SME0822 Análise Multivariada e Aprendizado Não-Supervisionado

Aula 10a: **Análise de Agrupamentos (clustering, conglomerados)**

Prof. Cibele Russo

cibele@icmc.usp.br

http://www.icmc.usp.br/~cibele

Johnson, R. A., & Wichern, D. W. (2007). Applied Multivariate Statistical Analysis. Prentice Hall.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013) An Introduction to Statistical Learning. New York: Springer.

Análise de agrupamentos

Objetivos:

Dividir os elementos da amostra ou população em **grupos**, de forma que os elementos pertencentes a um grupo específico sejam **similares** entre si, com respeito às variáveis que neles foram medidas, e que os elementos em grupos distintos sejam **dissimilares** em relação às mesmas variáveis.

Análise de agrupamentos

Exemplos:

- Classificação de pessoas quanto à sua personalidade,
- Identificação do posicionamento de produtos em relação aos seus concorrentes no mercado,
- Segmentação de clientes de acordo com perfis de consumo,
- Classificação de cidades de acordo com variáveis físicas, demográficas e econômicas, entre outros.

Análise de agrupamentos

Exemplos:

- Classificação de pessoas quanto à sua personalidade,
- Identificação do posicionamento de produtos em relação aos seus concorrentes no mercado,
- Segmentação de clientes de acordo com perfis de consumo,
- Classificação de cidades de acordo com variáveis físicas, demográficas e econômicas, entre outros.

Análise de agrupamentos

Exemplos:

- Classificação de pessoas quanto à sua personalidade,
- Identificação do posicionamento de produtos em relação aos seus concorrentes no mercado,
- Segmentação de clientes de acordo com perfis de consumo,
- Classificação de cidades de acordo com variáveis físicas, demográficas e econômicas, entre outros.

Análise de agrupamentos

Exemplos:

- Classificação de pessoas quanto à sua personalidade,
- Identificação do posicionamento de produtos em relação aos seus concorrentes no mercado,
- Segmentação de clientes de acordo com perfis de consumo,
- Classificação de cidades de acordo com variáveis físicas, demográficas e econômicas, entre outros.

Análise de agrupamentos

Medidas de similaridade e dissimilaridade:

Suponha que temos um conjunto de dados com n elementos amostrais, com medidas de p variáveis aleatórias em cada um deles.

Queremos agrupar n elementos em g grupos.

Para cada elemento amostral j, tem-se portanto o vetor de medidas

$$\underset{\sim}{X}_j = (X_{1j}, \ldots, X_{pj})^\top, \text{ com } j = 1, \ldots, n$$

em que X_{ij} é o valor observado da variável i no elemento amostral j.

Análise de agrupamentos

Para agrupar esses elementos, precisamos de uma medida de **similaridade ou dissimilaridade**.

As **medidas de dissimilaridade** mais utilizadas para **variáveis quantitativas** são:

- Distância euclidiana
- Distância generalizada ou ponderada
- Distância de Minkowski

Análise de agrupamentos

Para agrupar esses elementos, precisamos de uma medida de **similaridade ou dissimilaridade**.

As **medidas de dissimilaridade** mais utilizadas para **variáveis quantitativas** são:

- Distância euclidiana
- Distância generalizada ou ponderada
- Distância de Minkowski

Medidas de dissimilaridade

Distância Euclidiana:

$$d_E(\underset{\sim}{X}_k, \underset{\sim}{X}_l) = \sqrt{(\underset{\sim}{X}_k - \underset{\sim}{X}_l)^\top (\underset{\sim}{X}_k - \underset{\sim}{X}_l)} = \sqrt{\sum_{i=1}^{p} (X_{ik} - X_{il})^2}$$

Medidas de dissimilaridade

Distância generalizada ou ponderada:

$$d(\underset{\sim}{X}_k, \underset{\sim}{X}_l) = \sqrt{(\underset{\sim}{X}_k - \underset{\sim}{X}_l)^\top A (\underset{\sim}{X}_k - \underset{\sim}{X}_l)},$$

em que A é uma matriz de ponderação.

- Se $A = I$, tem-se a distância Euclidiana,
- Se $A = S^{-1}$, tem-se a distância de Mahalanobis,
- Se $A = diag\left(\frac{1}{p}\right)$, tem-se a distância Euclidiana média.

Medidas de dissimilaridade

Distância de Minkowski:

$$d_{M_k}(\underset{\sim}{X}_k, \underset{\sim}{X}_l) = \left\{ \sum_{i=1}^{p} |X_{il} - X_{ik}|^{\lambda} \right\}^{1/\lambda}, \text{ ou}$$

$$d^{\star}_{M_k}(\underset{\sim}{X}_k, \underset{\sim}{X}_l) = \left\{ \sum_{i=1}^{p} \omega_i |X_{il} - X_{ik}|^{\lambda} \right\}^{1/\lambda}.$$

Medidas de dissimilaridade

Distância de Minkowski:

$$d_{M_k}(\underset{\sim}{X}_k, \underset{\sim}{X}_l) = \left\{ \sum_{i=1}^{p} |X_{il} - X_{ik}|^{\lambda} \right\}^{1/\lambda}, \text{ ou}$$

$$d^{\star}_{M_k}(\underset{\sim}{X}_k, \underset{\sim}{X}_l) = \left\{ \sum_{i=1}^{p} \omega_i |X_{il} - X_{ik}|^{\lambda} \right\}^{1/\lambda}.$$

Matriz de distâncias

Definida a distância a ser utilizada, armazenam-se as distâncias amostrais numa matriz $n \times n$:

$$D = \begin{pmatrix} 0 & d_{12} & d_{13} & \dots & d_{1n} \\ & 0 & d_{23} & \dots & d_{2n} \\ & & & \ddots & \vdots \\ & & & \ddots & d_{n-1,n} \\ & & & & 0 \end{pmatrix}$$

Medidas de similaridade

Para variáveis nominais, usamos **medidas de similaridade**.

Exemplo: Em medicina, é comum coletar a presença ou ausência de características nos pacientes em estudo (0: ausência, 1: presença).

Paciente	Fumante	Álcool	Drogas ilícitas	Tranquilizantes
1	1	1	0	0
2	0	1	0	1
3	0	0	1	0

Medidas de similaridade

Coeficiente de concordância simples:

Para cada par de observações, conta-se quantas vezes as respostas concordam. Para o exemplo,

$$S_{1,2} = \frac{2}{4} = 0.5$$

$$S_{1,3} = \frac{1}{4} = 0.25$$

$$S_{2,3} = \frac{1}{4} = 0.25.$$

Quanto maior S, maior a similaridade entre os pacientes.

Dessa forma, para essa aplicação os pacientes 1 e 2 são mais próximos do que 1 e 3, por exemplo.

Medidas de similaridade

Coeficiente de concordância de Jaccard:

Avalia-se quantas variáveis assumem o mesmo valor em diferentes elementos, com exceção dos pares (0,0):

$S_{1,2} = \frac{1}{3}$

$S_{1,3} = 0$

$S_{2,3} = 0.$

Quanto maior S, maior a similaridade entre os pacientes.

Medidas de similaridade e dissimilaridade

Considere a **Distância Euclidiana média**:

$$d_{1,2} = \sqrt{\frac{\sum_{i=1}^{4}(X_{i1} - X_{i2})^2}{4}}$$

É comum transformar distâncias em coeficientes de similaridade fazendo

$$S_{A,B} = 1 - d^{\star}(A, B)$$

com

$$d^{\star}(A, B) = \frac{d(A, B) - min(d)}{max(d) - min(d)}$$

Técnicas para a construção de conglomerados (clusters)

Técnicas hierárquicas

- O número de grupos é definido pós-análise
- Existem técnicas hierárquicas aglomerativas e divisivas

Técnicas não-hierárquicas

O número de grupos é definido previamente

Técnicas para a construção de conglomerados (clusters)

Técnicas hierárquicas aglomerativas

- Inicia-se o processo com n conglomerados, ou seja, cada elemento amostral é considerado um cluster de tamanho 1, e no último estágio existe um único cluster de tamanho n contendo todos os elementos amostrais.
- É comum construir um gráfico chamado **"Dendrograma"**, que representa a árvore do agrupamento.
- A escolha do número de grupos finais g é subjetiva, embora existam técnicas para definir g.

Técnicas para a construção de conglomerados (clusters)

Técnicas hierárquicas aglomerativas

- Inicia-se o processo com n conglomerados, ou seja, cada elemento amostral é considerado um cluster de tamanho 1, e no último estágio existe um único cluster de tamanho n contendo todos os elementos amostrais.
- É comum construir um gráfico chamado **"Dendrograma"**, que representa a árvore do agrupamento.
- A escolha do número de grupos finais g é subjetiva, embora existam técnicas para definir g.

Técnicas para a construção de conglomerados (clusters)

Técnicas hierárquicas aglomerativas

- Inicia-se o processo com n conglomerados, ou seja, cada elemento amostral é considerado um cluster de tamanho 1, e no último estágio existe um único cluster de tamanho n contendo todos os elementos amostrais.
- É comum construir um gráfico chamado **"Dendrograma"**, que representa a árvore do agrupamento.
- A escolha do número de grupos finais g é subjetiva, embora existam técnicas para definir g.

Métodos de ligação

Sejam os conglomerados $C_1 = \{\underset{\sim}{X}_1, \underset{\sim}{X}_3, \underset{\sim}{X}_7\}$ e $C_2 = \{\underset{\sim}{X}_2, \underset{\sim}{X}_6\}$.

Ligação simples

Em cada estágio do processo de agrupamento, combina-se os dois elementos (ou conglomerados) mais similares em um único cluster.

$D(C_1, C_2) = min\{D(\underset{\sim}{X}_l, \underset{\sim}{X}_k),\ l \neq k,\ l = 1, 3, 7 \text{ e } k = 2, 6\}$

Ligação completa

$D(C_1, C_2) = max\{D(\underset{\sim}{X}_l, \underset{\sim}{X}_k),\ l \neq k,\ l = 1, 3, 7 \text{ e } k = 2, 6\}$

Métodos de ligação

Média das distâncias

$$D(C_1, C_2) = \sum_{\underset{\sim}{X}_l \in C_1} \sum_{\underset{\sim}{X}_k \in C_2} \frac{1}{n_l n_k} D(\underset{\sim}{X}_l, \underset{\sim}{X}_k)$$

Centroide

$A = \{\underset{\sim}{X}_1, \underset{\sim}{X}_3, \underset{\sim}{X}_7\}$ e $B = \{\underset{\sim}{X}_2, \underset{\sim}{X}_6\}$

$$\overline{\underset{\sim}{X}}_1 = \frac{\underset{\sim}{X}_1 + \underset{\sim}{X}_3 + \underset{\sim}{X}_7}{3} \text{ e } \overline{\underset{\sim}{X}}_2 = \frac{\underset{\sim}{X}_2 + \underset{\sim}{X}_6}{2}$$

$$D(A, B) = (\overline{\underset{\sim}{X}}_1 - \overline{\underset{\sim}{X}}_2)^\top (\overline{\underset{\sim}{X}}_1 - \overline{\underset{\sim}{X}}_2)$$

Ward

A cada passo, agrupam-se os conglomerados que minimizem a variância dos grupos.

Método de Ward

Seja $\overset{\sim}{\overline{X}}_i$ o vetor de médias do i-ésimo conglomerado. Em cada passo, calcula-se a soma de quadrados dentro de cada conglomerado

$$SS_i = \sum_{j=1}^{n_i} (\tilde{X}_{ij} - \overset{\sim}{\overline{X}}_i)^\top (\tilde{X}_{ij} - \overset{\sim}{\overline{X}}_i)$$

Em cada passo k, a soma de quadrados total é dada por

$$SSR = \sum_{i=1}^{g_k} SS_i$$

onde g_k é o número de conglomerados existentes no passo k.

Método de Ward

A distância entre os conglomerados C_l e C_i é definida por

$$d(C_l, C_i) = \frac{n_l n_i}{n_l + n_i}(\overline{\underset{\sim}{X}}_l - \overline{\underset{\sim}{X}}_i)^\top(\overline{\underset{\sim}{X}}_l - \overline{\underset{\sim}{X}}_i).$$

Em cada passo, os dois conglomerados que levem à menor distância são agrupados.

Agrupamentos hierárquicos - Exemplo

Os dados da Tabela a seguir foram obtidos da publicação “USP em números" e apresentam estatísticas de publicações nas unidades USP que produziram pelo menos 100 publicações no exterior no ano de 2010:

Agrupamentos hierárquicos - Exemplo

Tabela 1: Publicações em unidades USP.

Unidade	P.BR	N.BR	P.EXT	N.EXT	N.AUT
FM	589	416	739	433	506
FMRP	303	157	655	341	491
ESALQ	446	349	165	134	227
EP	231	179	274	219	215
FMVZ	250	183	209	127	168
FOB	287	151	260	118	120
ICB	62	50	365	262	280
IFSC	45	31	384	284	162
IF	26	19	384	239	159
IQ	90	67	274	215	184
FO	161	112	218	132	142
FFCLRP	87	79	197	157	189
FCF	110	78	209	143	157
ICMC	41	31	169	140	111
IQSC	60	51	158	130	104
IB	104	88	145	107	127
EEFE	119	87	143	101	68

Agrupamentos hierárquicos - Exemplo

em que

Unidade: Unidade USP

P.BR: Participações em publicações no Brasil

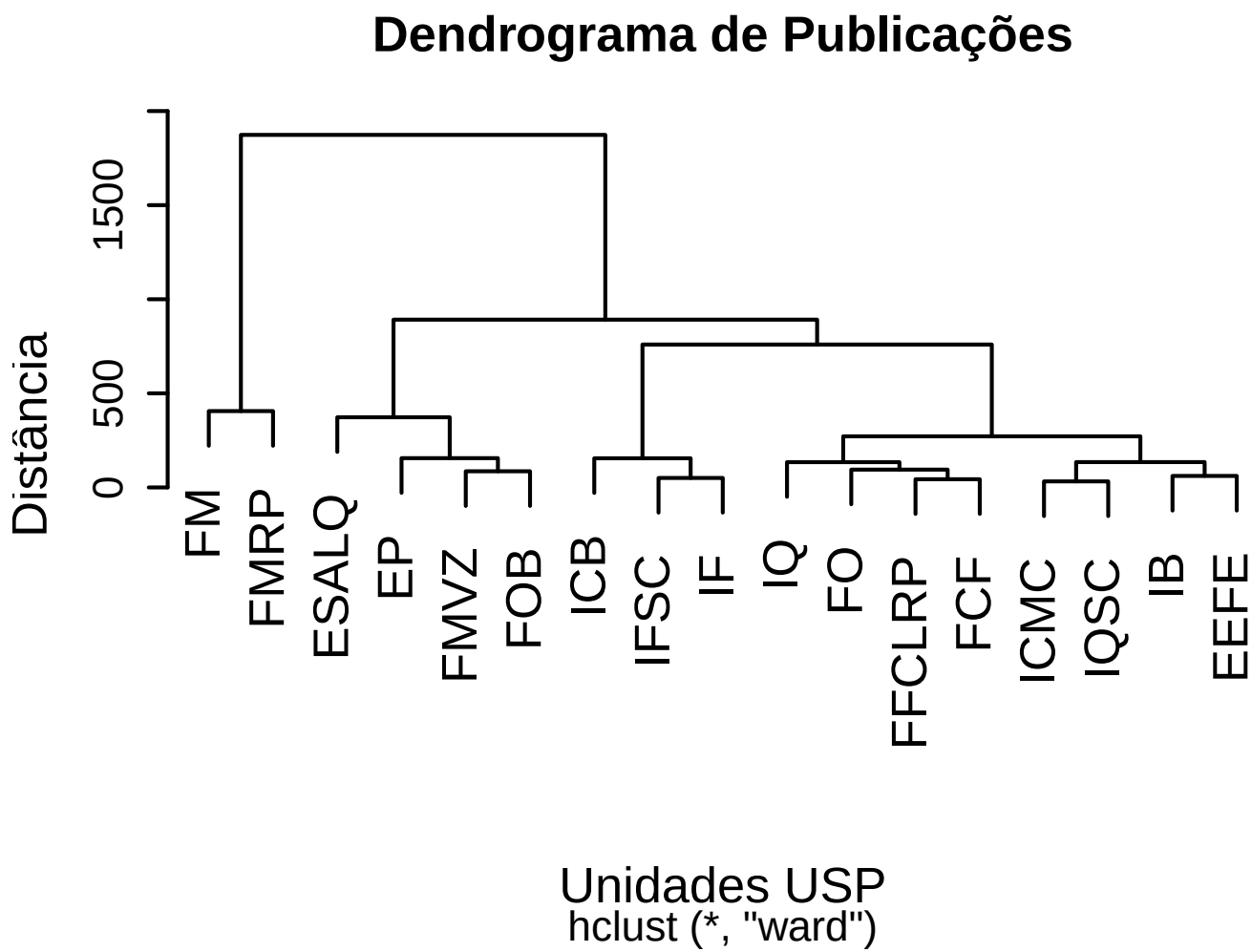
N.BR: Número de publicações no Brasil

P.EXT: Participações em publicações no exterior

N.EXT: Número de publicações no exterior

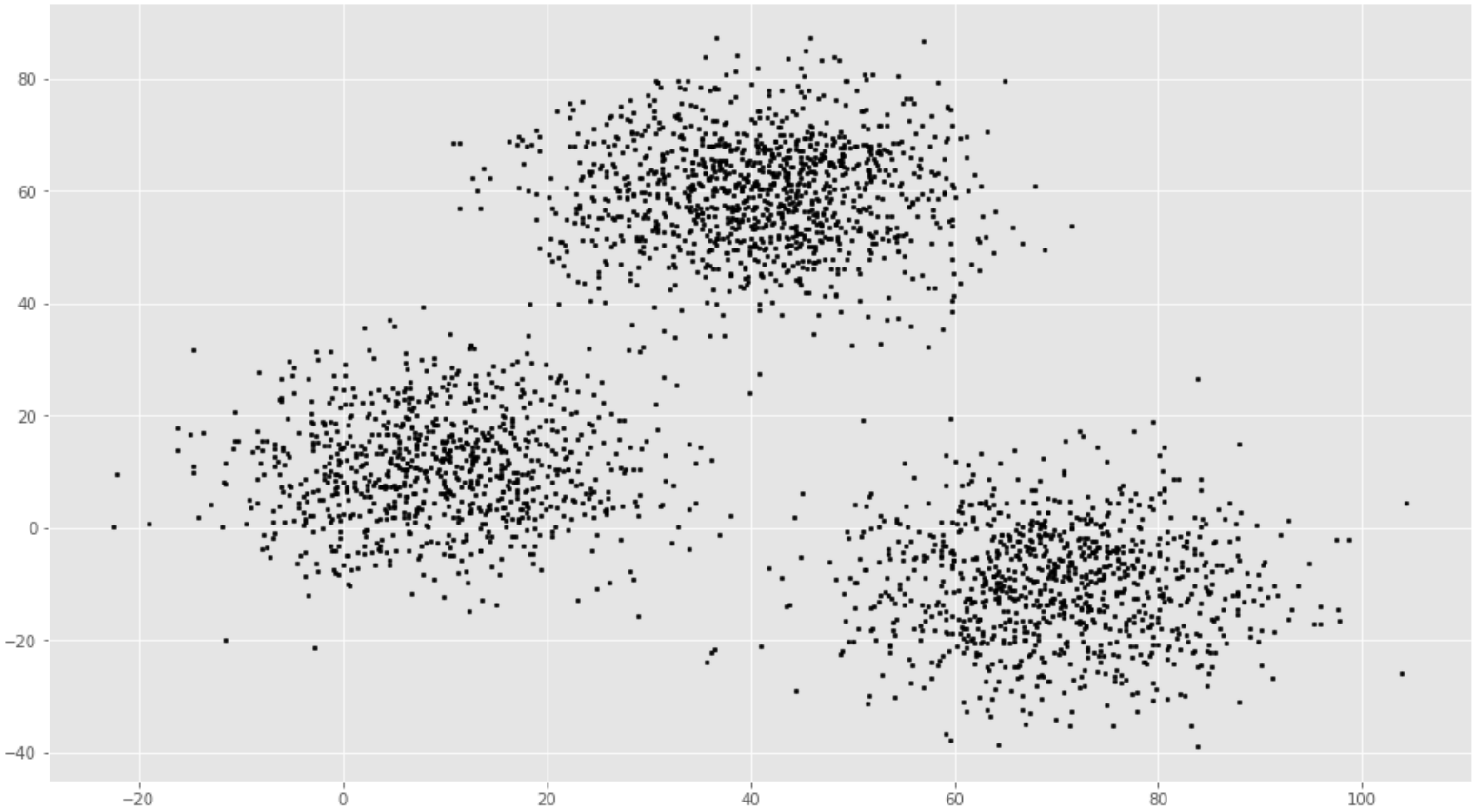
N.AUT: Número de autores

Agrupamentos hierárquicos - Exemplo

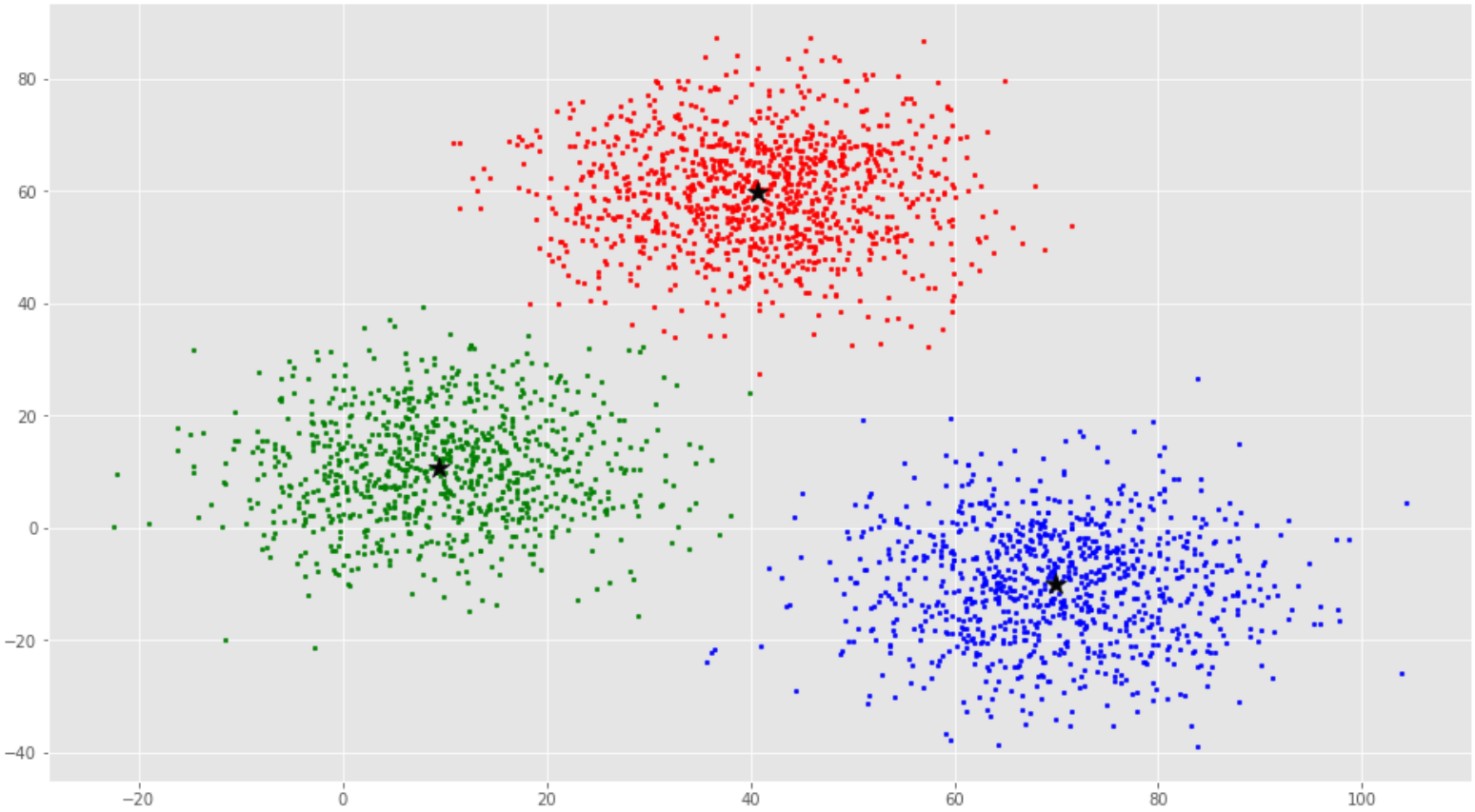


Agrupamento hierárquico via dendrograma.

Métodos de agrupamento não-hierárquico



Métodos de agrupamento não-hierárquico



Métodos de agrupamento não-hierárquico

K-médias

- Desenvolvido por MacQueen (1967)
- O objetivo é classificar cada elemento no cluster cujo centroide (vetor de médias amostrais) é o mais próximo de seu valor observado

MacQueen, James et al. Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. 1967. p. 281-297.

Métodos de agrupamento não-hierárquico

Algoritmo K-médias

1. Primeiramente são escolhidos k centroides, chamados de sementes ou protótipos, para iniciar o processo de partição dos elementos amostrais.
2. Cada elemento do conjunto de dados é comparado a cada centroide inicial, por meio de uma medida de distância (por exemplo, distância euclidiana). Cada elemento é alocado no grupo menos distante a ele.
3. Os centroides dos grupos são recalculados.
4. Os passos 2 e 3 se repetem até que todos os elementos amostrais estão alocados no grupo correto.

Métodos de agrupamento não-hierárquico

Algoritmo K-médias

1. Primeiramente são escolhidos k centroides, chamados de sementes ou protótipos, para iniciar o processo de partição dos elementos amostrais.
2. Cada elemento do conjunto de dados é comparado a cada centroide inicial, por meio de uma medida de distância (por exemplo, distância euclidiana). Cada elemento é alocado no grupo menos distante a ele.
3. Os centroides dos grupos são recalculados.
4. Os passos 2 e 3 se repetem até que todos os elementos amostrais estão alocados no grupo correto.

Métodos de agrupamento não-hierárquico

Algoritmo K-médias

1. Primeiramente são escolhidos k centroides, chamados de sementes ou protótipos, para iniciar o processo de partição dos elementos amostrais.
2. Cada elemento do conjunto de dados é comparado a cada centroide inicial, por meio de uma medida de distância (por exemplo, distância euclidiana). Cada elemento é alocado no grupo menos distante a ele.
3. Os centroides dos grupos são recalculados.
4. Os passos 2 e 3 se repetem até que todos os elementos amostrais estão alocados no grupo correto.

Métodos de agrupamento não-hierárquico

Algoritmo K-médias

1. Primeiramente são escolhidos k centroides, chamados de sementes ou protótipos, para iniciar o processo de partição dos elementos amostrais.
2. Cada elemento do conjunto de dados é comparado a cada centroide inicial, por meio de uma medida de distância (por exemplo, distância euclidiana). Cada elemento é alocado no grupo menos distante a ele.
3. Os centroides dos grupos são recalculados.
4. Os passos 2 e 3 se repetem até que todos os elementos amostrais estão alocados no grupo correto.

Métodos de agrupamento não-hierárquico

A escolha das sementes iniciais pode ser feita por meio de diferentes técnicas

- Técnicas hierárquicas aglomerativas
- Escolha aleatória
- Escolha via variáveis aleatórias
- Observações discrepantes
- Escolha pré-fixada
- k primeiras observações do conjunto de dados

Métodos de agrupamento não-hierárquico

A escolha das sementes iniciais pode ser feita por meio de diferentes técnicas

- Técnicas hierárquicas aglomerativas
- Escolha aleatória
- Escolha via variáveis aleatórias
- Observações discrepantes
- Escolha pré-fixada
- k primeiras observações do conjunto de dados

Métodos de agrupamento não-hierárquico

A escolha das sementes iniciais pode ser feita por meio de diferentes técnicas

- Técnicas hierárquicas aglomerativas
- Escolha aleatória
- Escolha via variáveis aleatórias
- Observações discrepantes
- Escolha pré-fixada
- k primeiras observações do conjunto de dados

Métodos de agrupamento não-hierárquico

A escolha das sementes iniciais pode ser feita por meio de diferentes técnicas

- Técnicas hierárquicas aglomerativas
- Escolha aleatória
- Escolha via variáveis aleatórias
- Observações discrepantes
- Escolha pré-fixada
- k primeiras observações do conjunto de dados

Métodos de agrupamento não-hierárquico

A escolha das sementes iniciais pode ser feita por meio de diferentes técnicas

- Técnicas hierárquicas aglomerativas
- Escolha aleatória
- Escolha via variáveis aleatórias
- Observações discrepantes
- Escolha pré-fixada
- k primeiras observações do conjunto de dados

Métodos de agrupamento não-hierárquico

A escolha das sementes iniciais pode ser feita por meio de diferentes técnicas

- Técnicas hierárquicas aglomerativas
- Escolha aleatória
- Escolha via variáveis aleatórias
- Observações discrepantes
- Escolha pré-fixada
- k primeiras observações do conjunto de dados