierárquicos

Exemplo

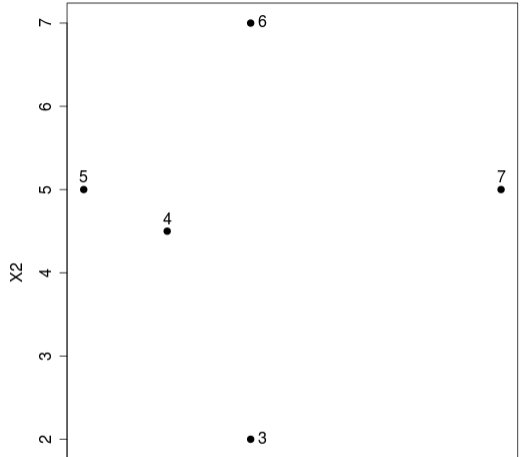
A fim de exemplificar a aplicação dos métodos de agrupamento, considere os dados da Tabela ao lado. Os sete casos são considerados as observações de cada indivíduo para as variáveis X_1 e X_2.

Caso	X_1	X_2
1	1	1
2	2	1
3	3	2
4	2	4,5
5	1	5
6	3	7
7	6	5

Métodos de Agrupamento Hierárquicos

Exemplo

Representação dos casos no plano



Exemplo

Matriz de distâncias

$$D_0 = \begin{bmatrix} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 0,000 & 1,000 & 2,236 & 3,640 & 4,000 & 6,325 & 6,403 \\ 2 & & 0,000 & 1,414 & 3,500 & 4,123 & 6,083 & 5,657 \\ 3 & & & 0,000 & 2,693 & 3,606 & 5,000 & 4,243 \\ 4 & & & & 0,000 & 1,118 & 2,693 & 4,031 \\ 5 & & & & & 0,000 & 2,828 & 5,000 \\ 6 & & & & & & 0,000 & 3,606 \\ 7 & & & & & & & 0,000 \end{bmatrix}$$

Métodos de Agrupamento Hierárquicos

Exemplo

Matriz de distâncias

$$D_0 = \begin{bmatrix} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 0,000 & \mathbf{1,000} & 2,236 & 3,640 & 4,000 & 6,325 & 6,403 \\ 2 & & 0,000 & 1,414 & 3,500 & 4,123 & 6,083 & 5,657 \\ 3 & & & 0,000 & 2,693 & 3,606 & 5,000 & 4,243 \\ 4 & & & & 0,000 & 1,118 & 2,693 & 4,031 \\ 5 & & & & & 0,000 & 2,828 & 5,000 \\ 6 & & & & & & 0,000 & 3,606 \\ 7 & & & & & & & 0,000 \end{bmatrix}$$

$$d_{12} = \sqrt{(X_{11} - X_{21})^2 + (X_{12} - X_{22})^2} = \sqrt{(1-1)^2 + (2-1)^2} = \sqrt{1} = 1$$

Métodos de Agrupamento Hierárquicos

Exemplo

Passo 1: juntar os casos 1 e 2

- Redefinir a matriz de distâncias considerando os casos mais parecidos como se fossem um único grupo.
- **Aqui os métodos se diferenciam!**
- **Método do vizinho mais próximo**

Exemplo

Construção da nova matriz de distâncias

$$d_{((1,2)3)} = \min(d_{13}; d_{23}) = \min(2,236; 1,414) = 1,414$$

$$d_{((1,2)4)} = \min(d_{14}; d_{24}) = \min(3,640; 3,500) = 3,500$$

$$d_{((1,2)5)} = \min(d_{15}; d_{25}) = \min(4,000; 4,123) = 4,000$$

$$d_{((1,2)6)} = \min(d_{16}; d_{26}) = \min(6,325; 6,083) = 6,083$$

$$d_{((1,2)7)} = \min(d_{17}; d_{27}) = \min(6,403; 5,657) = 5,657$$

Exemplo

Matriz de distâncias

$$D_1 = \begin{bmatrix} & (1,2) & 3 & 4 & 5 & 6 & 7 \\ (1,2) & 0,000 & 1,414 & 3,500 & 4,000 & 6,083 & 5,657 \\ 3 & & 0,000 & 2,693 & 3,606 & 5,000 & 4,243 \\ 4 & & & 0,000 & 1,118 & 2,693 & 4,031 \\ 5 & & & & 0,000 & 2,828 & 5,000 \\ 6 & & & & & 0,000 & 3,606 \\ 7 & & & & & & 0,000 \end{bmatrix}$$

Métodos de Agrupamento Hierárquicos

Exemplo

Matriz de distâncias

$$D_1 = \begin{bmatrix} & (1,2) & 3 & 4 & 5 & 6 & 7 \\ (1,2) & 0,000 & 1,414 & 3,500 & 4,000 & 6,083 & 5,657 \\ 3 & & 0,000 & 2,693 & 3,606 & 5,000 & 4,243 \\ 4 & & & 0,000 & \mathbf{1,118} & 2,693 & 4,031 \\ 5 & & & & 0,000 & 2,828 & 5,000 \\ 6 & & & & & 0,000 & 3,606 \\ 7 & & & & & & 0,000 \end{bmatrix}$$

Exemplo

Passo 2: juntar os casos 4 e 5

Redefinir a matriz de distâncias considerando os casos mais parecidos como se fossem um único grupo.

$$
\begin{aligned}
d_{((4,5)(1,2))} &= \min(d_{14}; d_{24}; d_{15}; d_{25}) = \min(3,640; 3,500; 4,000; 4,123) = 3,500 \\
&= \min(d_{(1,2)4}; d_{(1,2)5}) = \min(3,500; 4,000) \\
d_{((4,5)3)} &= \min(d_{34}; d_{35}) = \min(2,693; 3,606) = 2,693 \\
d_{((4,5)6)} &= \min(d_{46}; d_{56}) = \min(2,693; 2,828) = 2,693 \\
d_{((4,5)7)} &= \min(d_{47}; d_{57}) = \min(4,031; 5,000) = 4,031
\end{aligned}
$$

Exemplo

Matriz de distâncias

$$D_2 = \begin{array}{c} \\ (1,2) \\ 3 \\ (4,5) \\ 6 \\ 7 \end{array} \begin{bmatrix} (1,2) & 3 & (4,5) & 6 & 7 \\ 0,000 & 1,414 & 3,500 & 6,083 & 5,657 \\ & 0,000 & 2,693 & 5,000 & 4,243 \\ & & 0,000 & 2,693 & 4,031 \\ & & & 0,000 & 3,606 \\ & & & & 0,000 \end{bmatrix}$$

Exemplo

Matriz de distâncias

$$D_2 = \begin{array}{c} \\ (1,2) \\ 3 \\ (4,5) \\ 6 \\ 7 \end{array} \begin{bmatrix} (1,2) & 3 & (4,5) & 6 & 7 \\ 0,000 & \mathbf{1,414} & 3,500 & 6,083 & 5,657 \\ & 0,000 & 2,693 & 5,000 & 4,243 \\ & & 0,000 & 2,693 & 4,031 \\ & & & 0,000 & 3,606 \\ & & & & 0,000 \end{bmatrix}$$

Exemplo

Passo 3: juntar o grupo $(1,2)$ como caso 3
Redefinir a matriz de distâncias considerando os casos mais parecidos como se fossem um único grupo.

$$
\begin{aligned}
d_{(((1,2,3)(4,5))} &= \min(d_{14}; d_{24}; d_{34}; d_{15}; d_{25}; d_{35}) \\
&= \min(3,640; 3,500; 2,693; 4,000; 4,123; 3,606) = 2,693 \\
&= \min(d_{(1,2)(4,5)}; d_{(4,5)3}) = \min(3,500; 2,693) \\
d_{((1,2,3)6)} &= \min(d_{16}; d_{26}; d_{36}) = \min(6,325; 6,083; 5,000) = 5,000 \\
&= \min(d_{((1,2)6)}; d_{36}) = \min(6,083; 5,000) \\
d_{((1,2,3)7)} &= \min(d_{17}; d_{27}; d_{37}) = \min(6,403; 5,657; 4,243) = 4,243 \\
&= \min(d_{((1,2)7)}; d_{37}) = \min(5,657; 4,243)
\end{aligned}
$$

Exemplo

Matriz de distâncias

$$D_3 = \begin{bmatrix} & (1,2,3) & (4,5) & 6 & 7 \\ (1,2,3) & 0,000 & 2,693 & 5,000 & 4,243 \\ (4,5) & & 0,000 & 2,693 & 4,031 \\ 6 & & & 0,000 & 3,606 \\ 7 & & & & 0,000 \end{bmatrix}$$

Exemplo

Matriz de distâncias

$$D_3 = \begin{bmatrix} & (1,2,3) & (4,5) & 6 & 7 \\ (1,2,3) & 0,000 & \mathbf{2,693} & 5,000 & 4,243 \\ (4,5) & & 0,000 & \mathbf{2,693} & 4,031 \\ 6 & & & 0,000 & 3,606 \\ 7 & & & & 0,000 \end{bmatrix}$$

Métodos de Agrupamento Hierárquicos

Exemplo

Passo 4: juntar o grupo $(4,5)$ como caso 6
Redefinir a matriz de distâncias considerando os casos mais parecidos como se fossem um único grupo.

$$
\begin{aligned}
d_{(((4,5,6)(1,2,3))} &= \min(d_{14}; d_{24}; d_{34}; d_{15}; d_{25}; d_{35}; d_{16}; d_{26}; d_{36}) \\
&= \min(3,640; 3,500; 2,693; 4,000; 4,123; 3,606; 6,325; 6,083; 5,000) = 2,693 \\
&= \min(d_{(1,2,3)(4,5)}; d_{(1,2,3)6}) = \min(2,693; 5,000) \\
d_{((4,5,6)7)} &= \min(d_{47}; d_{57}; d_{67}) = \min(4,031; 5,000; 3,606) = 3,606 \\
&= \min(d_{((4,5)7)}; d_{67}) = \min(4,031; 3,606)
\end{aligned}
$$

Métodos de Agrupamento Hierárquicos

Exemplo

Matriz de distâncias

$$D_4 = \begin{bmatrix} & (1,2,3) & (4,5,6) & 7 \\ (1,2,3) & 0,000 & 2,693 & 4,243 \\ (4,5,6) & & 0,000 & 3,606 \\ 7 & & & 0,000 \end{bmatrix}$$

Métodos de Agrupamento Hierárquicos

Exemplo

Matriz de distâncias

$$D_4 = \begin{bmatrix} & (1,2,3) & (4,5,6) & 7 \\ (1,2,3) & 0,000 & \mathbf{2,693} & 4,243 \\ (4,5,6) & & 0,000 & 3,606 \\ 7 & & & 0,000 \end{bmatrix}$$

Exemplo

Passo 5: juntar o grupo $(1, 2, 3)$ com o grupo $(4, 5, 6)$
Redefinir a matriz de distâncias considerando os casos mais parecidos como se fossem um único grupo.

$$\begin{aligned} d_{((1,2,3,4,5,6)7)} &= \min(d_{17}; d_{27}; d_{37}; d_{47}; d_{57}; d_{67}) \\ &= \min(6,403; 5,657; 4,243; 4,031; 5,000; 3,606) = 3,606 \\ &= \min(d_{(1,2,3)7}; d_{(4,5,6)7}) = \min(4,243; 3,606) \end{aligned}$$

Métodos de Agrupamento Hierárquicos

Exemplo

Matriz de distâncias

$$D_5 = \begin{bmatrix} & (1,2,3,4,5,6) & 7 \\ (1,2,3,4,5,6) & 0,000 & 3,606 \\ 7 & & 0,000 \end{bmatrix}$$

Métodos de Agrupamento Hierárquicos

Exemplo

Matriz de distâncias

$$D_5 = \begin{bmatrix} & (1,2,3,4,5,6) & 7 \\ (1,2,3,4,5,6) & 0,000 & \mathbf{3,606} \\ 7 & & 0,000 \end{bmatrix}$$

Métodos de Agrupamento Hierárquicos

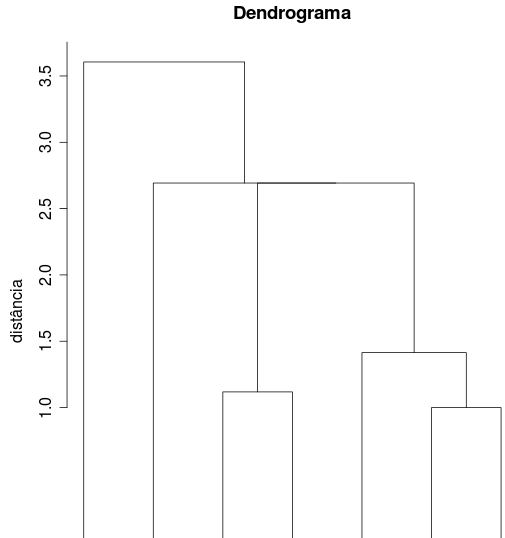
Exemplo

Resumo do método de agrupamento

Passo	Nº de grupos	Grupos	Distância
1	7	1, 2, 3, 4, 5, 6, 7	0,000
2	6	(1,2), 3, 4, 5, 6, 7	1,000
3	5	(1,2), 3, (4,5), 6, 7	1,118
4	4	(1,2,3), (4,5), 6, 7	1,414
5	3	(1,2,3), (4,5,6), 7	2,693
6	2	(1,2,3,4,5,6), 7	2,693
7	1	(1,2,3,4,5,6,7)	3,606

Métodos de Agrupamento Hierárquicos

Exemplo



Métodos de Agrupamento Hierárquicos

Determinação do número de grupos

- O dendrograma permite ao pesquisador consultar a distância em que os clusters foram combinados para formar um novo cluster.
- Clusters que são semelhantes entre si são combinados a baixas distâncias, enquanto grupos que são mais dissimilares são combinados em altas distâncias.
- A diferença de distâncias define como os clusters próximos são um do outro.

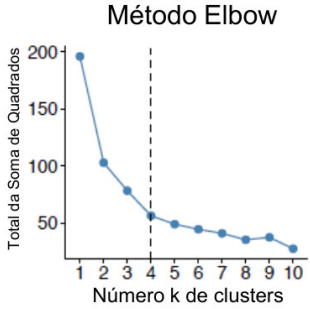
Métodos de Agrupamento Hierárquicos

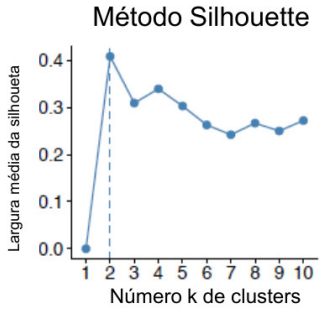
Determinação do número de grupos

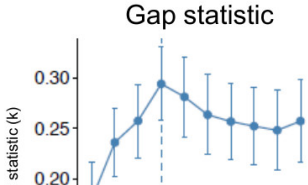
- Uma partição dos dados em um número especificado de grupos pode ser obtida "cortando" o dendograma a uma distância apropriada.
- Se traçarmos uma linha horizontal no dendograma a uma determinada distância, então o número k das linhas verticais cortadas por essa linha horizontal identificará uma solução k-cluster.
- A intersecção da linha horizontal e uma dessas linhas verticais representa um cluster, e os itens localizados no final de todos os ramos abaixo interseção constituem os membros do cluster.

Métodos de Agrupamento Hierárquicos

Determinação do número de grupos







Métodos hierárquicos: comentários finais

- Fontes de erros e de variação não são formalmente considerados nos procedimentos hierárquicos: Significa que esses métodos são sensíveis a outliers ou pontos de perturbação

Métodos hierárquicos: comentários finais

- Fontes de erros e de variação não são formalmente considerados nos procedimentos hierárquicos: Significa que esses métodos são sensíveis a outliers ou pontos de perturbação
- Deve-se sempre verificar a sensibilidade da configuração dos grupos: Os métodos não permitem a realocação de objetos que possam ter sido agrupados incorretamente nos estágios iniciais

Métodos hierárquicos: comentários finais

- Fontes de erros e de variação não são formalmente considerados nos procedimentos hierárquicos: Significa que esses métodos são sensíveis a outliers ou pontos de perturbação
- Deve-se sempre verificar a sensibilidade da configuração dos grupos: Os métodos não permitem a realocação de objetos que possam ter sido agrupados incorretamente nos estágios iniciais
- É recomendado tentar vários métodos de agrupamento e de atribuição de distâncias (similaridades)

Métodos hierárquicos: comentários finais

- Fontes de erros e de variação não são formalmente considerados nos procedimentos hierárquicos: Significa que esses métodos são sensíveis a outliers ou pontos de perturbação
- Deve-se sempre verificar a sensibilidade da configuração dos grupos: Os métodos não permitem a realocação de objetos que possam ter sido agrupados incorretamente nos estágios iniciais
- É recomendado tentar vários métodos de agrupamento e de atribuição de distâncias (similaridades)
- Empates na matriz de distâncias podem produzir múltiplas soluções ao problema de agrupamento hierárquico

Métodos hierárquicos: comentários finais

- A maioria dos métodos produz clusters esféricos ou elípticos

Métodos hierárquicos: comentários finais

- A maioria dos métodos produz clusters esféricos ou elípticos
- O método de **ligação simples** é um dos poucos métodos que pode delinear cluster não-elípticos
  - Tem a capacidade de gerar estruturas geométricas diferentes
  - Entretanto, ele é incapaz de perceber grupos pouco separados

Métodos hierárquicos: comentários finais

- A maioria dos métodos produz clusters esféricos ou elípticos
- O método de **ligação simples** é um dos poucos métodos que pode delinear cluster não-elípticos
  - Tem a capacidade de gerar estruturas geométricas diferentes
  - Entretanto, ele é incapaz de perceber grupos pouco separados
- Os clusters formados pelo método de ligação simples não serão modificados por qualquer atribuição de distância (similaridade) que dá as mesmas ordenações relativas

Métodos hierárquicos: comentários finais

- O método de **ligação completa** tende a produzir conglomerados de aproximadamente mesmo diâmetro
  - Tem a tendência de isolar os valores discrepantes nos estágios iniciais do agrupamento

Métodos hierárquicos: comentários finais

- O método de **ligação completa** tende a produzir conglomerados de aproximadamente mesmo diâmetro
  - Tem a tendência de isolar os valores discrepantes nos estágios iniciais do agrupamento
- O método da **média das distâncias** tende a produzir conglomerados de aproximadamente mesma variância interna
  - Em geral, produz melhores partições que os métodos de ligação simples e completa

Métodos hierárquicos: comentários finais

- Os métodos de ligação simples, completa e da média podem ser utilizados tanto para **variáveis quantitativas** quanto para **variáveis qualitativas**

Métodos hierárquicos: comentários finais

- Os métodos de ligação simples, completa e da média podem ser utilizados tanto para **variáveis quantitativas** quanto para **variáveis qualitativas**
- Os métodos do centróide e de Ward são apropriados apenas para **variáveis quantitativas**

Métodos hierárquicos: comentários finais

- Os métodos de ligação simples, completa e da média podem ser utilizados tanto para **variáveis quantitativas** quanto para **variáveis qualitativas**
- Os métodos do centróide e de Ward são apropriados apenas para **variáveis quantitativas**
- O **método de Ward** tende a produzir grupos com aproximadamente o mesmo número de elementos

Métodos hierárquicos: comentários finais

- Os métodos de ligação simples, completa e da média podem ser utilizados tanto para **variáveis quantitativas** quanto para **variáveis qualitativas**
- Os métodos do centróide e de Ward são apropriados apenas para **variáveis quantitativas**
- O **método de Ward** tende a produzir grupos com aproximadamente o mesmo número de elementos
- Espera-se sempre que haja uma certa consistência entre as soluções obtidas por métodos diferentes: pode não ocorrer a igualdade das soluções apresentadas pelos vários métodos

Validação dos agrupamentos

- MANOVA

Validação dos agrupamentos

- MANOVA
- Análise Discriminante

Validação dos agrupamentos

- MANOVA
- Análise Discriminante
- Correlação Cofenética

Validação dos agrupamentos

- MANOVA
- Análise Discriminante
- Correlação Cofenética
- Gráfico da Silhueta

Validação dos agrupamentos

Correlação Cofenética

- Medida de validação usada nos métodos hierárquicos principalmente
- **Idéia:** realizar uma comparação das distâncias observadas e preditas (via a formação de agrupamentos) entre os objetos

Validação dos agrupamentos

Dist. Observada

	SJRP	RP	Bauru	Campinas	Sorocaba
SJRP	0				
RP	0,59	0			
Bauru	0,55	1,05	0		
Campinas	2,74	2,27	2,89	0	
Sorocaba	2,37	2,17	2,24	1,37	0

Dist. Predita – Ligação Completa

Passo	Grupo	Distância
1	**SJRP,Bauru**	0,55
2	**SJRP,Bauru,RP**	1,05
3	**Campinas,Sorocaba**	1,37
4	**SJRP,Bauru,RP,Campinas,Sorocaba**	2,89

Validação dos agrupamentos

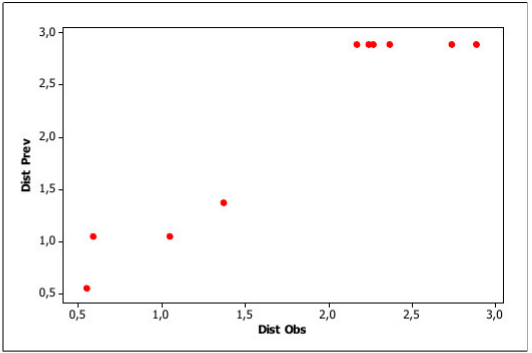
Passo	Grupo	Distância
1	SJRP,Bauru	0,55
2	SJRP,Bauru,RP	1,05
3	Campinas,Sorocaba	1,37
4	SJRP,Bauru,RP,Campinas,Sorocaba	2,89

Matriz de Distâncias Preditas – Ligação Completa

	SJRP	RP	Bauru	Campinas	Sorocaba
SJRP	0				
RP	1,05	0			
Bauru	0,55	1,05	0		
Campinas	2,89	2,89	2,89	0	
Sorocaba	2,89	2,89	2,89	1,37	0

Validação dos agrupamentos

Pares de Região		Dist. Observada	Dist. Prevista
SJRP	RP	0,59	1,05
SJRP	Bauru	0,55	0,55
SJRP	Campinas	2,74	2,89
SJRP	Sorocaba	2,37	2,89
RP	Bauru	1,05	1,05
RP	Campinas	2,27	2,89
RP	Sorocaba	2,17	2,89
Bauru	Campinas	2,89	2,89
Bauru	Sorocaba	2,24	2,89
Campinas	Sorocaba	1,37	1,37



r = 0,954

Correlação cofenética

$\Rightarrow$ O agrupamento formado via o algoritmo do vizinho mais longe é de boa qualidade!

Validação dos agrupamentos

Correlação Cofenética

- Em um bom agrupamento, espera-se que as distâncias previstas respeitem a ordem determinada pelas distâncias observadas.
- Para avaliar a ocorrência desse comportamento, define-se a correlação cofenética como sendo a correlação entre as distâncias efetivamente observadas e as previstas.

Validação dos agrupamentos

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 quão bem cada elemento está alocado 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.

Validação dos agrupamentos

Gráfico da Silhueta

- $a_{(i)}$ = distância média do objeto i para os elementos de seu próprio grupo
- $b_{(i)}$ = distância média do objeto i para os elementos do grupo mais próximo

$$s_{(i)} = \frac{b_{(i)} - a_{(i)}}{\max(b_{(i)}, a_{(i)})}, \quad i = 1, \cdots, n$$

- Por definição, $-1 \leqslant s_{(i)} \leqslant 1$.

Validação dos agrupamentos

Gráfico 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 (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 (i está mal alocado);
- Se $a_{(i)} \approx b_{(i)}, s_{(i)} \approx 0$, indicando que i está na fronteira de seu grupo e do grupo vizinho.

Validação dos agrupamentos

Gráfico da Silhueta

Silhueta Média	**Interpretação Sugerida**
$0{,}71 - 1{,}00$	Grupos encontrados possuem estrutura muito robusta
$0{,}51 - 0{,}70$	Grupos razoavelmente unidos
$0{,}26 - 0{,}50$	A estrutura encontrada é fraca, tente outros métodos de agrupamento
$\leqslant 0{,}25$	Nenhuma estrutura encontrada

Métodos de agrupamento não-hierárquicos - MANH

- Os MANH não se baseiam em sucessivas aglomerações (ou partições) dos elementos com base numa matiz de distâncias;

Métodos de agrupamento não-hierárquicos - MANH

- Os MANH não se baseiam em sucessivas aglomerações (ou partições) dos elementos com base numa matiz de distâncias;
- Nos MANH tem-se a possibilidade de realocar indivíduos entre clusters ao longo do processo, diferentemente do que ocorre nos métodos hierárquicos;

Métodos de agrupamento não-hierárquicos - MANH

- Os MANH não se baseiam em sucessivas aglomerações (ou partições) dos elementos com base numa matiz de distâncias;
- Nos MANH tem-se a possibilidade de realocar indivíduos entre clusters ao longo do processo, diferentemente do que ocorre nos métodos hierárquicos;
- Há diferentes técnicas não hierárquicas de agrupamento, dentre as quais aquelas baseadas em partições amostrais (particularmente o algoritmo k-means) são as mais populares.

Método k-médias (*k-means*)

a) Selecionar aleatoriamente k indivíduos como centroides iniciais (ou escolha os centroides iniciais de alguma forma);

Método k-médias (*k-means*)

a) Selecionar aleatoriamente k indivíduos como centroides iniciais (ou escolha os centroides iniciais de alguma forma);

b) Calcular as distâncias entre cada indivíduo e o centro de cada um dos k grupos e classificar o indivíduo no grupo mais próximo.

Método k-médias (*k-means*)

a) Selecionar aleatoriamente k indivíduos como centroides iniciais (ou escolha os centroides iniciais de alguma forma);

b) Calcular as distâncias entre cada indivíduo e o centro de cada um dos k grupos e classificar o indivíduo no grupo mais próximo.

c) Calcule o centroide de cada grupo;

Método k-médias (*k-means*)

a) Selecionar aleatoriamente k indivíduos como centroides iniciais (ou escolha os centroides iniciais de alguma forma);

b) Calcular as distâncias entre cada indivíduo e o centro de cada um dos k grupos e classificar o indivíduo no grupo mais próximo.

c) Calcule o centroide de cada grupo;

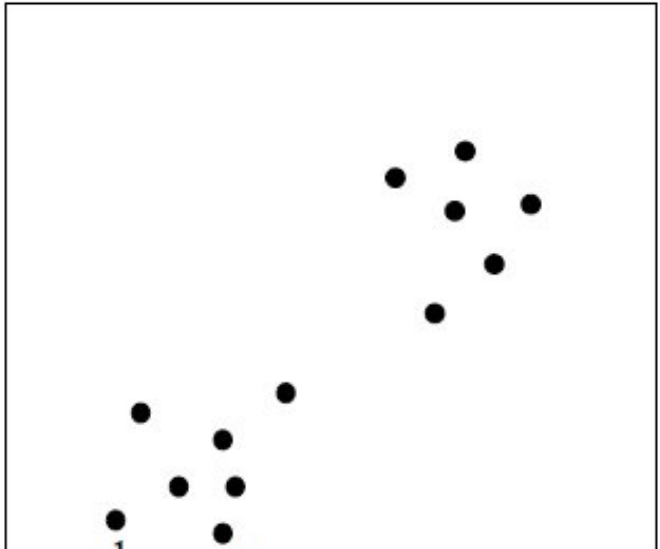
d) Repita os passos $b)$ e $c)$ até que os centroides não apresentem mais mudanças.

Método k-médias (*k-means*)

Esquematicamente

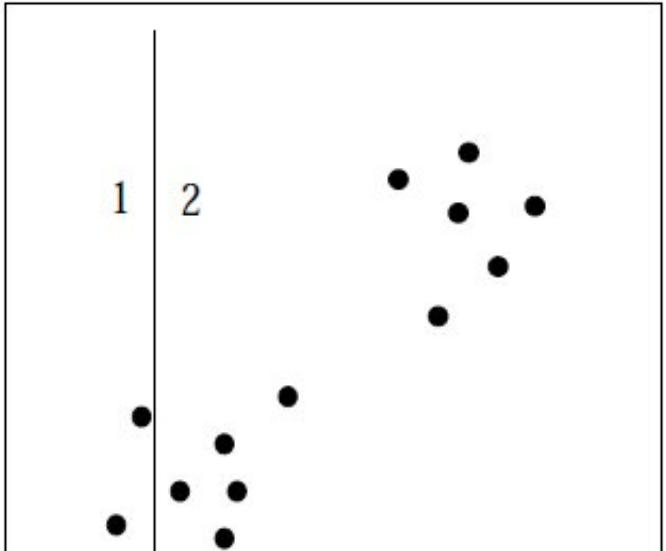
Método k-médias (*k-means*)

Esquematicamente



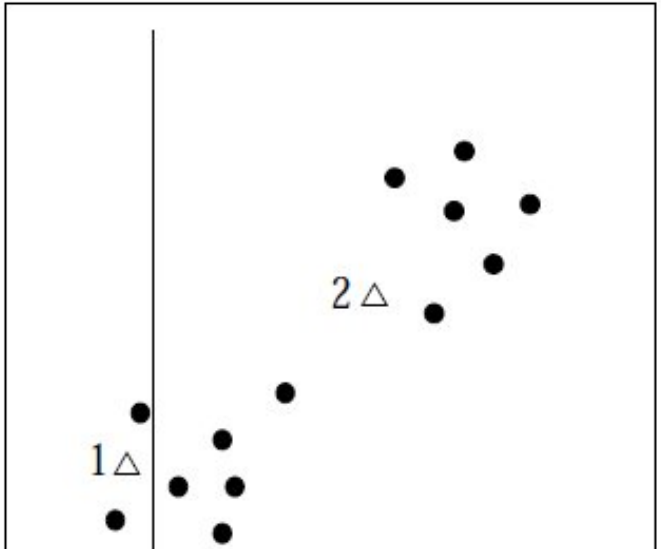
Método k-médias (*k-means*)

Esquematicamente



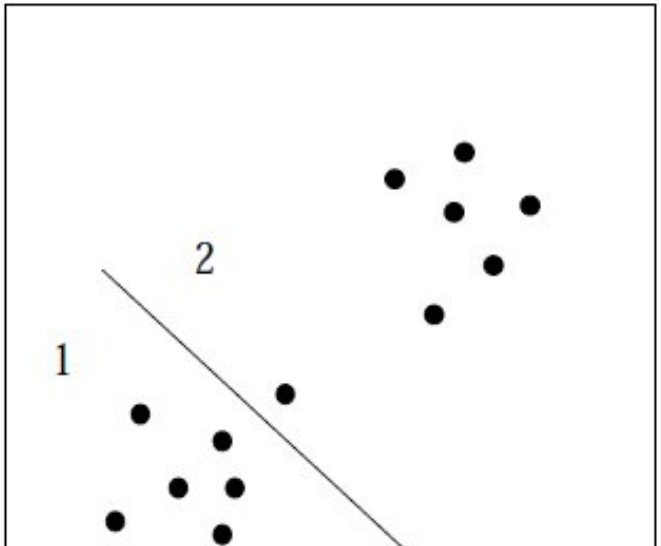
Método k-médias (*k-means*)

Esquematicamente



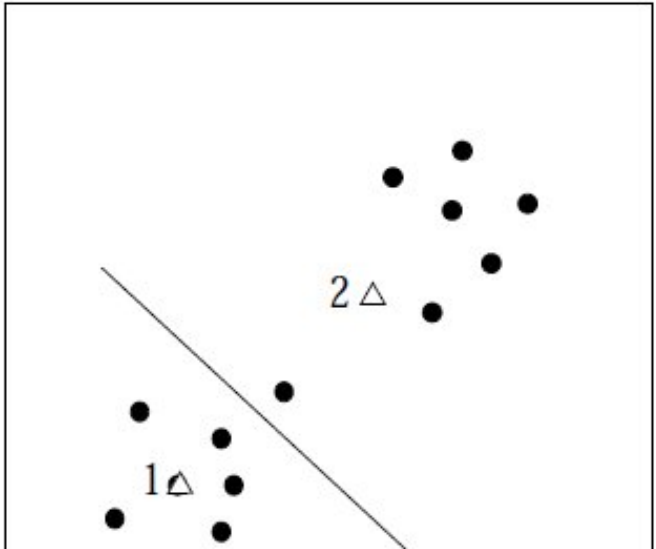
Método k-médias (*k-means*)

Esquematicamente



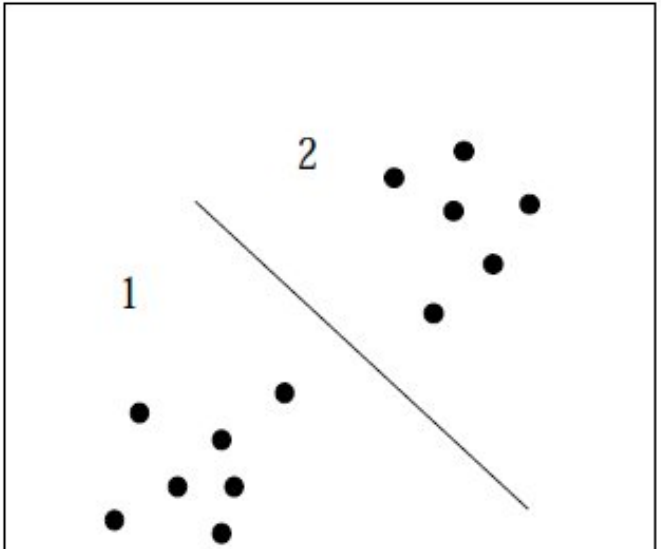
Método k-médias (*k-means*)

Esquematicamente



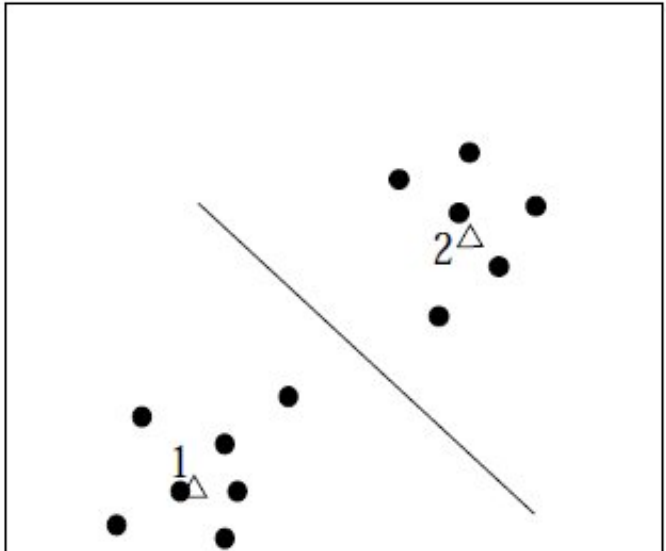
Método k-médias (*k-means*)

Esquematicamente



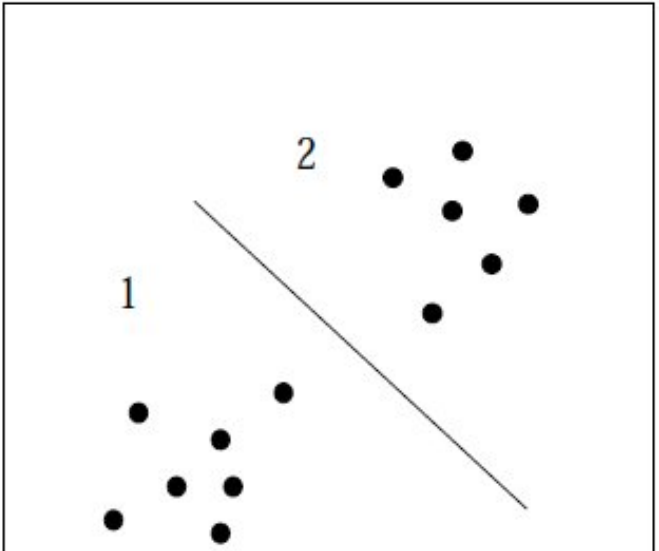
Método k-médias (*k-means*)

Esquematicamente



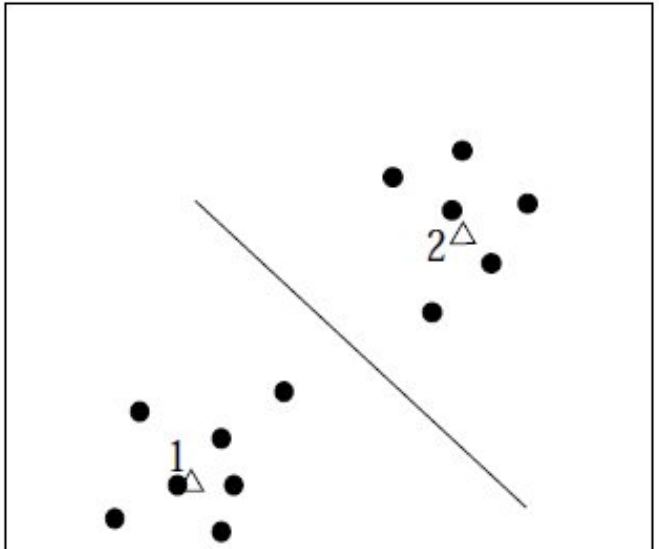
Método k-médias (*k-means*)

Esquematicamente



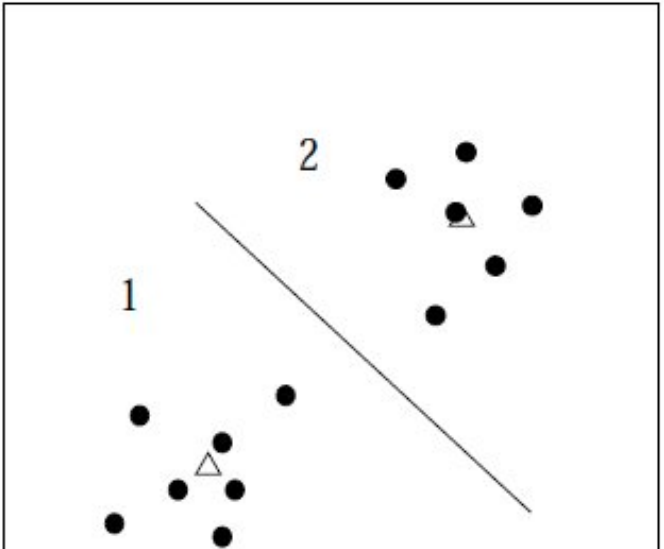
Método k-médias (*k-means*)

Esquematicamente



Método k-médias (*k-means*)

Esquematicamente



Método k-médias: comentários finais

- O algoritmo *k-means* é sensível à configuração inicial dos clusters (passo a)), podendo produzir resultados diferentes mediante diferentes partições iniciais.

Método k-médias: comentários finais

- O algoritmo *k-means* é sensível à configuração inicial dos clusters (passo a)), podendo produzir resultados diferentes mediante diferentes partições iniciais.
- O usual é considerar, inicialmente, k "sementes", que seriam k pontos definidos em R^p.

Método k-médias: comentários finais

- O algoritmo *k-means* é sensível à configuração inicial dos clusters (passo a)), podendo produzir resultados diferentes mediante diferentes partições iniciais.
- O usual é considerar, inicialmente, k "sementes", que seriam k pontos definidos em R^p.
- O processo inicia alocando cada elemento à semente mais próxima, e atualizando o ponto de referência do cluster (centroide) cada vez que ele incorpora um novo elemento.

Método k-médias: comentários finais

- O algoritmo *k-means* é sensível à configuração inicial dos clusters (passo a)), podendo produzir resultados diferentes mediante diferentes partições iniciais.
- O usual é considerar, inicialmente, k "sementes", que seriam k pontos definidos em R^p.
- O processo inicia alocando cada elemento à semente mais próxima, e atualizando o ponto de referência do cluster (centroide) cada vez que ele incorpora um novo elemento.
- Podemos definir as sementes como os vetores de observações de k elementos selecionados aleatoriamente da base.

Método k-médias: comentários finais

- **Nota:** Estabelecer alguma restrição quanto à distância das sementes, garantindo alguma distância mínima entre elas, pode ajudar na performance do método.

Método k-médias: comentários finais

- **Nota:** Estabelecer alguma restrição quanto à distância das sementes, garantindo alguma distância mínima entre elas, pode ajudar na performance do método.
- As sementes podem ser definidas como k vetores quaisquer definidos em R^p;

Método k-médias: comentários finais

- **Nota:** Estabelecer alguma restrição quanto à distância das sementes, garantindo alguma distância mínima entre elas, pode ajudar na performance do método.
- As sementes podem ser definidas como k vetores quaisquer definidos em R^p;
- Caso se considere algum mecanismo aleatório de seleção de sementes, é recomendável que o processo seja repetido um número m de vezes, identificando-se, dentre as m soluções obtidas, a solução ótima;

Método k-médias: comentários finais

- Uma alternativa é rodar, inicialmente, uma análise hierárquica, e definir k clusters;

Método k-médias: comentários finais

- Uma alternativa é rodar, inicialmente, uma análise hierárquica, e definir k clusters;
- Os centroides dos k clusters obtidos podem servir de semente para o algoritmo *k - means*.