Análise Multivariada de Dados - Aula 03

Análise de Conglomerados II: Esquemas de Aglomeração

Kaique Matias de Andrade Roberto

Ciências Atuariais

HECSA - Escola de Negócios

FIAM-FAAM-FMU

Conteúdo

Conceitos que aprendemos em Aulas anteriores

Conceitos que aprendemos em Aulas anteriores

- medidas de distância;
- medidas de similaridade.

Esquemas de Aglomeração em Análise de Conglomerados

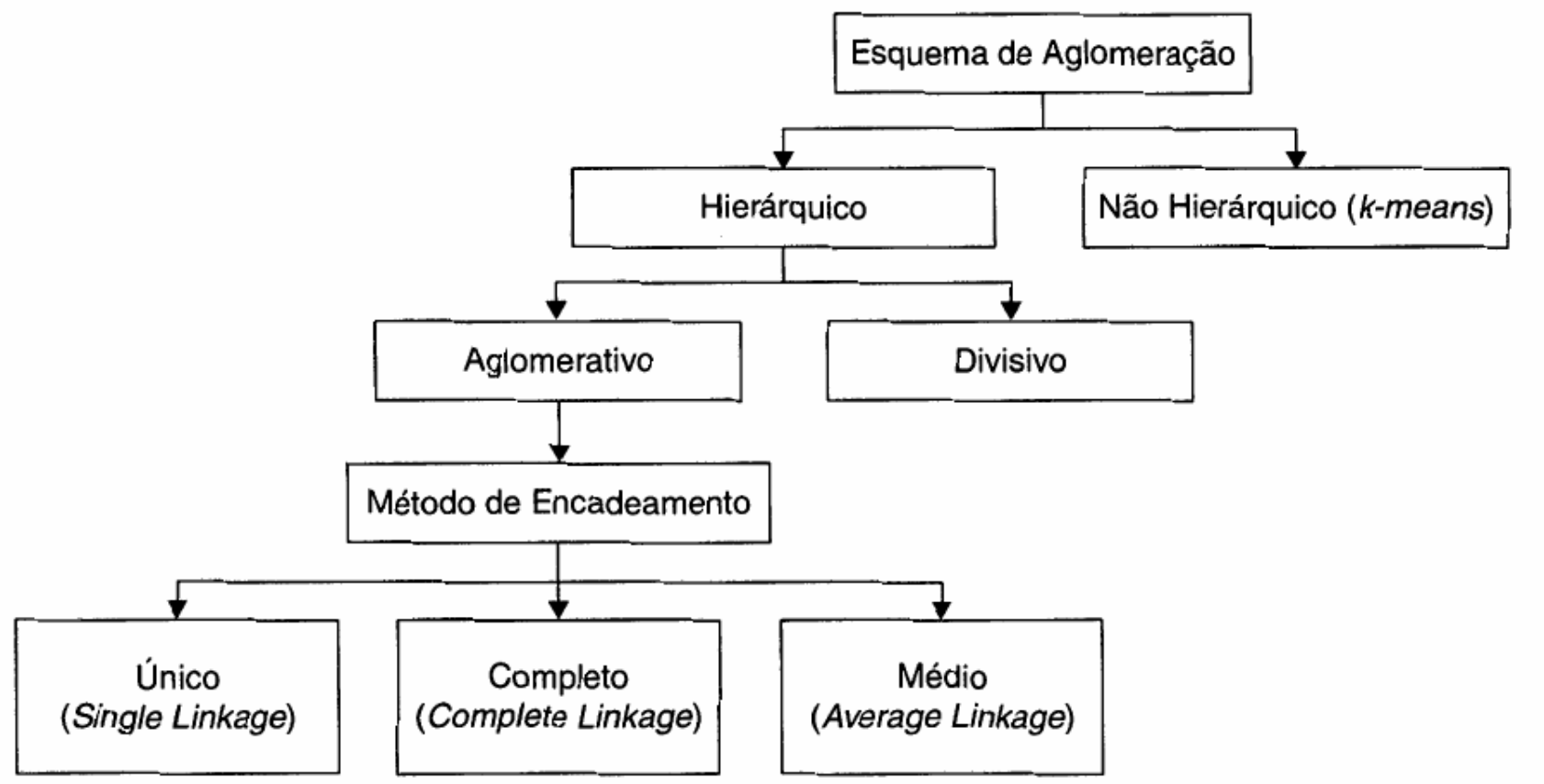
Esquemas de Aglomeração em Análise de Conglomerados

Na análise de agrupamentos, a escolha do método de aglomeração, também conhecido como esquema de aglomeração, é tão importante quanto a definição da medida de distância (ou de semelhança).

Esquemas de Aglomeração em Análise de Conglomerados

Essa decisão também precisa ser tomada com base naquilo que o pesquisador pretende em termos de objetivos de pesquisa.

Esquemas de Aglomeração em Análise de Conglomerados



Esquemas de Aglomeração em Análise de Conglomerados

Enquanto os esquemas hierárquicos caracterizam-se por privilegiar uma estrutura hierárquica (passo a passo) para a formação dos agrupamentos, os esquemas não hierárquicos utilizam algoritmos para maximizar a homogeneidade dentro de cada agrupamento, sem que haja um processo hierárquico para tal.

Esquemas de Aglomeração em Análise de Conglomerados

Os esquemas de aglomeração hierárquicos podem ser aglomerativos ou divisivos, dependendo do modo como é iniciado o processo.

Esquemas de Aglomeração em Análise de Conglomerados

Caso todas as observações sejam consideradas separadas e, a partir de suas distâncias (ou semelhanças), sejam formados grupos até que se chegue a um estágio final com apenas um agrupamento, então esse processo é conhecido como aglomerativo.

Esquemas de Aglomeração em Análise de Conglomerados

Dentre os esquemas hierárquicos aglomerativos, são mais comumente utilizados aqueles que apresentam método de encadeamento do tipo único (nearest neighbor ou single linkage), completo (furthest neighbor ou complete linkage) ou médio (between groups ou average linkage).

Esquemas de Aglomeração em Análise de Conglomerados

Por outro lado, caso todas as observações sejam consideradas agrupadas e, estágio após estágio, sejam formados grupos menores pela separação de cada observação, até que essas subdivisões gerem grupos individuais (ou seja, observações totalmente separadas), então, estaremos diante de um processo divisivo.

Esquemas de Aglomeração em Análise de Conglomerados

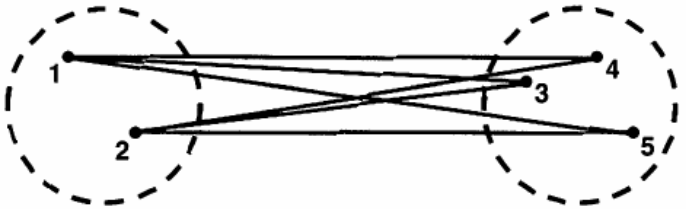
Já os esquemas de aglomeração não hierárquicos, entre os quais o mais popular é o procedimento k-means, ou k-médias, referem-se a processos em que são definidos centros de aglomeração a partir dos quais são alocadas as observações pela proximidade a eles.

Esquemas Hierárquicos

Esquemas Hierárquicos

Três são os principais métodos de encadeamento em esquemas hierárquicos aglomerativos: método de encadeamento único (nearest neighbor ou single linkage), completo (furthest neighbor ou complete linkage) e médio (between groups ou average línkage).

Esquemas Hierárquicos

Método de Encadeamento	Ilustração	Distância (Dissimilaridade)
Único *(Nearest Neighbor* ou *Single Linkage)*		d_{23}
Completo *(Furthest Neighbor* ou *Complete Linkage)*		d_{15}
Médio *(Between Groups* ou *Average Linkage)*		$\frac{d_{13}+d_{14}+d_{15}+d_{23}+d_{24}+d_{25}}{6}$

Esquemas Hierárquicos

O método de encadeamento único privilegia as menores distâncias (daí vem a nomenclatura nearest neighbor) para que sejam formados novos agrupamentos a cada estágio de aglomeração pela incorporação de observações ou grupos.

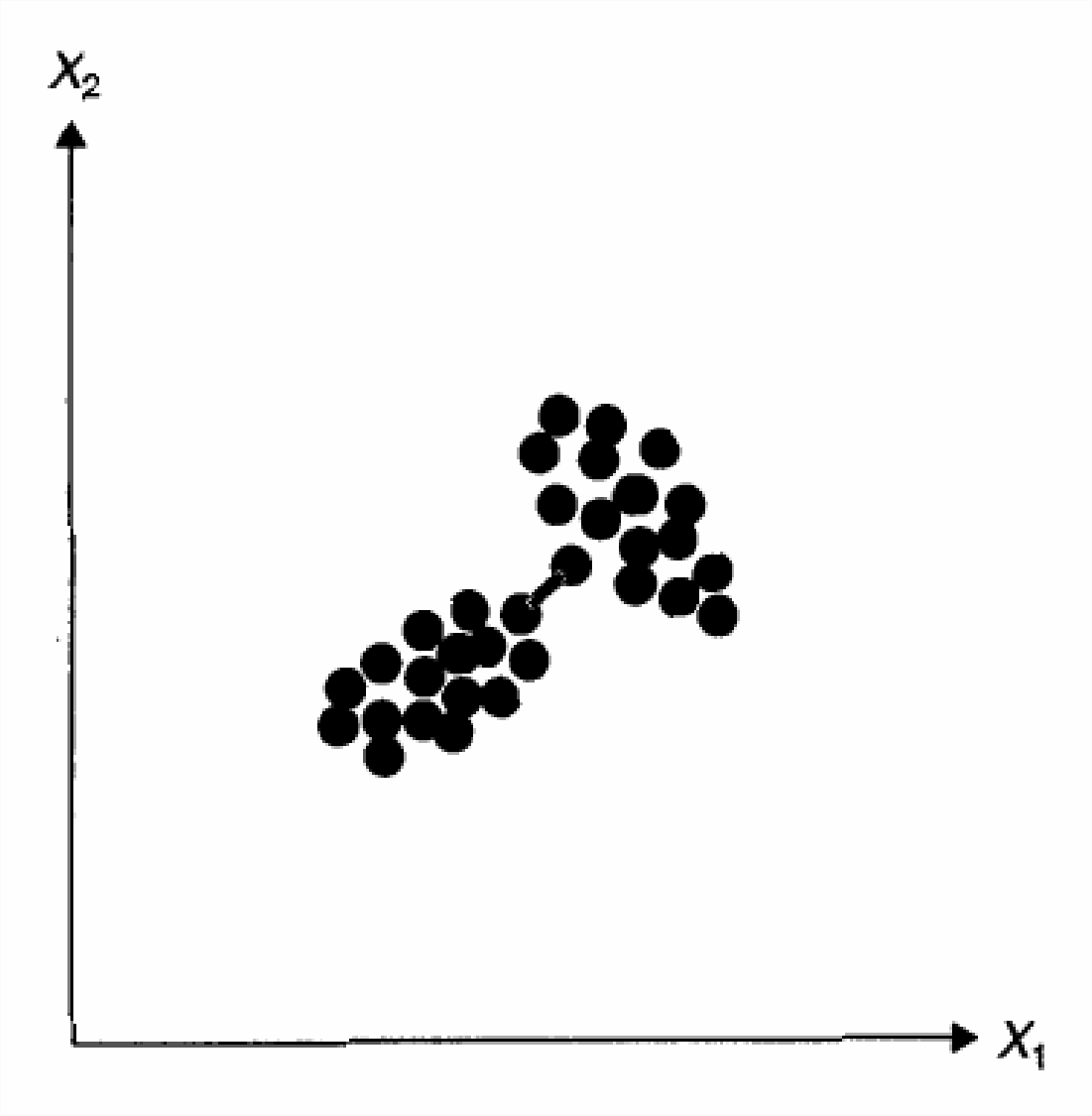
Esquemas Hierárquicos

Nesse sentido, sua aplicação é recomendável para os casos em que as observações sejam relativamente afastadas, isto é, diferentes, e deseja-se formar agrupamentos levando-se em consideração um mínimo de homogeneidade.

Esquemas Hierárquicos

Por outro lado, sua análise fica prejudicada quando da existência de observações ou agrupamentos pouco afastados entre si.

Esquemas Hierárquicos



Esquemas Hierárquicos

Já o método de encadeamento completo vai em direção contrária, ou seja, privilegia as maiores distâncias entre as observações ou grupos para que sejam formados novos agrupamentos (daí, a nomenclatura furthest neighbor).

Esquemas Hierárquicos

Dessa maneira, sua adoção é recomendável para os casos em que não exista considerável afastamento entre as observações e o pesquisador tenha a necessidade de identificar heterogeneidades entre elas.

Esquemas Hierárquicos

Por fim, no método de encadeamento médio dois grupos sofrem fusão com base na distância média entre todos os pares de observações pertencentes a esses grupos (daí, a nomenclatura average línkage).

Esquemas Hierárquicos

Dessa forma, embora ocorram alterações no cálculo das medidas de distância entre os agrupamentos, o método de encadeamento médio acaba por preservar a solução de ordenamento das observações em cada grupo, oferecida pelo método de encadeamento único, caso haja um considerável afastamento entre as observações.

Esquemas Hierárquicos

O mesmo vale em relação à solução de ordenamento oferecida pelo método de encadeamento completo, caso as observações sejam bastante próximas entre si.

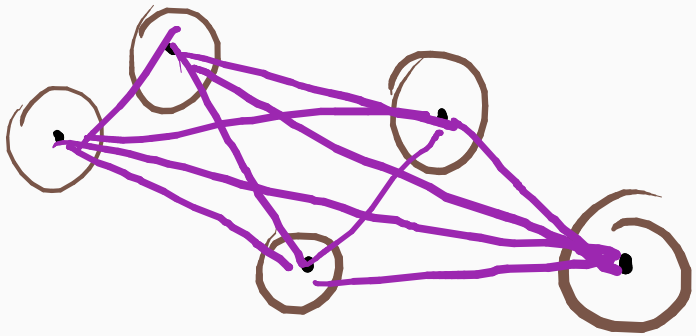
Esquemas Hierárquicos

Agora vamos entender "como fazer as contas" (isto é, qual é o algoritmo que usaremos):

Esquemas Hierárquicos

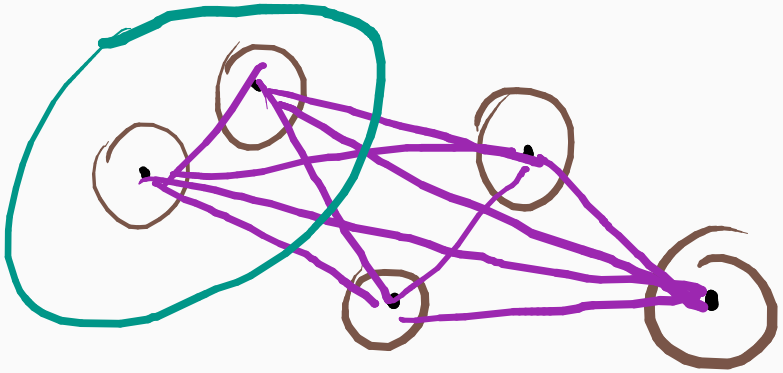
Passo 1

Sendo n a quantidade de observações de um banco de dados, devemos dar início ao esquema de aglomeração com exatamente n grupos individuais (estágio 0), de modo que teremos inicialmente uma matriz de distâncias (ou de semelhanças) D_0 composta pelas distâncias entre cada par de observações.



Passo 2

No primeiro estágio, devemos escolher a menor distância entre todas as que compõem a matriz D_0, ou seja, aquela que une as duas observações mais similares. Nesse exato momento, deixamos de ter n grupos individuais para termos $(n-1)$ grupos, sendo um deles formado por duas observações.



Passo 3

No estágio de aglomeração seguinte, devemos repetir o estágio anterior, porém agora levando em consideração a distância entre cada par de observações e entre o primeiro grupo já formado e cada uma das demais observações, com base em um dos métodos de encadeamento adotado. Em outras palavras, teremos, após o primeiro estágio de aglomeração, uma matriz D_1 com dimensões $(n-1) \times (n-1)$, em que uma das linhas será representada pelo primeiro par agrupado de observações. No segundo estágio, consequentemente, um novo grupo será formado pelo agrupamento de duas novas observações ou pela junção de determinada observação ao primeiro grupo já formado anteriormente, no primeiro estágio.

Esquemas Hierárquicos

Passo 4

O processo anterior deve ser repetido $(n-1)$ vezes, até que reste apenas um único grupo formado por todas as observações. Em outras palavras, no estágio $(n-2)$ teremos uma matriz D_{n-2} que conterá apenas a distância entre os dois últimos grupos remanescentes, antes da fusão final.

Esquemas Hierárquicos

Passo 5

Por fim, a partir dos estágios de aglomeração e das distâncias entre os agrupamentos formados, é possível construir um gráfico em formato de árvore, que resume o processo de aglomeração e explicita a alocação de cada observação em cada agrupamento. Esse gráfico é conhecido como dendrograma ou fenograma.

Esquemas Hierárquicos

Portanto, os valores que compõem as matrizes D de cada um dos estágios serão função da medida de distância escolhida e do método de encadeamento adotado.

Esquemas Hierárquicos

Imagine, em determinado estágio de aglomeração s, que um pesquisador agrupe dois clusters M e N já formados anteriormente, contendo, respectivamente, m e n observações, a fim de que seja formado o cluster MN.

Esquemas Hierárquicos

Na sequência, tem a intenção de agrupar MN com outro cluster W, com w observações.

Esquemas Hierárquicos

Como sabemos que a decisão de escolha do próximo agrupamento será sempre a menor distância entre cada par de observações ou grupos nos métodos hierárquicos aglomerativos, o esquema de aglomeração será de fundamental importância para que sejam analisadas as distâncias que comporão cada matriz D_s.

Esquemas Hierárquicos

A partir dessa lógica, apresentamos o critério de cálculo da distância inserida na matriz D_s entre os clusters MN e W (em função do método de encadeamento):

Esquemas Hierárquicos

- Método de Encadeamento Único (Nearest Neighbor ou Single Linkage)

$$d_{(MN)W} = \min\{d_{MW}, d_{NW}\}$$

em que d_{MW} e d_{NW} são distâncias entre as observações mais próximas dos clusters M e W e dos clusters N e W, respectivamente.

Esquemas Hierárquicos

- Método de Encadeamento Completo (Furthest Neighbor ou Complete Linkage)

$$d_{(MN)W} = \max\{d_{MW}, d_{NW}\}$$

em que d_{MW} e d_{NW} são distâncias entre as observações mais distantes dos clusters M e W e dos clusters N e W, respectivamente.

Esquemas Hierárquicos

- Método de Encadeamento Médio (Between Groups ou Average Linkage)

$$d_{(MN)W} = \frac{\sum_{p=1}^{m+n} \sum_{q=1}^{w} d_{pq}}{(m+n) \cdot w}$$

em que d_{pq} representa a distância entre qualquer observação p do cluster MN e qualquer observação q do cluster W, e $m+n$ e w representam, respectivamente, a quantidade de observações nos clusters MN e W.

Um Primeiro Exemplo

Um Primeiro Exemplo

Imagine que o professor de uma faculdade, bastante preocupado com a capacidade de aprendizado dos alunos em sua disciplina de métodos quantitativos, tenha o interesse em alocá-los em grupos com a maior homogeneidade possível, com base nas notas obtidas no vestibular em disciplinas consideradas quantitativas (Matemática, Física e Química).

Um Primeiro Exemplo

Nesse sentido, o professor fez um levantamento sobre essas notas, que variam de 0 a 10, e, dado que realizará uma análise de agrupamentos inicialmente de maneira algébrica, resolveu trabalhar, para efeitos didáticos, apenas com cinco alunos.

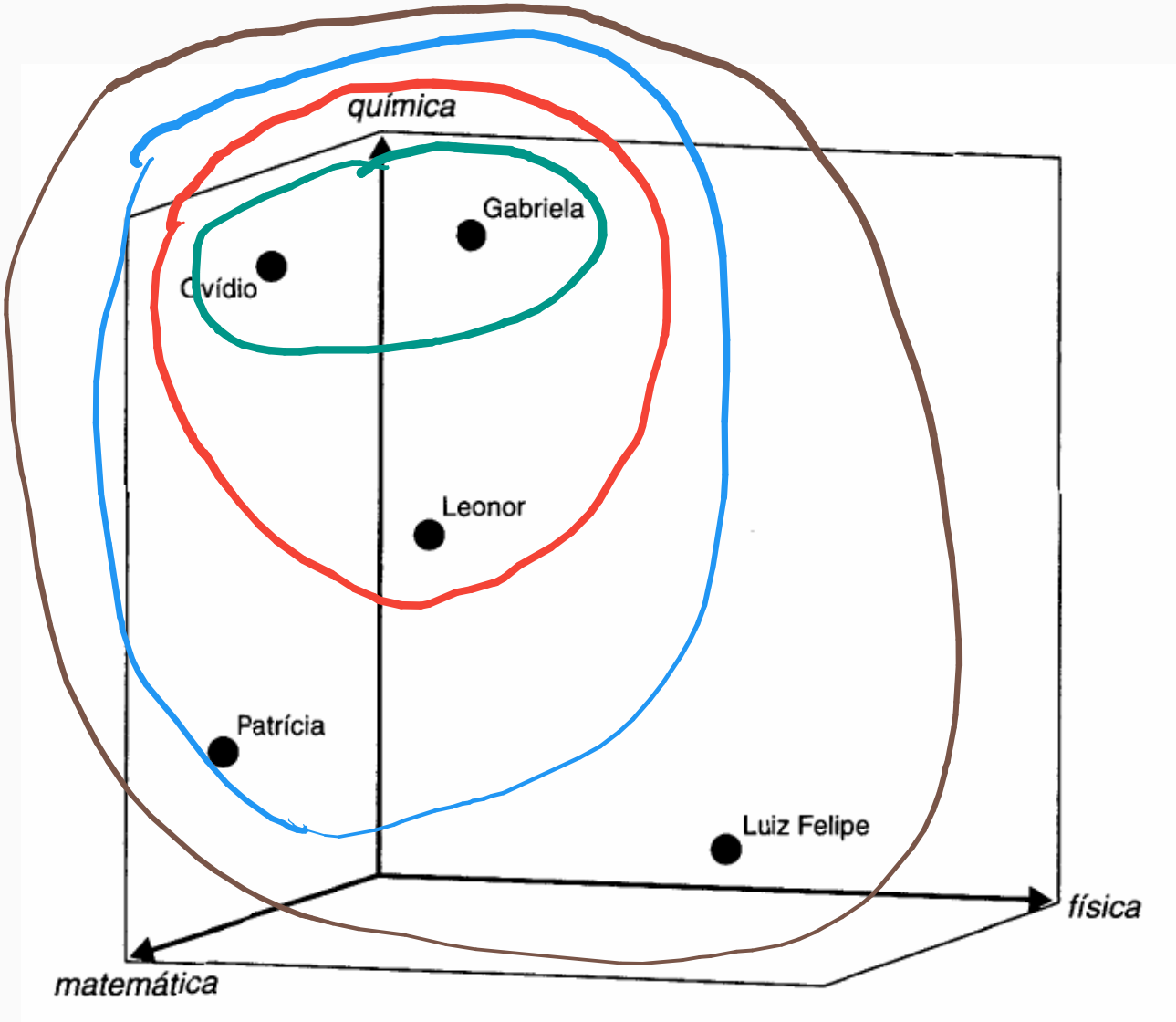
Um Primeiro Exemplo

Estudante (Observação)	Nota de Matemática (X_{1i})	Nota de Física (X_{2i})	Nota de Química (X_{3i})
Gabriela	3,7	2,7	9,1
Luiz Felipe	7,8	8,0	1,5
Patrícia	8,9	1,0	2,7
Ovídio	7,0	1,0	9,0
Leonor	3,4	2,0	5,0

Um Primeiro Exemplo

Com base nos dados obtidos, é construído o gráfico a seguir, e, como as variáveis são métricas, será adotada a medida de dissimilaridade conhecida por distância euclidiana para a análise de agrupamentos.

Um Primeiro Exemplo



Um Primeiro Exemplo

Além disso, como todas as variáveis apresentam valores na mesma unidade de medida (notas de 0 a 10), não será necessária, nesse caso, a elaboração da padronização pelo procedimento zscore.

Um Primeiro Exemplo

A partir dos dados apresentados, iremos, neste momento, elaborar uma análise de agrupamentos por meio de um esquema de aglomeração hierárquico com método de encadeamento único. Inicialmente, definimos a matriz D_0, composta pelas distâncias euclidianas entre cada par de observações, conforme segue:

Um Primeiro Exemplo

$D_0 =$

	Gabriela	Luiz Felipe	Patrícia	Ovídio	Leonor
Gabriela	0,000				
Luiz Felipe	10,132	0,000			
Patrícia	8,420	7,187	0,000		
Ovídio	3,713	10,290	6,580	0,000	
Leonor	4,170	8,223	6,045	5,474	0,000

Um Primeiro Exemplo

É importante mencionar que, neste momento inicial, cada observação é considerada um cluster individual, ou seja, no estágio 0, temos 5 clusters (tamanho da amostra).

Um Primeiro Exemplo

Em destaque, na matriz D_0, está a menor distância entre todas as observações e, portanto, serão inicialmente agrupadas, no primeiro estágio, as observações Gabriela e Ovídio, que passam a formar um novo cluster.

Um Primeiro Exemplo

Para que seja elaborado o próximo estágio de aglomeração, devemos construir a matriz D_1. em que são calculadas as distâncias entre o cluster Gabriela-Ovídio e as demais observações, ainda isoladas.

Um Primeiro Exemplo

Dessa forma, por meio do método de encadeamento único temos que:

$$d_{(\text{Gabriela-Ovídio})\text{Luiz Felipe}} = \min\{10,132; 10,290\} = 10,132$$

$$d_{(\text{Gabriela-Ovídio})\text{Patrícia}} = \min\{8,420; 6,580\} = 6,580$$

$$d_{(\text{Gabriela-Ovídio})\text{Leonor}} = \min\{4,170; 5,474\} = 4,170$$

A matriz D_1 encontra-se a seguir:

Um Primeiro Exemplo

$D_1 =$

	Gabriela Ovídio	Luiz Felipe	Patrícia	Leonor
Gabriela Ovídio	0,000			
Luiz Felipe	10,132	0,000		
Patrícia	6,580	7,187	0,000	
Leonor	4,170	8,223	6,045	0,000

Um Primeiro Exemplo

Da mesma forma, na matriz D_1 está em destaque a menor distância entre todas. Portanto, no segundo estágio, é inserida a observação Leonor no cluster já formado Gabriela-Ovídio. As observações Luiz Felipe e Patrícia permanecem ainda isoladas.

Um Primeiro Exemplo

Para que possamos dar o próximo passo, devemos construir a matriz D_2, em que são calculadas as distâncias entre o cluster Gabriela-Ovídio-Leonor e as duas observações remanescentes. Analogamente, temos que:

$$d_{(\text{Gabriela-Ovídio-Leonor})\text{Luiz Felipe}} = \min\{10,132; 8,223\} = 8,223$$

$$d_{(\text{Gabriela-Ovídio-Leonor})\text{Patrícia}} = \min\{6,580; 6,045\} = 6,045$$

A matriz D_2 pode ser escrita como:

Um Primeiro Exemplo

$D_2 =$

	Gabriela Ovídio Leonor	Luiz Felipe	Patrícia
Gabriela Ovídio Leonor			
Luiz Felipe	8,223	0,000	
Patrícia	6,045	7,187	0,000

Um Primeiro Exemplo

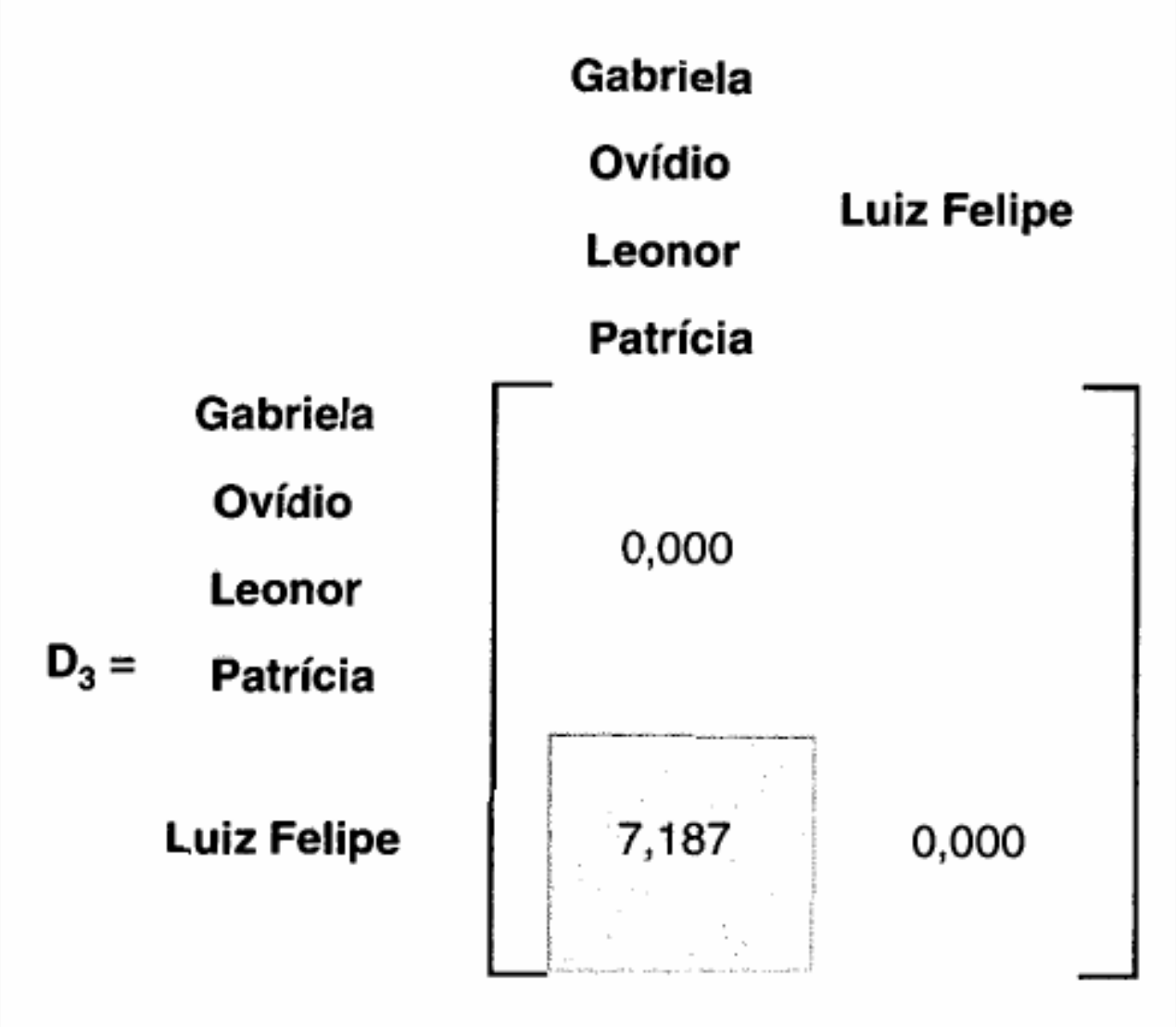
No terceiro estágio de aglomeração, é incorporada a observação Patrícia no cluster Gabriela-Ovídio-Leonor, visto que a correspondente distância é a menor entre todas as apresentadas na matriz D_2.

Um Primeiro Exemplo

Portanto, podemos escrever a matriz D_3, que se encontra na sequência, levando em consideração o seguinte critério:

$$^{d}(\text{Gabriela-Ovídio-Leonor-Patrícia})\text{Luiz Felipe} = \min\{8,223; 7,187\} = 7,187$$

Um Primeiro Exemplo



Um Primeiro Exemplo

Por fim, no quarto e último estágio, todas as observações estão alocadas no mesmo agrupamento, encerrando-se, assim, o processo hierárquico. A seguir apresenta um resumo desse esquema de aglomeração elaborado por meio do método de encadeamento único.

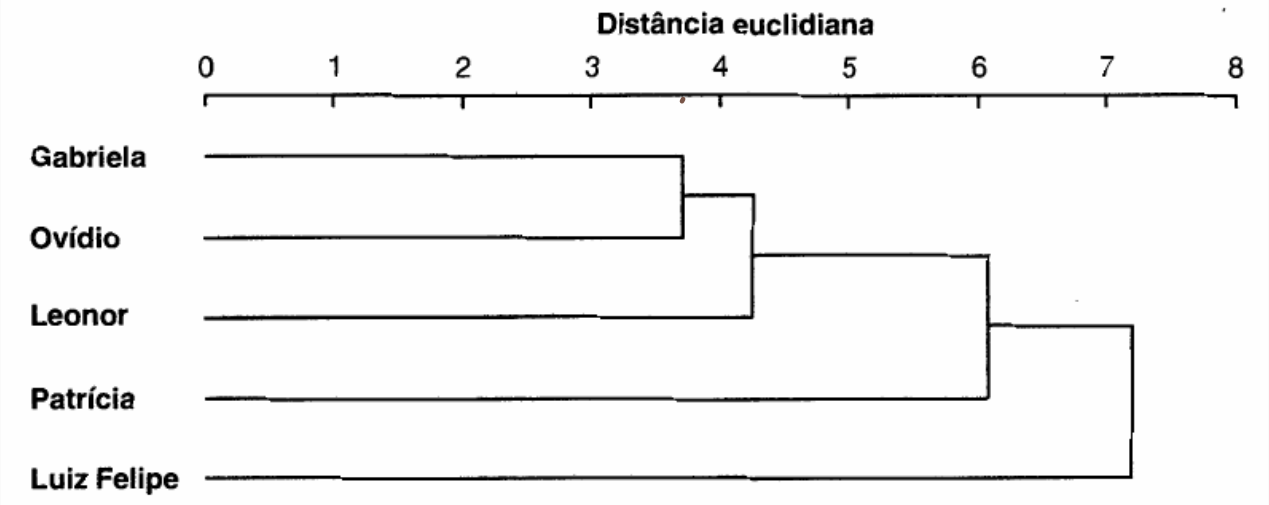
Um Primeiro Exemplo

Estágio	Agrupamento	Observação Agrupada	Menor Distância Euclidiana
1	Gabriela	Ovídio	3,713
2	Gabriela – Ovídio	Leonor	4,170
3	Gabriela – Ovídio – Leonor	Patrícia	6,045
4	Gabriela – Ovídio – Leonor – Patrícia	Luiz Felipe	7,187

Um Primeiro Exemplo

Com base nesse esquema de aglomeração, podemos construir um gráfico em formato de árvore, conhecido como dendrograma ou fenograma, cujo intuito é ilustrar o passo a passo dos agrupamentos e facilitar a visualização da alocação de cada observação em cada estágio.

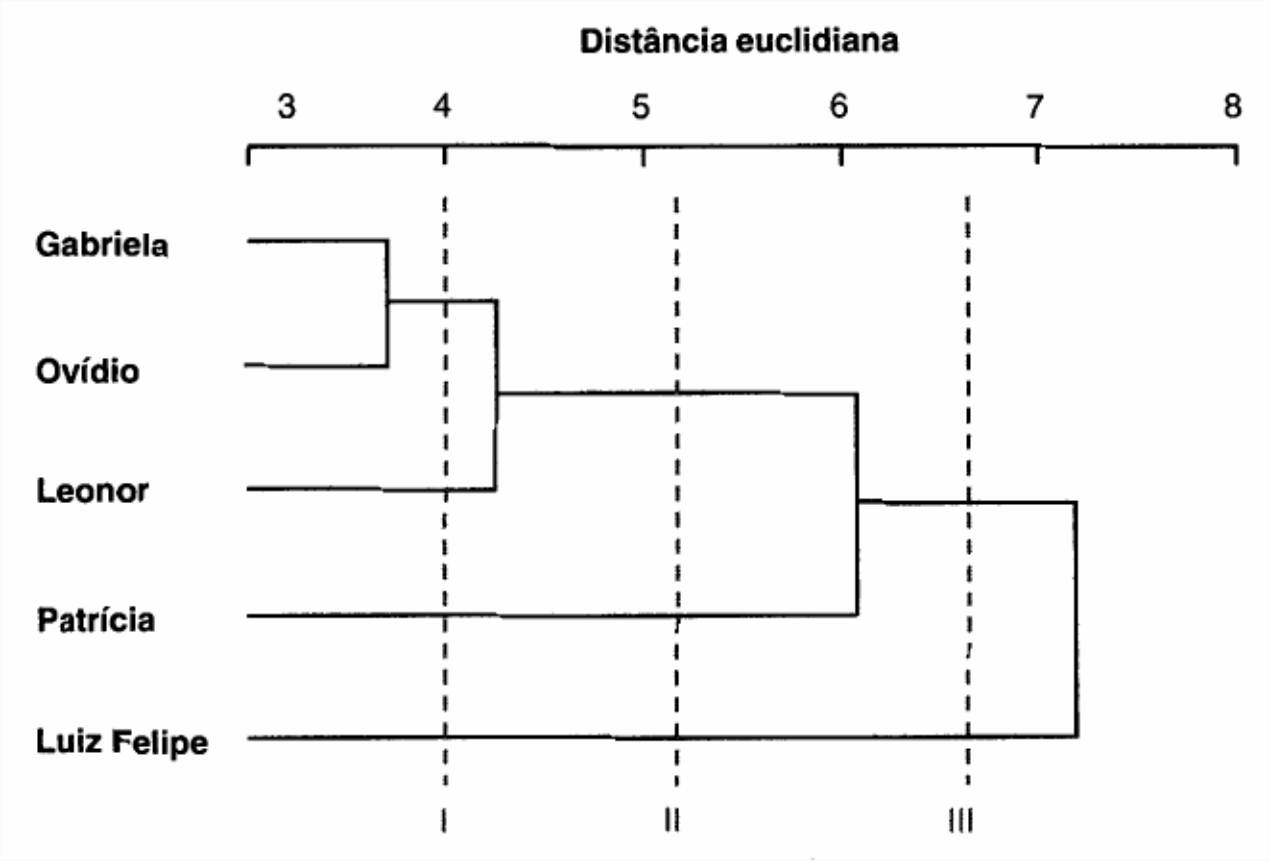
Um Primeiro Exemplo



Um Primeiro Exemplo

Inicialmente, traçamos três linhas (I, II e III) ortogonais às linhas do dendrograma que permitem identificar as quantidades de agrupamentos em cada estágio de aglomeração, bem como as observações em cada cluster.

Um Primeiro Exemplo



Um Primeiro Exemplo

Assim, a linha l "corta"o dendrograma imediatamente após o primeiro estágio de aglomeração e, neste momento, podemos verificar que existem quatro clusters (quatro encontros com as linhas horizontais do dendrograma), um deles formado pelas observações Gabriela e Ovídio, e os demais, pelas observações individuais.

Um Primeiro Exemplo

Já a linha II encontra três linhas horizontais do dendrograma, o que significa que, após o segundo estágio, em que foi incorporada a observação Leonor ao agrupamento já formado Gabriela-Ovídio, existem três clusters.

Um Primeiro Exemplo

Por fim, a linha III é desenhada imediatamente após o terceiro estágio, em que ocorre o agrupamento da observação Patrícia com o cluster Gabriela-Ovídio-Leonor.

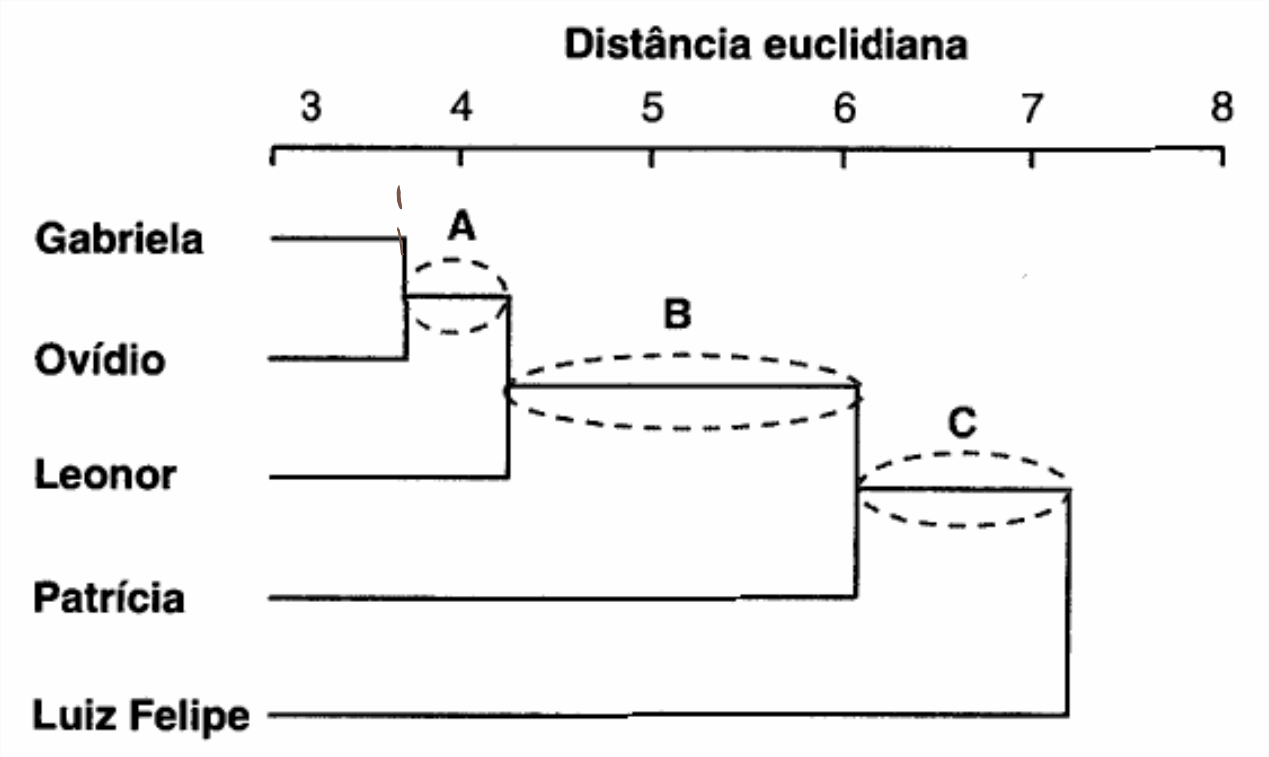
Um Primeiro Exemplo

Como são identificados dois encontros entre essa linha e as linhas horizontais do dendrograma, verificamos que a observação Luiz Felipe permanece isolada, enquanto as demais formam um único agrupamento.

Um Primeiro Exemplo

Além de propiciar o estudo sobre a quantidade de clusters em cada estágio de aglomeração, bem como sobre a alocação das observações, o dendrograma também permite que o pesquisador analise a magnitude dos saltos de distância para que se estabeleçam os agrupamentos.

Um Primeiro Exemplo



Um Primeiro Exemplo

Um salto com magnitude elevada, em comparação aos demais, pode indicar que determinada observação ou cluster consideravelmente distintos estejam incorporados a agrupamentos já formados, o que fornece subsídios ao estabelecimento de uma solução da quantidade de agrupamentos sem a necessidade de um próximo estágio de aglomeração.

Um Primeiro Exemplo

Além disso, como a quantidade de agrupamentos é importante para a elaboração de esquemas de aglomeração não hierárquicos, essa informação (considerada output do esquema hierárquico) pode servir de input para o procedimento k-means.

Um Primeiro Exemplo

A Figura apresenta três saltos de distância (A, B e C), referentes a cada um dos estágios de aglomeração, e, a partir de sua análise, podemos verificar que o salto B, que representa a incorporação da observação Patrícia ao cluster já formado Gabriela-Ovídio-Leonor, é o maior dos três.

Um Primeiro Exemplo

Assim, caso haja a intenção de definir uma quantidade interessante de agrupamentos nesse exemplo, o pesquisador pode optar pela solução com três clusters, sem o estágio em que é incorporada a observação Patrícia, visto que possivelmente apresenta características não tão homogêneas que inviabilizam sua inclusão no cluster já formado, dado o grande salto de distância.

Um Primeiro Exemplo

Nesse caso, portanto, teríamos um agrupamento formado por Gabriela, Ovídio e Leonor, outro formado apenas por Patrícia e um terceiro formado apenas por Luiz Felipe.

Um Primeiro Exemplo

Um critério muito útil para a identificação da quantidade de clusters, consiste em identificar um considerável salto de distância (quando possível) e definir a quantidade de agrupamentos formados no estágio de aglomeração imediatamente anterior ao grande salto, visto que saltos muito elevados podem incorporar observações com características não tão homogêneas.

Um Primeiro Exemplo

Além disso, é relevante também comentar que, caso os saltos de distância de um estágio para outro sejam pequenos, pela existência de variáveis com valores muito próximos para as observações, o que pode dificultar a leitura do dendrograma, o pesquisador poderá fazer uso da distância quadrática euclidiana, a fim de que os saltos fiquem mais nítidos e explicitados, facilitando a identificação dos agrupamentos no dendrograma e propiciando melhores argumentos para a tomada de decisão.

Esquemas Não-Hierárquicos

Esquemas Não-Hierárquicos

Um esquema de aglomeração não hierárquico requer a estipulação, a priori, da quantidade de clusters a partir da qual serão definidos os centros de aglomeração e alocadas as observações.

Esquemas Não-Hierárquicos

É por essa razão que se recomenda a elaboração de um esquema de aglomeração hierárquico preliminarmente à de um esquema não hierárquico, quando não há uma estimativa razoável da quantidade de clusters que podem ser formados a partir das observações do banco de dados e com base nas variáveis em estudo.

Esquemas Não-Hierárquicos: Kmeans

Dentre os esquemas de aglomeração não hierárquicos, o procedimento k-means é o mais utilizado por pesquisadores em diversos campos do conhecimento.

Esquemas Não-Hierárquicos: Kmeans

Dado que a quantidade de clusters é definida preliminarmente pelo pesquisador, esse procedimento pode ser elaborado após a aplicação de um esquema hierárquico aglomerativo quando não se tem ideia da quantidade de clusters que podem ser formados e, nessa situação, o output obtido por esse procedimento pode servir de input para o não hierárquico.

Esquemas Não-Hierárquicos: Kmeans

Agora vamos entender "como fazer a conta" para o kmeans:

Esquemas Não-Hierárquicos: Kmeans

Passo 1

Definimos a quantidade inicial de clusters e os respectivos centroides. O objetivo é dividir as observações do banco de dados em K clusters, de modo que aquelas dentro de cada cluster estejam mais próximas entre si se comparadas a qualquer outra pertencente a um diferente. Para tal, as observações precisam arbitrariamente ser alocadas nos K clusters, a fim de que possam ser calculados os respectivos centroides.

Esquemas Não-Hierárquicos: Kmeans

Passo 2

Devemos selecionar determinada observação que se encontra mais próxima de um centroide e realocá-la nesse cluster. Neste momento, outro cluster acaba de perder aquela observação, e, portanto, devem ser recalculados os centroides do cluster que a recebe e os do cluster que a perde.

Esquemas Não-Hierárquicos: Kmeans

Passo 3

Devemos proceder com o passo anterior até que não seja mais possível realocar observação alguma por maior proximidade a um centroide de outro cluster.

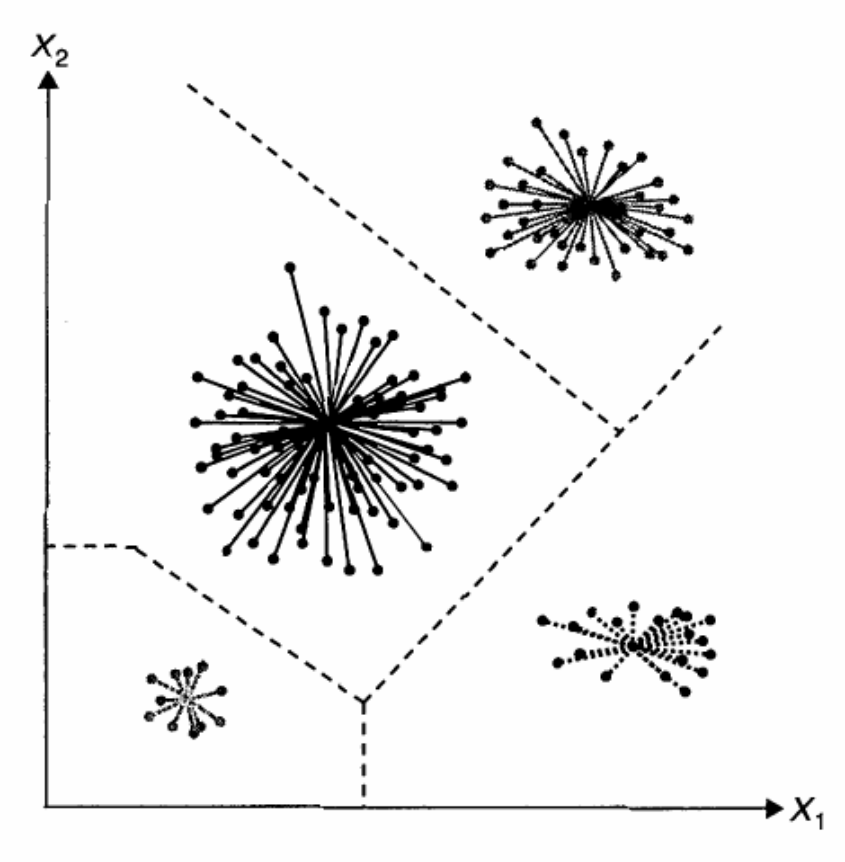
Esquemas Não-Hierárquicos: Kmeans

A coordenada $\overline{x}$ de um centroide deve ser recalculada quando da inclusão ou exclusão de determinada observação p no respectivo cluster.

Esquemas Não-Hierárquicos: Kmeans

A Figura abaixo apresenta, para duas variáveis (X_1 e X_2), uma situação hipotética que representa o término do procedimento k-means, em que não é mais possível realocar observação alguma pelo fato de não mais haver maiores proximidades a centroides de outros agrupamentos.

Esquemas Não-Hierárquicos: Kmeans



Esquemas Não-Hierárquicos: Kmeans

Além disso, lembramos que as variáveis devem ser padronizadas antes da elaboração do procedimento k-means, assim como nos esquemas de aglomeração hierárquicos, caso os respectivos valores não estejam na mesma unidade de medida.

Esquemas Não-Hierárquicos: Kmeans

Finalmente, após a conclusão desse procedimento, é importante que o pesquisador estude se os valores de determinada variável métrica diferem-se entre os grupos definidos, ou seja, se a variabilidade entre os clusters é significativamente superior à variabilidade interna a cada cluster.

Esquemas Não-Hierárquicos: Kmeans

O teste F da análise de variância de um fator (em inglês, one-way analysis of variance ou one-way ANOVA) permite que seja elaborada essa análise.

Comentários Finais

Comentários Finais

Em resumo, na aula de hoje nós:

- descrevemos os esquemas de aglomeração;
- descrevemos os algoritmos para esquemas hierárquicos;
- realizamos alguns exemplos para esquemas hierárquicos
- descrevemos o algoritmo para o esquema não-hierárquico k-means.

Comentários Finais

Nas próximas aulas vamos continuar lidando com Esquemas de Aglomeração.

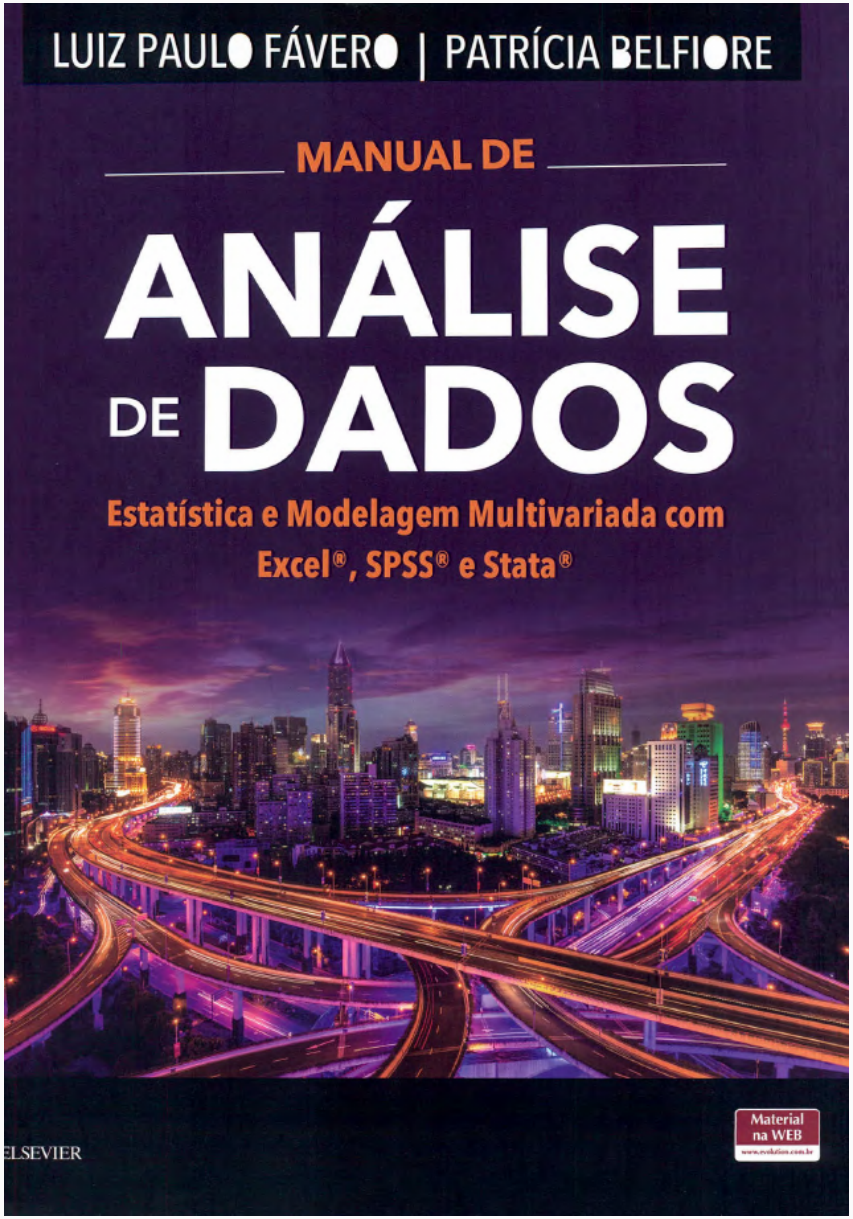
Comentários Finais

ATIVIDADE PARA ENTREGAR (E COMPOR A NOTA N1)

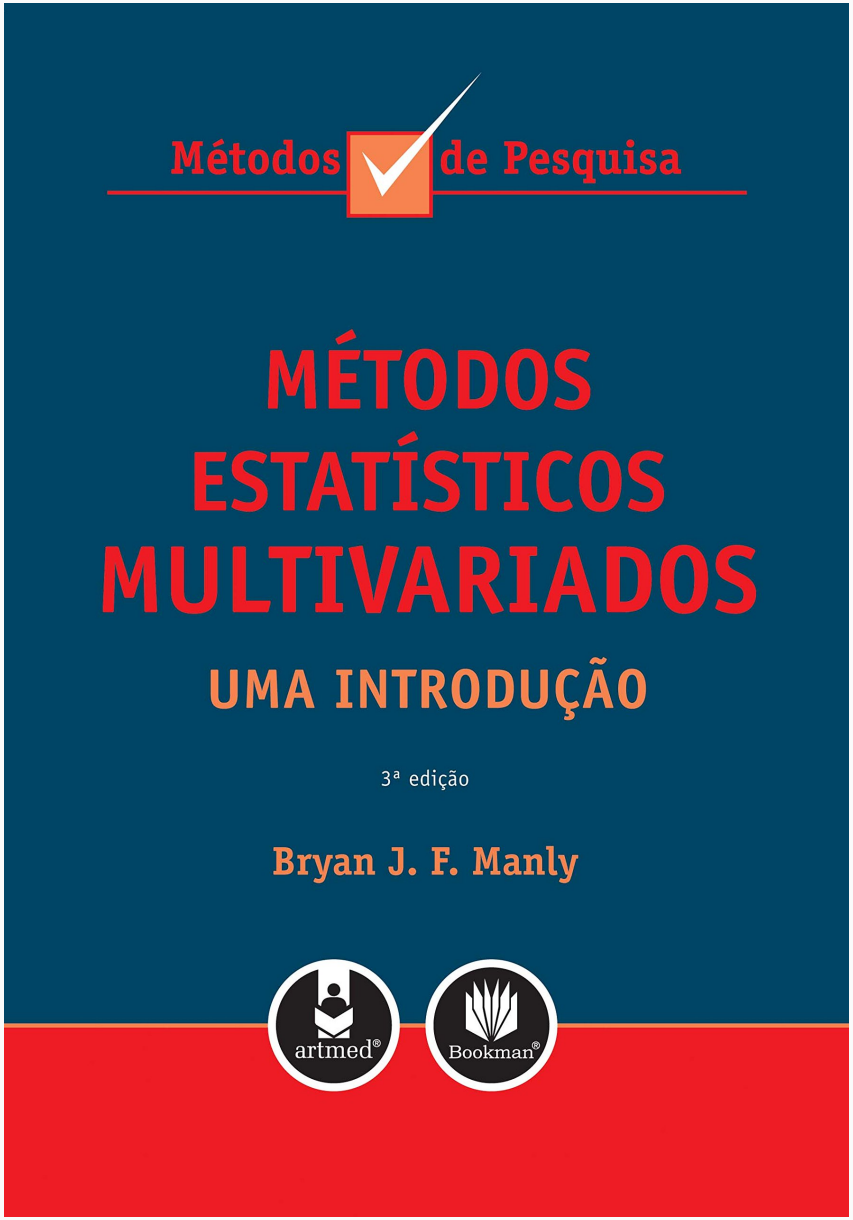
Resolva em grupos de até 4 integrantes os Exercícios 3.1, 3.2, 3.3.

Referências

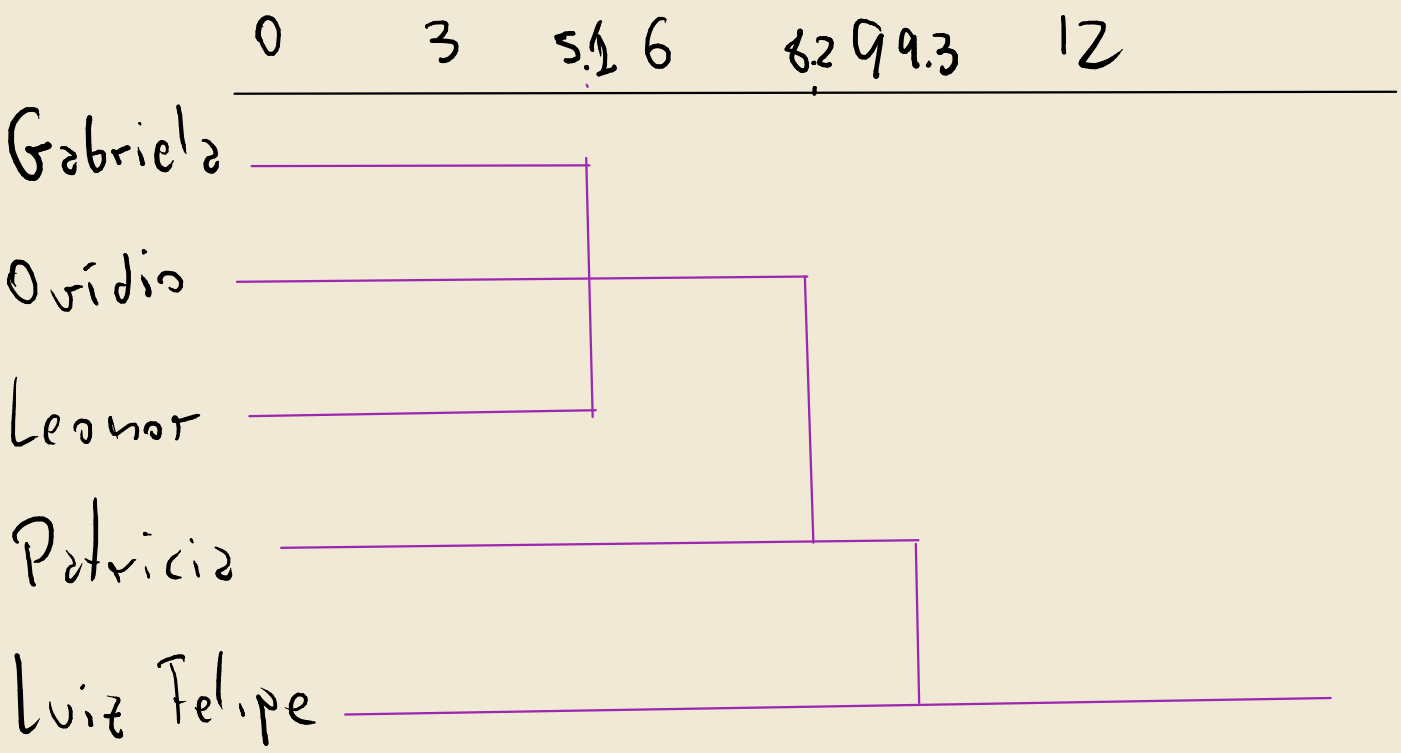
Referências

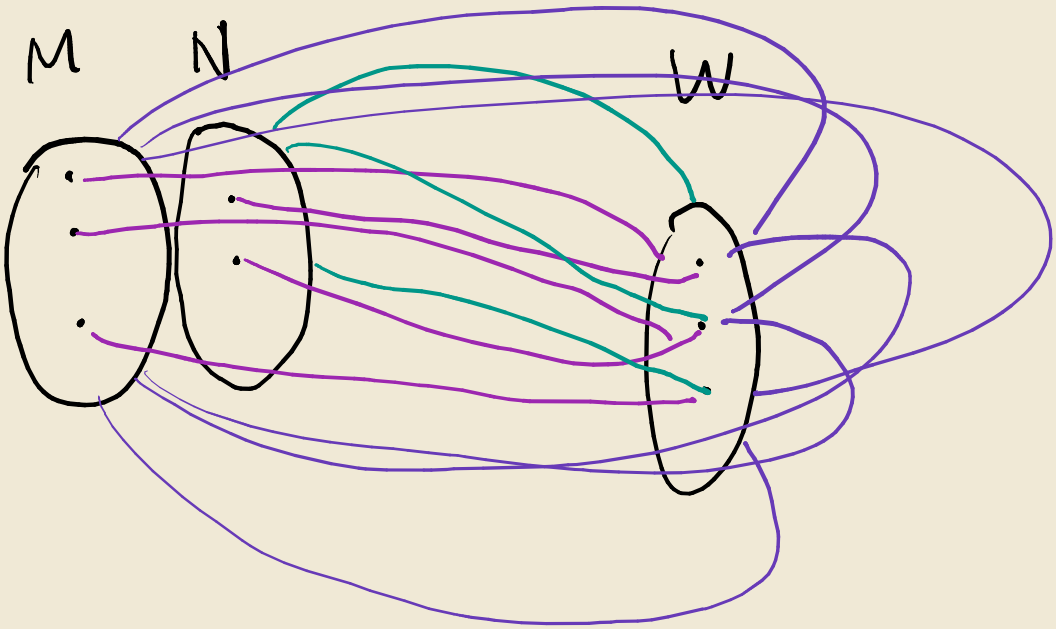


Referências



Bons Estudos!


0
3
5.1
6
8.2
9
9.3
12
Gabriela
Ovídio
Leonor
Patrícia
Luiz Felipe


M
N
W