9. ANÁLISE DE AGUPAMENTOS (*CLUSTER*)

9.1 INTRODUÇÃO

A Análise de Agrupamentos é uma técnica distinta dos Métodos de Classificação (Análise Discriminante, Regressão Logística). Na Classificação temos um número de grupos conhecidos, e o objetivo era alocar uma nova observação em um destes grupos (Supervisionado). Agrupar é uma técnica mais primitiva, no sentido de que nenhuma suposição é feita quanto ao número de grupos ou estrutura de agrupamento. O agrupamento (não supervisionado) é feito com base na similaridade ou distância.

9.2 MEDIDAS DE SIMILARIDADE E DISSIMILARIDADE

Quando itens (unidades ou casos) são agrupados, a proximidade é usualmente indicada por uma espécie de **distância**. Por outro lado, as variáveis são usualmente agrupadas com base nos **coeficientes de correlação** ou outras medidas de associação.

- **Similaridade:** quanto maior o valor observado mais parecidos são os objetos. Ex.: o coeficiente de correlação.

- **Dissimilaridade:** quanto maior o valor observado menos parecidos (mais dissimilares) serão os objetos. Ex.: distância Euclidiana.

Algumas distâncias:

(1) **Distância Euclidiana:** essa é provavelmente a mais conhecida e usada medida de distância. Ela simplesmente é a distância geométrica no espaço multidimensional. Ela é calculada como:

$$d(\underline{\mathbf{x}},\underline{\mathbf{y}}) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

(2) Quadrado da distância Euclidiana:

$$d(\underline{\mathbf{x}},\underline{\mathbf{y}}) = \sum_{i=1}^{p} (x_i - y_i)^2$$

(3) Distância city-block (Manhattan)

$$d(\underline{\mathbf{x}},\underline{\mathbf{y}}) = \sum_{i=1}^{p} |x_i - y_i|$$

(4) Distância de Mahalanobis (distância estatística)

$$d(\underline{\mathbf{x}},\underline{\mathbf{y}}) = \sqrt{(\underline{\mathbf{x}}-\underline{\mathbf{y}})'S^{-1}(\underline{\mathbf{x}}-\underline{\mathbf{y}})} = \sqrt{\frac{(x_1-y_1)^2}{s_1^2}+\ldots+\frac{(x_p-y_p)^2}{s_p^2}}$$

(5) Métrica de Minkowski

$$d(\underline{x},\underline{y}) = \sqrt[n]{|x_1-y_1|^n+|x_2-y_2|^n+\ldots+|x_p-y_p|^n} = \sqrt[n]{\sum_{i=1}^{p}|x_i-y_i|^n}$$

9.3 MÉTODO DE AGRUPAMENTO HIERÁRQUICO

Neste método, no início existe tantos grupos quanto objetos (itens). Diversos objetos semelhantes são agrupados primeiro, estes grupos iniciais são fundidos de acordo com as suas similaridades, eventualmente, relaxando no critério de similaridade os sub-grupos vão se unindo a outros sub-grupos até formar um grupo único.

O procedimento é o seguinte:

(1) No início tem-se n grupos, sendo que cada um é formado por um único objeto; calcula-se a **matriz simétrica de distâncias** n×n, D = (d_{ij}), onde d_{ij} é a distância ou similaridade entre o objeto i e o objeto j.

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

onde: $d_{11} = d_{22} = \ldots = d_{nn} = 0$

(2) Na matriz D, acha-se o par de grupos mais próximo (menor distância) e junta-se estes grupos.

(3) O novo grupo formado é denominado, por ex., (A,B), se os grupos primitivos do par são A e B. Nova matriz de distâncias é construída, simplesmente apagando-se as linhas e colunas correspondentes aos grupos A e B e adicionando-se a linha e a coluna dadas pelas distâncias entre (AB) e os grupos remanescentes.

(4) Repete-se os passos 2 e 3 (n-1) vezes observando-se as identidades dos grupos que são agrupados.

9.4 LIGAÇÕES

No item anterior descreveu-se o Método Aglomerativo Hierárquico e ali é feito referência ao modo de se agrupar os objetos semelhantes, sendo este agrupamento feito por meio de **ligações**. Os tipos de ligações mais comuns são: Ligações Simples (vizinho mais próximo), Ligações Completas (vizinho mais distante), Método das Médias das Distâncias, Método do Centróide, Método de Ward. Vamos ver detalhadamente os dois primeiros tipos de ligações:

(1º) Ligações simples (vizinho mais próximo)

Nas ligações simples o agrupamento é feito juntando-se **dois grupos** com **menor distância ou maior similaridade**. Uma vez formado o novo grupo, por exemplo, (AB), na ligação simples, a distância entre (AB) e algum outro grupo C é calculado:

$$d_{(AB)C} = \text{mín}\{d_{AC}, d_{BC}\}$$

Os resultados obtidos são dispostos graficamente em um **diagrama em árvore ou dendrograma** que possui uma escala para se observar os níveis.

Exemplo 1: Seja a matriz de distâncias $D = \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{bmatrix}$. Construa o dendrograma.

(2º) Ligações completas (vizinho mais distante)

Na ligação completa o procedimento é muito semelhante ao da ligação simples, com uma única exceção. O algoritmo aglomerativo começa determinando a menor distância d_{ik}, constrói-se a matriz de distâncias $D = (d_{ik})$ e os grupos vão se juntando. Se A e B são dois grupos de um único elemento, tem-se (A,B) como novo grupo. A distância entre (A,B) e outro grupo C é dada por

$$d[(A,B),C] = \max\{d_{(AC)}, d_{(BC)}\}$$

Exemplo 2: Com os dados do exemplo 1 construa o dendrograma, adotando agora as ligações completas.

Exemplo 3: Para os dados seguintes:

Objeto	X	Y
1	1	2
2	2,5	4,5
3	2	2
4	4	1,5
5	4	2,5

(a) Calcular as distâncias Euclidianas.

(b) Montar a matriz de distâncias.

(c) Utilizando a ligações simples (vizinho mais próximo), construir o dendrograma.

9.5 AVALIAÇÃO DA FORMAÇÃO DOS AGRUPAMENTOS

Uma forma de avaliar a validade da informação gerada pela função ligação é compará-la com os dados originais da distância. Se o agrupamento é válido, a ligação dos objetos no agrupamento tem uma forte correlação com as distâncias entre objetos no vetor de distâncias. A **função cofenética** compara esses dois conjuntos de valores e calcula sua correlação. A melhor solução para um agrupamento tem **correlação cofenética** mais próxima de 1.

Exemplo 4: Avaliar a formação dos agrupamentos para o exemplo 3.

9.6 MÉTODO DE AGRUPAMENTO NÃO-HIERÁRQUICO

O agrupamento não-hierárquico é uma técnica usada quando se deseja formar k grupos de itens ou objetos. O Método Aglomerativo Não-Hierárquico mais usado é o algoritmo das **k-médias.**

O método das k-médias é composto por 3 etapas:

1ª) Partição arbitrária dos itens em k grupos iniciais;

2ª) Re-alocar cada item no grupo cuja média (centróide) esteja mais próximo. Em geral é usada a distância Euclidiana. O centróide é recalculado para o grupo que recebeu novo item e para o grupo que perdeu algum item;

3ª) Repete-se a 2ª etapa até que não restem mais re-alocações a serem feitas.

Exemplo 5: São conhecidas as medidas de duas variáveis X_1 e X_2 para cada um dos itens: A, B, C e D. Os dados estão na tabela seguinte. Agrupe os itens em 2 grupos, de modo que os itens de cada grupo estejam o mais próximo possível um dos outros.

ITEM	X_1	X_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

Exemplo 6: Para os dados abaixo

Marca do Carro	Custo (×1000)	Pot. Motor (CV)	Consumo (km/l)
A	15	60	10
B	20	70	8
C	18	65	10
D	25	80	7

Padronizar cada uma das variáveis e determinar:

(a) A matriz de distâncias (distância Euclidiana).

(b) Agrupar segundo o Método das Ligações Simples (vizinho mais próximo).

(c) Construir o dendrograma.

(d) Agrupe os carros em dois grupos utilizando o método das k-médias.

Solução:

Médias: $\overline{X}_1 = 19{,}5$; $\overline{X}_2 = 68{,}75$; $\overline{X}_3 = 8{,}75$

Desvios padrões: $s_1 = 3{,}64$; $s_2 = 7{,}40$; $s_3 = 1{,}30$

Marca do Carro	Z_1	Z_2	Z_3
A	-1,24	-1,18	0,96
B	0,14	0,17	-0,58
C	-0,41	-0,51	0,96
D	1,51	1,52	-1,35