Aline Maria Soares de Albuquerque

Técnicas de Agrupamento: Uma Proposta de Atividades Para o Ensino Básico

Maringá-PR, Brasil

2019

Aline Maria Soares de Albuquerque

Técnicas de Agrupamento: Uma Proposta de Atividades Para o Ensino Básico

Dissertação de Mestrado apresentada ao Departamento de Matemática da Universidade Estadual de Maringá, como parte dos requisitos exigidos para obtenção do título de Mestre pelo Programa de Mestrado Profissional em Matemática em Rede Nacional.

Universidade Estadual de Maringá- UEM

Departamento de Matemática

Programa de Mestrado Profissional em Matemática em Rede Nacional

Orientador: Prof. Dr. Rodrigo Martins

Maringá-PR, Brasil

2019

Dados Internacionais de Catalogação-na-Publicação (CIP)
(Biblioteca Central - UEM, Maringá - PR, Brasil)

Soares de Albuquerque, Aline Maria

S676t Técnicas de agrupamento : uma proposta de atividades para o ensino básico / Aline Maria Soares de Albuquerque. -- Maringá, PR, 2019.
63 f.: il. color.

Orientador: Prof. Dr. Rodrigo Martins.
Dissertação (Mestrado Profissional) - Universidade Estadual de Maringá, Centro de Ciências Exatas, Departamento de Matemática, Programa de Pós-Graduação em Matemática (PROFMAT) - Mestrado Profissional, 2019.

1. Estatística - Técnicas de agrupamento. 2. Estatística - Medidas de dissimilaridade. 3. Educação básica - Proposta de atividades. 4. Método hierárquico (Matemática). 5. Método não hierárquico (Matemática). I. Martins, Rodrigo, orient. II. Universidade Estadual de Maringá. Centro de Ciências Exatas. Departamento de Matemática. Programa de Pós-Graduação em Matemática (PROFMAT) - Mestrado Profissional. III. Título.

CDD 23.ed. 519.5

Márcia Regina Paiva de Brito - CRB-9/1267

TÉCNICAS DE AGRUPAMENTO: UMA PROPOSTA DE ATIVIDADES PARA O ENSINO MÉDIO

Dissertação apresentada ao Programa de Mestrado Profissional em Matemática em Rede Nacional do Departamento de Matemática, Centro de Ciências Exatas da Universidade Estadual de Maringá, como parte dos requisitos necessários para a obtenção do título de Mestre em Matemática tendo a Comissão Julgadora composta pelos membros:

COMISSÃO JULGADORA:

Prof. Dr. Rodrigo Martins
DMA/Universidade Estadual de Maringá (Orientador)

Profa. Ms. Aline Edlâine de Medeiros
DMA/Universidade Estadual de Maringá

Prof. Ms. Fabio Crivelli de Avila
Universidade de São Paulo – Avaré/SP

Aprovada em: 02 de agosto de 2019.
Local de defesa: Auditório do DMA, Bloco F67, campus da Universidade Estadual de Maringá.

Agradecimentos

A Deus, por me dar condições para que eu pudesse realizar esse sonho e por ter colocado pessoas no meu caminho que me ajudaram a prosseguir.

Ao meu orientador, Rodrigo Martins, pela paciência, ensinamentos e enorme contribuição na confecção deste trabalho.

Aos meus pais, Raimunda e Natalino, pelo estímulo e apoio ao longo de toda minha vida acadêmica e a toda família, em especial meu esposo, Dircinei, pela compreensão e pelo fundamental apoio.

Aos professores do PROFMAT pelos ensinamentos e apoio recebido.

Às amizades construídas no curso: Bruna, Elaine, Nelidy, Sandra e Simone pela ajuda nos estudos e pelo companheirismo ao longo da jornada.

Por fim, à todos que, direta ou indiretamente, contribuíram para a minha vida acadêmica e profissional, os meus sinceros agradecimentos.

"Sem sonhos, a vida não tem brilho.
Sem metas, os sonhos não tem alicerces.
Sem prioridades, os sonhos não se tornam reais.
Sonhe, trace metas, estabeleça prioridades
e corra riscos para executar seus sonhos."
(Augusto Cury)

Resumo

Em qualquer decisão que tomamos em nosso cotidiano, sempre levamos em consideração uma grande quantidade de variáveis. Para tomarmos tais vereditos, nos baseamos geralmente na intuição, não identificamos quais as variáveis que afetaram a nossa decisão. Podemos associar esse fato ao treinamento que temos de analisar apenas uma variável. Com intuito de melhorar este aspecto, visamos apresentar neste trabalho, técnicas de análise de agrupamentos (cluster analysis), também conhecidas como técnicas de agrupamento que fazem parte da Análise Multivariada de dados, ramo da estatística que analisa dados com mais de uma variável. De modo geral, as técnicas de agrupamento consistem em separar objetos em grupos, baseando-se nas características que estes objetos possuem em comum, além de que essas técnicas podem ser utilizadas para classificar e analisar dados estatísticos que apresentam mais de uma variável. A justificativa para a escolha desse tema está na grande utilização destas técnicas em pesquisas de diversas áreas do conhecimento, embora, nas ementas do ensino de estatística no ensino básico, seja contemplado apenas o estudo de dados com uma variável. Desse modo, apresentamos neste trabalho as principais técnicas de agrupamento, bem como sua utilização e, em seguida, apresentamos algumas atividades que podem ser aplicadas no ensino básico.

Palavras-chave: Técnicas de agrupamento. Método hierárquico. Método não hierárquico. Medidas de dissimilaridade.

Abstract

In any decision we take in our daily life, we always consider a great number of variables ou a lot of variables. In order to make a decision, we usually do it by intuition, and it is not identified which variables affected our decision. We can associate this fact to the training that we have to analyze just one variable. With the motif of improvin this aspect, our aim in this essay was present analysis techniques of grouping (cluster analysis), also known as grouping techniques that belong to the Multivariate Data Analysis, a Statistic's workstream which studies data with more than one variable. In a general vision, the Grouping Techniques consist in split objects in groups, basing in the characteristics that the objects have in common; these techniques can be used to classify and analyze statistics data which present more than one variable. The defense to this theme choice is the large use of these methods in researches of many knowledge areas, although the syllabus of statistics in elementary teaching contemplates just the data study with one variable. By these means, we present, in this essay, the main grouping techniques, as well as its application and, forthwith, some activities that can be executed in the elementary teaching.

Keywords: Cluster analysis. Hierarchical method. Non-hierarchical method. Dissimilarity measurements.

Lista de ilustrações

Lista de tabelas

Sumário

Introdução

A estatística está inserida em várias áreas do conhecimento, por isso é necessária a sua aplicação, o seu entendimento e a sua interpretação como ferramenta de pesquisa. Agrupar dados semelhantes é uma das características intrínsecas do ser humano, podemos observar classificações em nosso cotidiano, por exemplo, em supermercados, em lojas, em revistas e jornais, em campeonatos dos mais diversos jogos. Por esse motivo, o ensino de classificação e agrupamentos já começa nas séries iniciais, como podemos evidenciar no trecho da Base Nacional Comum Curricular do ensino fundamental: "organização e comparação, em contexto local ou global, é importante para a melhor compreensão de si, do outro, da escola, da comunidade, do Estado, do país e do mundo."(BNCC 2017, p.352). Genericamente, agrupamentos e classificações são de extrema importância em nosso cotidiano, por esse pretexto eles também são trabalhados, de certo modo, até no ensino infantil. Além disso, agrupamentos e classificações apresentam um papel fundamental em diversas áreas da ciência, podemos citar como exemplo a classificação na biologia, conhecida como Taxionomia, e, na química, a classificação dos elementos da tabela periódica.

Outro documento norteador da educação, os Parâmetros Nacionais Curriculares, também evidencia a importância do ensino de estatística: “Com relação à estatística, a finalidade é fazer com que o aluno venha a construir procedimentos para coletar, organizar, comunicar e interpretar dados, utilizando tabelas, gráficos e representações que aparecem frequentemente em seu dia-a-dia. (PCN, 1998, p. 40)”.

Segundo Vincini (2005), os métodos estatísticos, para analisar variáveis, estão dispostos em dois grupos: um que trata da estatística, que olha as variáveis de maneira isolada – a estatística univariada – e outro que olha as variáveis de forma conjunta – a estatística multivariada. Quando um fenômeno depende de muitas variáveis, a estatística univariada falha, pois não basta conhecer informações estatísticas isoladas, mas é necessário, também, conhecer a totalidade dessas informações fornecidas pelo conjunto das variáveis e suas relações.

De acordo com Quintal (2006), entre os objetivos dessas técnicas estão a análise da

estrutura dos dados, a relação dessa estrutura e, posteriormente, uma classificação. Assim a análise de agrupamento é um procedimento da estatística multivariada que visa agrupar um conjunto de dados em subgrupos homogêneos interiormente e heterogêneos em relação a outros grupos.

As técnicas de agrupamento incluem uma série de procedimentos estatísticos que detalharemos a seguir, estas fazem um elo entre a estatística e as áreas da matemática como a álgebra e a geometria. Além disso, a matemática envolvida nas técnicas de agrupamento não demanda de muitos requisitos prévios, o que torna possível sua apresentação para alunos do ensino básico com uma quantidade propícia de dados. Nesse sentido, o objetivo desse trabalho é apresentar algumas técnicas de agrupamento e, em seguida, apresentar uma sequência de atividades direcionadas a alunos da educação básica que proporcionem, de forma simplificada, o entendimento das técnicas. Vale salientar que o intuito é de propiciar aos alunos o entendimento de que a análise estatística não necessariamente precisa analisar as variáveis de forma separada (univariada), podemos analisar de forma múltipla, com a estatística multivariada, utilizando, por exemplo, as técnicas de agrupamento.

A organização deste trabalho foi realizada da seguinte forma: no Capítulo 1, apresentaremos o conceito das técnicas de agrupamento e a sua utilização em pesquisas científicas, fazendo um breve relato de alguns trabalhos; no Capítulo 2, serão apresentadas as medidas de dissimilaridade mais utilizadas nas técnicas de agrupamento; nos Capítulos 3 e 4, apresentaremos as técnicas de agrupamento Hierárquico e Não Hierárquico; no Capítulo 5, apresentaremos propostas de atividades para propiciar o entendimento das técnicas aos alunos do ensino básico; no Apêndice A, por fim, apresentamos um breve resumo da utilização dos software R para realizar os agrupamentos.

Capítulo 1

Análise de Agrupamentos

A estatística mostra-se como uma poderosa ferramenta para a análise e a avaliação de dados nas mais diversas áreas de conhecimento. Segundo Vicini (2005), embora a estatística multivariada tenha surgido por volta de 1901, foi atualmente, com o avanço dos computadores, que foi possível desenvolver e aprimorar essas técnicas. Ainda de acordo com o mesmo autor, as técnicas de análise de agrupamentos ou técnicas de agrupamento são técnicas matemáticas, com grande fundamentação na álgebra e na geometria.

De modo geral, as técnicas de agrupamento permitem agrupar objetos em grupos com base em suas próprias características, buscando uma classificação natural entre estes objetos. Segundo Figueiredo Filho (2012), uma forma bastante intuitiva de compreender a lógica dessas técnicas e visualizar os agrupamentos é na organização de um supermercado. Em geral, itens semelhantes são agrupados em um mesmo setor: cerveja, vinho e refrigerantes se agrupam no setor de bebidas, enquanto que banana, maçã e laranja se agrupam no setor de hortifrutigranjeiro.

Segundo Favero et. al. (2009), essas técnicas podem ser utilizadas em todas as áreas do conhecimento humano cujo objetivo seja segmentar as observações em grupos homogêneos internamente e heterogêneos entre si. De acordo com Favero et. al. (2009) e Quintal (2006), as técnicas de agrupamento podem ser utilizadas em diversas situações de pesquisa, como, por exemplo:

- Um pesquisador que visa separar empresas com base em seus indicadores financeiros (rentabilidade, margem, custos, entre outros);
- Uma seguradora que visa identificar grupos de segurados de menor risco;
- Um educador que visa identificar grupos de alunos mais propensos à evasão escolar;
- Na área médica, agrupamento de doenças por sintomas ou curas pode levar a taxonomias muito úteis;

- Em marketing, a análise de agrupamento pode ser utilizada para encontrar grupos mais propensos a consumir determinado produto;
- Na biologia e na química, as técnicas podem contribuir para uma definição de classificação tal como a taxonomia relativa a minerais, insetos, plantas etc.

De modo geral, toda vez que se faz necessário que se classifique uma quantidade de dados em grupos de acordo com suas semelhanças, se utiliza métodos de agrupamento.

Uma característica importante presente nas diversas técnicas de agrupamento é o tratamento dos dados estatísticos com diversas variáveis como sendo pontos do Espaço $\mathbb{R}^n$ e, em seguida, o cálculo da distância escolhida.

Segundo Vincini (2005), em um conjunto de dados, é muito importante a escolha de um coeficiente que mostre a semelhança entre os objetos. Esse coeficiente pode ser dividido em duas categorias, as quais dizem respeito à estimação de uma medida de similaridade (semelhança) ou dissimilaridade (distância) entre os indivíduos, ou populações, a serem agrupados. Na medida de similaridade, quanto maior for o valor observado, mais parecidos serão os objetos. Já na medida de dissimilaridade, quanto maior for o valor observado, menos parecidos serão os objetos. Um exemplo de medida de similaridade é o coeficiente de correlação, pois quanto maior seu valor, maior a associação; da medida de dissimilaridade, a distância euclidiana, pois quanto menor o valor, mais próximo os objetos estão uns dos outros.

As medidas de similaridade e de dissimilaridade são inter-relacionadas, sendo facilmente conversíveis de uma para a outra. Existe um grande número de coeficientes de similaridade e dissimilaridade disponíveis na literatura, tais coeficientes de similaridade podem ser facilmente convertidos em coeficientes de dissimilaridade. Se denominarmos a medida de similaridade por s, a medida de dissimilaridade será o seu complementar (1-s).

De acordo com Favero et. al (2009), a aplicação das técnicas de agrupamento podem ser divididas nas seguintes etapas:

- Análise das variáveis e dos objetos a serem agrupados.
- Seleção da medida de distância (dissimilaridade) ou semelhança (similaridade) entre cada par de objetos.
- Escolha da quantidade de agrupamentos formados.
- Interpretação e validação dos agrupamentos.

A seguir descrevemos brevemente cada uma das etapas.

Análise das Variáveis: a seleção das variáveis deve ser feita com cautela, pois os grupos a serem formados refletirão a estrutura das variáveis e as técnicas de agrupamento

não diferenciam as variáveis relevantes para o estudo, cabe ao pesquisador identificá-las de acordo com o problema a ser respondido. Outro fato para o qual devemos nos atentar é a padronização dos dados, pois como é utilizado na maioria das técnicas o cálculo da distância, a variável que apresenta maior dispersão terá um peso elevado no cálculo da distância escolhida, assim o ideal seria utilizar a padronização dos dados.

Uma das formas de realizar a padronização dos dados é transformar cada variável em *escore* padrão (Z *escores*) permitindo que seja eliminada a diferença entre as escalas que pode prejudicar a análise dos grupos. O método Z *escores* padroniza cada variável (x) de maneira a apresentar média 0 e desvio padrão amostral 1, sendo calculado da seguinte maneira:

$$Z = \frac{\text{(x-média)}}{\text{desvio-padrão}}.$$

Seleção da medida de distância: após a análise das variáveis e a verificação da necessidade ou não da padronização das variáveis, é a hora de escolher a medida mais adequada aos seus dados. Vale lembrar que as medidas mais utilizadas são as de distância, como a Euclidiana, que detalhamos no próximo capítulo.

Escolha da quantidade de agrupamentos formados: conforme Favero et. al. (2009), a formação dos grupos decorre da distância escolhida e do método de agrupamento escolhido. As técnicas de agrupamento são divididas em dois tipos, a saber, os hierárquicos e os não hierárquicos. Nos algoritmos hierárquicos, a quantidade de grupos pode ser escolhida após o algoritmo, enquanto que nos não hierárquicos a quantidade de grupos deve ser escolhida previamente. Nos capítulos 3 e 4 vamos discutir esses dois tipos de agrupamentos apontando as vantagens e desvantagens de cada um.

Interpretação e validação dos agrupamentos: após a formação dos grupos de acordo com o método escolhido, o pesquisador deve realizar a interpretação dos resultados, analisando as características em comum em cada grupo, buscando a solução de seu problema inicial de pesquisa.

Podemos dizer que, nas mais diversas áreas de pesquisa, são mensurados inúmeras variáveis. Atualmente, temos diversas pesquisas que utilizam as técnicas de agrupamento para analisar essas variáveis em conjunto, descreveremos brevemente o estudo de algumas delas para exemplificar a importância destas técnicas.

Dentre os trabalhos analisados no decorrer deste estudo, podemos citar o de título "Análise de clusters aplicada ao sucesso/insucesso em matemática" (QUINTAL, 2006). Nele, a autora utiliza as técnicas de agrupamento para analisar os fatores que levam os alunos a terem sucesso ou insucesso na disciplina de matemática, utilizando para isto uma amostra de alunos do nono ano de escolas públicas e particulares, na cidade de Funchal, em Portugal, concluindo, por meio dessa análise, que as causas principais do insucesso escolar não estão apenas concentradas nos estudantes, mas estão principalmente associadas

a outros fatores tais como habilitações dos pais e critérios de avaliação pouco exigentes baseados nas leis em vigor.

No trabalho "Uma análise multivariada do sucesso ou fracasso em matemática dos alunos do 8º Ano do ensino fundamental" (SILVA JÚNIOR, 2014), foi realizado um estudo aplicando a análise de agrupamento hierárquico, para avaliar as variáveis que influenciam no desempenho em matemática de alunos do 9º ano do ensino fundamental II, em quatro escolas da rede pública de ensino localizadas nos estados de Pernambuco e da Bahia. O autor concluiu, em sua pesquisa, que o acesso dos alunos às tecnologias mostrou uma relação positiva com o desempenho na disciplina de matemática, o que evidencia a necessidade de aproximação do ensino de matemática com as tecnologias educacionais; outros fatores que influenciaram positivamente foi a parceria entre a família, a escola e a presença de assistente social na escola.

Outro trabalho analisado foi o intitulado "Aplicações de técnicas de análise multivariada em experimentos agropecuários utilizando o software R" (SARTORIO, 2008), no qual a autora afirma que a disseminação das técnicas da análise multivariada pode melhorar a qualidade das pesquisas, proporcionando uma relativa economia de tempo e custo e diminuindo também a perda de informação. Sartorio apresentou, nesse trabalho, diversos experimentos utilizando a análise multivariada e sobre a análise de agrupamentos, afirma que podem auxiliar o pesquisador na construção de grupos de animais baseando-se em informações de mais de uma característica e na decisão pela melhor solução, recomenda que o pesquisador avalie as variâncias internas dos grupos obtidos. Segundo a autora, o software R mostrou-se como uma ferramenta poderosa para a aplicação das técnicas de análise multivariada.

No capítulo a seguir, apresentamos as medidas de dissimilaridade, que são as mais utilizadas nas técnicas de agrupamento, as suas propriedades e algumas distâncias ou métricas que podemos utilizar com esta finalidade.

Capítulo 2

Medidas de Dissimilaridade

As medidas de dissimilaridade são distâncias (também chamadas de métricas). Para realizar os cálculos, vamos considerar os objetos como pontos no Espaço $\mathbb{R}^n$ (cada característica observada é considerada uma variável), onde os pontos são as listas $x = (x_1, \cdots, x_n)$. De acordo com Lima (1977), uma métrica em um conjunto M é uma função $d : M \times M \to \mathbb{R}$, que associa cada par ordenado de elementos $x, y \in M$ a um número real $d(x, y)$, de modo que sejam satisfeitas as seguintes condições para quaisquer $x, y, z \in M$:

$d_1)$ $d(x, x) = 0$;

$d_2)$ $d(x, y) > 0$, se $x \neq y$;

$d_3)$ $d(x, y) = d(y, x)$;

$d_4)$ $d(x, z) \leq d(x, y) + d(y, z)$.

Analisando as quatro condições acima, observamos que (d_1) e (d_2) nos dizem que a distância entre dois pontos x e y será igual a zero se, e somente se, $x = y$; a condição (d_4) é conhecida como desigualdade triangular e tem origem no plano euclidiano, geometricamente representa que o comprimento de um dos lados de um triângulo não excede a soma dos outros dois, caso temos a igualdade, os pontos x, y e z não formam um triângulo.

Exemplo 2.1. *Como exemplo, vamos citar a métrica do "zero-um", onde qualquer conjunto M pode tornar-se um espaço métrico de maneira muito simples. Para isto, definimos a métrica* $d : M \times M \to \mathbb{R}$ *pondo* $d(x, x) = 0$ *e* $d(x, y) = 1$*, se* $x \neq y$*. Note que todas as condições são facilmente verificadas.*

Elencamos a seguir algumas das distâncias que são utilizadas nas técnicas de agrupamentos. Para simplificar a notação, vamos realizar a demonstração das propriedades de métrica no $\mathbb{R}^2$, a demonstração para o caso no $\mathbb{R}^n$ segue de modo análogo.

1) **Distância Euclidiana:** provém da fórmula para distância entre dois pontos do plano em coordenadas cartesianas, a qual se prova pelo Teorema de Pitágoras. É considerada uma distância natural, pois fornece a distância da Geometria Euclidiana, ou seja, a menor distância entre dois pontos é uma reta. Sendo $x = (x_1, ..., x_n)$ e $y = (y_1, ..., y_n)$, essa distância é dada por:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2} = [\sum_{i=1}^{n}(x_n - y_n)^2]^{\frac{1}{2}}.$$

Demonstração: considere $x = (x_1, x_2), y = (y_1, y_2)$ e $z = (z_1, z_2)$ pontos do $\mathbb{R}^2$ diferentes da origem.

d_1) Queremos mostrar que $d(x, x) = 0$.

De fato,

$$d(x, x) = \sqrt{(x_1 - x_1)^2 + (x_2 - x_2)^2} = 0.$$

d_2) Queremos mostrar que $d(x, y) > 0$ com $x \neq y$.

De fato,

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2},$$

Note que, como $x \neq y$, temos que $(x_1 - y_1)^2 > 0$ assim como $(x_2 - y_2)^2 > 0$, logo

$$(x_1 - y_1)^2 + (x_2 - y_2)^2 > 0,$$

e consequentemente

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} > 0.$$

d_3) Queremos mostrar que $d(x, y) = d(y, x)$.

De fato,

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

e

$$d(y, x) = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2}.$$

Note que

$$(x_1 - y_1)^2 = (y_1 - x_1)^2$$

e

$$(x_2 - y_2)^2 = (y_2 - x_2)^2.$$

já que estão elevados ao quadrado. Logo,

$$(x_1 - y_1)^2 + (x_2 - y_2)^2 = (y_1 - x_1)^2 + (y_2 - x_2)^2,$$

$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} = \sqrt{(y_1 - x_1)^2 + (y_2 - x_2)^2}.$$

$d_4)$ $d(x, z) \leq d(x, y) + d(y, z)$.

A desigualdade acima pode ser ilustrada na geometria euclidiana pelo triângulo abaixo:

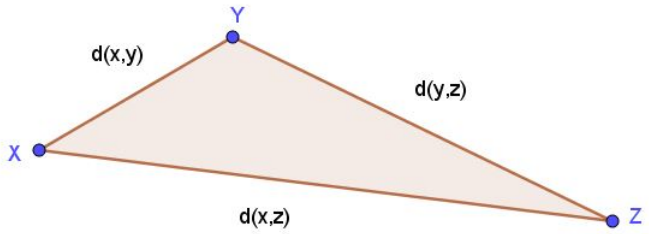


Figura 1 – Representação da desigualdade triangular na distância Euclidiana

Intuitivamente, percebemos que $d(x, z) = d(x, y) + d(y, z)$ se y está na mesma semirreta e entre x e z e $d(x, z) < d(x, y) + d(y, z)$ se y estiver fora desta semirreta.

Antes de demonstrarmos essa propriedade, mostraremos a **Desigualdade de Cauchy-Schwartz** no $\mathbb{R}^2$, cujo enunciado é o seguinte:

Se a_1, a_2 e b_1, b_2 são números reais arbitrários, então

$$(a_1^2 + a_2^2).(b_1^2 + b_2^2) \leq (a_1 b_1 + a_2 b_2)^2.$$

Demonstração: Considere a seguinte função :

$$f(x) = (a_1 - b_1 x)^2 + (a_2 - b_2 x)^2.$$

Note que $f(x)$ é uma função do segundo grau, para $b_1 \neq 0$ e $b_2 \neq 0$, com valores maiores ou iguais a zero, logo $\Delta \leq 0$. Desenvolvendo $f(x)$ obtemos:

$$f(x) = a_1^2 - 2a_1 b_1 x + b_1^2 x^2 + a_2^2 - 2a_2 b_2 x + b_2^2 x^2,$$

$$f(x) = (b_1^2 + b_2^2)x^2 - 2(a_1 b_1 + a_2 b_2)x + (a_1^2 + a_2^2).$$

Assim temos,

$$\Delta = (-2(a_1b_1 + a_2b_2))^2 - 4.(b_1^2 + b_2^2).(a_1^2 + a_2^2) \leq 0,$$
$$4(a_1b_1 + a_2b_2)^2 - 4.(b_1^2 + b_2^2).(a_1^2 + a_2^2) \leq 0,$$
$$\Rightarrow 4(a_1b_1 + a_2b_2)^2 \leq 4.(b_1^2 + b_2^2).(a_1^2 + a_2^2),$$
$$\therefore (a_1b_1 + a_2b_2)^2 \leq (b_1^2 + b_2^2).(a_1^2 + a_2^2).$$

ou equivalentemente:

$$(a_1b_1 + a_2b_2) \leq (b_1^2 + b_2^2)^{\frac{1}{2}}.(a_1^2 + a_2^2)^{\frac{1}{2}}.$$

Vamos agora demonstrar a desigualdade triangular para a distância euclidiana, queremos mostrar que $d(x, z) \leq d(x, y) + d(y, z)$.
De fato,

$$d(x, z) = \sqrt{(x_1 - z_1)^2 + (x_2 - z_2)^2},$$
$$d(x, z)^2 = (x_1 - z_1)^2 + (x_2 - z_2)^2,$$
$$d(x, z)^2 = (x_1 - y_1 + y_1 - z_1)^2 + (x_2 - y_2 + y_2 - z_2)^2,$$
$$d(x, z)^2 = (x_1-y_1)^2+2(x_1-y_1)(y_1-z_1)+(y_1-z_1)^2+(x_2-y_2)^2+2((x_2-y_2)(y_2-z_2)+(y_2-z_2)^2.$$
$$d(x, z)^2 = (x_1-y_1)^2+(x_2-y_2)^2+2[(x_1-y_1)(y_1-z_1)+(x_2-y_2)(y_2-z_2)]+(y_1-z_1)^2+(y_2-z_2)^2.$$

Aplicando a desigualdade de Cauchy-Schwartz, temos

$$d(x, z)^2 \leq (x_1-y_1)^2+(x_2-y_2)^2+2[(x_1-y_1)^2+(x_2-y_2)^2]^{\frac{1}{2}}[(y_1-z_1)^2+(y_2-z_2)^2]^{\frac{1}{2}}+(y_1-z_1)^2$$
$$+(y_2 - z_2)^2$$
$$= [\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} + \sqrt{(y_1 - z_1)^2 + (y_2 - z_2)^2}]^2$$
$$= [d(x, y) + d(y, z)]^2.$$

Assim, $d(x, z) \leq d(x, y) + d(y, z)$ e, portanto, d é uma métrica em $\mathbb{R}^2$.

2) **Distância do Taxista:** provém da associação com a ideia de "trafegar por ruas". A distância entre dois pontos no plano cartesiano é calculada assumindo-se que só se possa fazer trajetos horizontais e verticais. Sendo $x = (x_1, ..., x_n)$ e $y = (y_1, ..., y_n)$, essa distância é dada por:

$$d'(x, y) = |x_1 - y_1| + \cdots + |x_n - y_n| = \sum_{i=1}^{n} |x_n - y_n|.$$

Demonstração: considere $x = (x_1, x_2)$, $y = (y_1, y_2)$ e $z = (z_1, z_2)$ pontos do $\mathbb{R}^2$ diferentes da origem.

d_1) Queremos mostrar que $d'(x, x) = 0$.

De fato,

$$d'(x,x) = |x_1 - x_1| + |x_2 - x_2| = 0.$$

d_2) Queremos mostrar que $d'(x,y) > 0$ com $x \neq y$.

De fato,

$$d'(x,y) = |x_1 - y_1| + |x_2 - y_2|.$$

Como $x \neq y$ e ambos são diferentes da origem, temos que $|x_1 - y_1| > 0$ ou $|x_2 - y_2| > 0$, neste caso,

$$d'(x,y) = |x_1 - y_1| + |x_2 - y_2| > 0.$$

d_3) Queremos mostrar que $d'(x,y) = d'(y,x)$.

Temos que:

$$d'(x,y) = |x_1 - y_1| + |x_2 - y_2|$$

e

$$d'(y,x) = |y_1 - x_1| + |y_2 - x_2|$$

Observe que $|x_1 - y_1| = |y_1 - x_1|$, pois estão em módulo, assim como $|x_2 - y_2| = |y_2 - x_2|$, logo

$$d'(x,y) = |x_1 - y_1| + |x_2 - y_2| = |y_1 - x_1| + |y_2 - x_2| = d'(y,x).$$

d_4) Queremos mostrar que $d'(x,z) \leq d'(x,y) + d'(y,z)$.

Temos que:

$$\begin{aligned} d(x,z) &= |x_1 - z_1| + |x_2 - z_2| \\ &= |x_1 - y_1 + y_1 - z_1| + |x_2 - y_2 + y_2 - z_2| \\ &\leq |x_1 - y_1| + |y_1 - z_1| + |x_2 - y_2| + |y_2 - z_2| \\ &\leq |x_1 - y_1| + |x_2 - y_2| + |y_1 - z_1| + |y_2 - z_2| \\ &= d'(x,y) + d'(y,z). \end{aligned}$$

Assim $d'(x,z) \leq d'(x,y) + d'(y,z)$ para todo x, y e $z \in \mathbb{R}^2$ e, portanto, d' é uma métrica em $\mathbb{R}^2$.

3) **Distância do Máximo:** considera a distância máxima entre as coordenadas. Sendo $x = (x_1, ..., x_n)$ e $y = (y_1, ..., y_n)$, essa distância é dada por:

$$d''(x,y) = max\{|x_1 - y_1|, \cdots, |x_n - y_n|\} = max\{|x_i - y_i|\}, 1 \leq i \leq n.$$

Demonstração: considere $x = (x_1, x_2), y = (y_1, y_2)$ e $z = (z_1, z_2)$ pontos do $\mathbb{R}^2$ diferentes da origem.

d_1) Queremos mostrar que $d"(x, x) = 0$.

De fato,

$$d''(x, x) = max\{|x_1 - x_1|, |x_2 - x_2|\} = 0.$$

d_2) Queremos mostrar que $d"(x, y) > 0$ com $x \neq y$.

Basta notar que $|x_1 - y_1| > 0$ ou $|x_2 - y_2| > 0$, pois estão em módulo, logo

$$d''(x, y) = max\{|x_1 - y_1|, |x_2 - y_2|\} > 0.$$

d_3) Queremos mostrar que $d"(x, y) = d"(y, x)$.

Temos que:

$$d''(x, y) = max\{|x_1 - y_1|, |x_2 - y_2|\}$$

e

$$d''(y, x) = max\{|y_1 - x_1|, |y_2 - x_2|\}.$$

Observe que $|x_1 - y_1| = |y_1 - x_1|$ e $|x_2 - y_2| = |y_2 - x_2|$, assim

$$d''(x, y) = d''(y, x).$$

d_4) Queremos mostrar que $d"(x, z) \leq d"(x, y) + d"(y, z)$.

Temos que:

$$\begin{aligned} d''(x, z) &= max\{|x_1 - z_1|, |x_2 - z_2|\} \\ &= max\{|x_1 - y_1 + y_1 - z_1|, |x_2 - y_2 + y_2 - z_2|\} \\ &\leq max\{|x_1 - y_1| + |y_1 - z_1|, |x_2 - y_2| + |y_2 - z_2|\} \\ &\leq max\{|x_1 - y_1|, |x_2 - y_2|\} + max\{|y_1 - z_1|, |y_2 - z_2|\} \\ &= max\{|x_1 - y_1|, |x_2 - y_2|\} + max\{|y_1 - z_1|, |y_2 - z_2|\} \\ &= d"(x, y) + d"(y, z). \end{aligned}$$

Assim $d"(x, z) \leq d"(x, y) + d"(y, z)$ para todo x, y e $z \in \mathbb{R}^2$ e, portanto, $d"$ é uma métrica em $\mathbb{R}^2$.

4) **Distância euclidiana quadrática:** é dada pelo quadrado da distância Euclidiana, ou seja,

$$d(x, y)^2 = (x_1 - y_1)^2 + \cdots + (x_n - y_n)^2 = \sum_{i=1}^{n} (x_n - y_n)^2.$$

Demonstração: é análoga a demonstração da métrica euclidiana realizada em (1).

Exemplo 2.2. *Fazendo uma analogia da diferença entre as distâncias Euclidiana e do Taxista, a distância Euclidiana seria o segmento de uma reta que liga os pontos A e B e a distância do Taxista seria um segmento de retas tanto na vertical quanto na horizontal, semelhante a uma rota de carro pelos quarteirões de uma cidade. Na imagem abaixo AB representa a distância Euclidiana, enquanto que AC + CB representa a distância do Taxista.*

Figura 2 – Representação da distância Euclidiana com a distância do Taxista.

Um conceito fundamental no estudo de distâncias é a noção de bola, para isso, vamos utilizar a definição presente em Lima (1977).

Seja a um ponto do espaço métrico M, dado um número real $r > 0$ definimos:

A **bola aberta** de centro a e raio r como o conjunto $B(a; r)$, formado pelos pontos de M que estão a uma distância menor a r no ponto a, ou seja,

$$B(a; r) = \{x \in M; d(x, a) < r\}.$$

A **bola fechada** de centro a e raio r como o conjunto $B[a; r]$, formado pelos pontos de M que estão a uma distância menor ou igual a r no ponto a, ou seja,

$$B[a; r] = \{x \in M; d(x, a) \leq r\}.$$

A **esfera** de centro a e raio r como o conjunto $S(a; r)$, formado pelos pontos pontos $x \in M$ tais que $d(x, a) = r$, ou seja,

$$S(a; r) = \{x \in M; d(x; a) = r\}.$$

Evidentemente que $B[a; r] = B(a; r) \cup S(a; r)$.

Para exemplificar essas definições, na distância Euclidiana, o conceito de Bola fechada é o que costumamos utilizar, isto é, seja $a = (a_1, a_2)$ um ponto do $\mathbb{R}^2$ e um raio r temos que a bola fechada com a distância euclidiana é dada por

$$(x - a_1)^2 + (y - a_2)^2 \leq r.$$

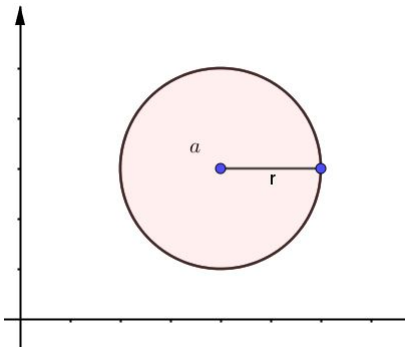


Figura 3 – Representação da Bola Fechada utilizando a distância Euclidiana.

Já na distância do Taxista, o conceito de Bola fechada é diferente do usual, pois, seja $a = (a_1, a_2)$ um ponto do $\mathbb{R}^2$ e um raio r, temos que a bola fechada com a distância do taxista é dada por

$$|x - a_1| + |y - a_2| \leq r.$$

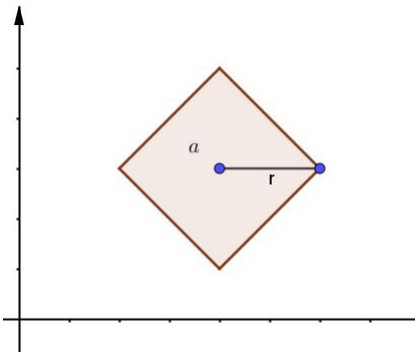


Figura 4 – Representação da Bola Fechada utilizando a distância do Taxista.

Na distância do Máximo, o conceito de Bola fechada também é diferente do usual, pois, para $a = (a_1, a_2)$ um ponto do $\mathbb{R}^2$ e um raio r, a bola fechada com a distância do máximo é dada por

$$max|x - a_1|, |y - a_2| \leq r$$

e assim $|x - a_1| \leq r$ e $|y - a_2| \leq r$.

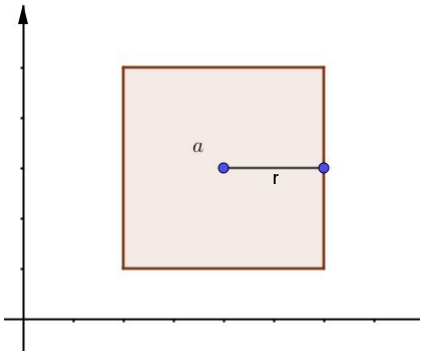


Figura 5 – Representação da Bola Fechada utilizando a distância do Máximo.

As três métricas mencionadas anteriormente (d, d' e d'') estão relacionadas, conforme veremos na proposição a seguir.

Proposição 1: Sejam d, d' e d'' as métricas definidas anteriormente. Quaisquer que sejam $x, y \in \mathbb{R}^n$, para $n \geq 2$ e $n \in \mathbb{N}$ tem-se:

$$d''(x, y) \leq d(x, y) \leq d'(x, y) \leq n.d''(x, y).$$

Demonstração: para demonstrar esta proposição vamos analisar separadamente cada uma das desigualdades. Sejam $x, y \in \mathbb{R}^2$, com $x = (x_1, x_2)$ e $y = (y_1, y_2)$.

1) Queremos provar que $d''(x, y) \leq d(x, y)$, ou seja,

$$max\{|x_1 - y_1|, |x_2 - y_2|\} \leq \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}.$$

Elevando o primeiro membro da desigualdade ao quadrado e supondo

$$|x_1 - y_1|$$

seja o máximo obtemos

$$|x_1 - y_1|^2.$$

Note que, elevando o segundo membro da desigualdade, obtemos

$$(x_1 - y_1)^2 + (x_2 - y_2)^2,$$

assim

$$|x_1 - y_1|^2 \leq (x_1 - y_1)^2 + (x_2 - y_2)^2,$$

já que $(x_2 - y_2)^2$ é positivo.

Logo,

$$max\{|x_1 - y_1|, |x_2 - y_2|\} \leq \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}.$$

2) Queremos provar que $d(x, y) \leq d'(x, y)$, ou seja,

$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \leq |x_1 - y_1| + |x_2 - y_2|.$$

Pelo item (1), temos que o primeiro membro da desigualdade elevado ao quadrado é igual a

$$(x_1 - y_1)^2 + (x_2 - y_2)^2.(*)$$

Note que, elevando o segundo membro da desigualdade, obtemos:

$$[|x_1 - y_1| + |x_2 - y_2|]^2 = |x_1 - y_1|^2 + 2.|x_1 - y_1|.|x_2 - y_2| + |x_2 - y_2|^2.(**)$$

De (*) e (**) segue que

$$(x_1 - y_1)^2 + (x_2 - y_2)^2 \leq |x_1 - y_1|^2 + 2.|x_1 - y_1|.|x_2 - y_2| + |x_2 - y_2|^2,$$

pois

$$2.|x_1 - y_1|.|x_2 - y_2|$$

é sempre positivo.

Logo,

$$\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \leq |x_1 - y_1| + |x_2 - y_2|.$$

3) Queremos mostrar que $d'(x, y) \leq n.d''(x, y)$, ou seja,

$$|x_1 - y_1| + |x_2 - y_2| \leq n.max\{|x_1 - y_1|, |x_2 - y_2|\}.$$

De fato, como a segunda desigualdade apresenta o máximo de

$$|x_1 - y_1|, |x_2 - y_2|$$

e $n = 2$, logo,

$$|x_1 - y_1| + |x_2 - y_2| \leq n.max\{|x_1 - y_1|, |x_2 - y_2|\}.$$

Pelos itens (1), (2) e (3) obtemos o desejado

$$d''(x, y) \leq d(x, y) \leq d'(x, y) \leq n.d''(x, y).$$

Seja M um espaço métrico, um ponto $a \in M$ chama-se *ponto isolado* de M quando ele é uma bola aberta em M, ou seja, quando existe $r > 0$, tal que $B(a; r) = \{a\}$. Dizer que um ponto $a \in M$ não é isolado significa afirmar que para todo $r > 0$ pode se encontrar um ponto $x \in M$ tal que $0 < d(a, x) < r$.

Exemplo 2.3. *Seja* $\mathbb{Z} = \{..., -2, -1, 0, 1, 2, ...\}$ *o conjunto dos números inteiros com a métrica euclidiana. Temos que todo ponto de* $\mathbb{Z}$ *é isolado, pois, tomando* $r = 1$*, se* $x \in \mathbb{Z}$ *e tal que* $x \in B(n; 1)$*, então* $|x - n| < 1$ *e, portanto,* $x = n$*.*

Um espaço métrico chama-se *discreto* quando todo ponto de M é isolado. Por exemplo, um espaço com a métrica zero-um é discreto, no exemplo anterior $\mathbb{Z}$ também é um conjunto discreto.

Proposição 2: dados os pontos $a \neq b$ em um espaço métrico M, sejam $r > 0$ e $s > 0$, tais que $r + s \leq d(a, b)$. Então as bolas abertas $B(a; r)$ e $B(b; s)$ são disjuntas.

Demonstração: vamos supor que exista um ponto $x \in B(a; r) \cap B(b; s)$, teríamos $d(a; x) < r$ e $d(b; x) < s$. Assim

$$d(a, b) \leq d(a; x) + d(b; x) < r + s \leq d(a; b),$$

um absurdo.

Capítulo 3

Agrupamento Hierárquico

Agrupamento hierárquico se define por técnicas que agrupam os elementos de acordo com as suas semelhanças ademais de estabelecer uma relação entre os grupos. As semelhanças entre os dados são calculadas, por exemplo, através de uma distância escolhida. Segundo Moita (2004), a análise de agrupamento hierárquico consiste no tratamento matemático de cada amostra como um ponto no espaço multidimensional descrito pelas variáveis escolhidas. Quando uma determinada amostra é tomada como um ponto no espaço das variáveis, é possível calcular a distância desse ponto a todos os outros pontos, constituindo-se, assim, uma matriz que descreve a proximidade entre todas as amostras estudadas.

A principal vantagem dos algoritmos hierárquicos é o fato de eles oferecerem não só os grupos obtidos, mas também toda a estrutura dos dados e permitirem obter facilmente subconjuntos dentro desses dados. Segundo Hair et al (2009), no agrupamento hierárquico, os resultados de um estágio anterior são sempre aninhados com os resultados de um estágio posterior, apresentando semelhança como na estrutura de uma árvore.

Nesses algoritimos distinguem-se dois tipos de procedimentos: os métodos aglomerativos (mais utilizados) e os métodos divisivos.

No método aglomerativo, cada elemento da amostra começa com o seu próprio agrupamento, em seguida novos agrupamentos serão formados com base na dissimilaridade ou similaridade desses pontos. Na etapa seguinte, os dois indivíduos mais próximos (de acordo com a distância escolhida) serão agrupados e, posteriormente, irão se agrupar com os demais grupos de acordo com a proximidade, conforme no exemplo 3.1 a seguir. Já no método divisivo, todas as observações começam em um grande grupo, sendo separadas primeiramente as observações mais distantes, até que cada observação se torne um grupo.

Ambos os métodos utilizam como base a **Matriz de Distâncias** para realizar os agrupamentos, que é definida como:

Considere A, B, C e D uma amostra com n variáveis. A cada objeto associamos a um ponto no plano $\mathbb{R}^n$ da forma $X = (x_1, \ldots, x_n)$. A Matriz distância é dada por:

$$D' = \begin{bmatrix} d(A,A) & d(A,B) & d(A,C) & d(A,D) \\ d(B,A) & d(B,B) & d(B,C) & d(B,D) \\ d(C,A) & d(C,B) & d(C,C) & d(C,D) \\ d(D,A) & d(D,B) & d(D,C) & d(D,D) \end{bmatrix}.$$

Note que, pela definição de distância sobre uma métrica, essa matriz sempre será simétrica, o que facilita os cálculos; além disso, a diagonal principal é nula. Ademais, no exemplo acima, temos uma matriz 4×4, pois temos 4 objetos, para uma amostra de m objetos teremos uma matriz $m \times m$, assim, para uma quantidade grande de objetos na amostra o cálculo manual das distâncias é inviável, por isso as pesquisas geralmente utilizam softwares para realizar os agrupamentos, dentre eles podemos citar o software R.

Segundo Linden (2009), o algorítimo de agrupamento hierárquico aglomerativo é realizado da seguinte forma:

1) Inicie com N grupos, onde N é a quantidade de objetos, e construa a matriz de distâncias de ordem N.

2) Identifique o menor elemento da matriz de distâncias para encontrar o par de grupos mais similares.

3) Reúna os dois grupos identificados na etapa 2 em um único grupo (utilizando a Menor Distância, Distância Média ou Maior Distância, conforme veremos a seguir) e atualize a matriz de distâncias, retirando as linhas e colunas relativas aos dois grupos identificados em 2 e incluindo a linha e coluna com as distâncias entre os demais grupos e o novo grupo formado. Note que a ordem da matriz de distâncias diminui de uma unidade a cada vez que a etapa 3 é executada.

4) Repita os passos 2 e 3 até que reste apenas um grupo. A cada iteração, guarde a identificação dos grupos que foram fundidos e também a distância entre eles, estas informações serão utilizadas na montagem do dendrograma.

Existem diferentes maneiras de medir a distância entre dois agrupamentos, descrevemos algumas a seguir.

- **Menor distância (ligação simples):** a distância entre dois agrupamentos é dada pela distância entre seus pontos mais próximos. É um método que prioriza os elementos mais próximos, deixando os mais distantes em segundo plano.

  Após identificar na matriz de distâncias os objetos u e v que são mais próximos (chamamos de d_{uv}), agrupamos u e v em um único cluster que é denominado uv e a matriz de distâncias é atualizada da seguinte forma:

i) Eliminamos a linha e a coluna referentes aos objetos u e v;
ii) Calculamos as distâncias entre os demais objetos ao grupo uv, tal que

$$d_{(uv)w} = min\{d_{uw}, d_{vw}\}.$$

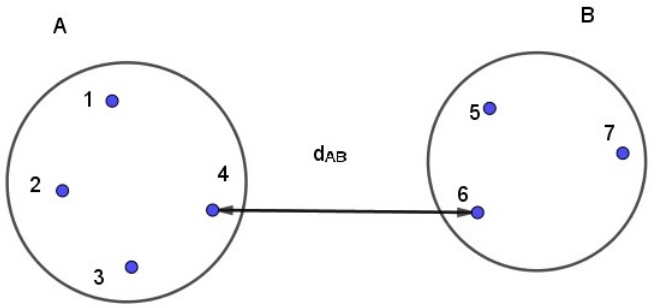


Figura 6 – Representação geométrica da menor distância.

- **Maior distância (ligação completa):** a distância entre os grupos é dada pela medida de seus elementos mais distantes. É semelhante ao anterior, com a diferença de que a matriz é atualizada com as distâncias entre os demais objetos ao grupo uv, tal que

$$d_{(uv)w} = max\{d_{uw}, d_{vw}\}.$$

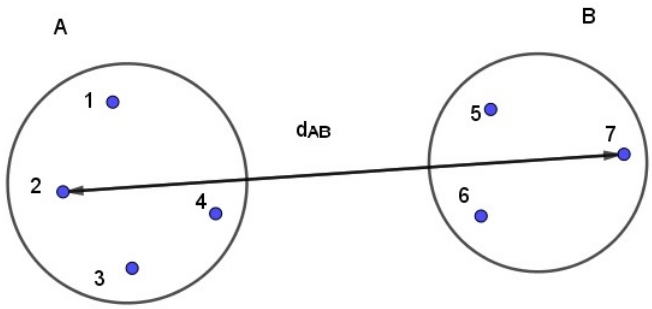


Figura 7 – Representação geométrica da maior distância.

- **Distância média (ligação média):** a distância entre dois grupos é a distância média entre todos os pares de indivíduos dos dois grupos, buscando reunir os grupos cuja distância média é menor. É semelhante as distâncias anteriores com a diferença de que, neste caso, as distâncias são atualizadas conforme a seguinte relação:

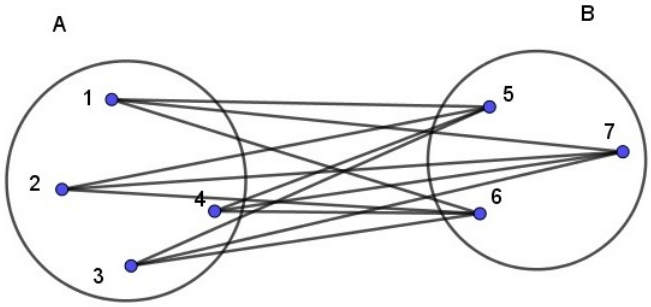
$$d_{(uv)w} = \frac{\sum_{i=1}^{i} \sum_{j=1}^{j} d_{ij}}{n_{(uv)} n_w}.$$

Onde:

d_{ij} = distância entre o objeto i, do cluster (uv) e o objeto j, do cluster w;

$n_{(uv)}$ = número de objetos do cluster (uv);

n_w = número de objetos do cluster w.



$$d_{AB} = \frac{1}{12}\left[(d_{15} + d_{16} + d_{17}) + (d_{25} + d_{26} + d_{27}) + (d_{35} + d_{36} + d_{37}) + (d_{45} + d_{46} + d_{47})\right]$$

Figura 8 – Representação geométrica da distância média.

De acordo com Linden (2009), essas três maneiras de medir a distância entre agrupamentos não são totalmente equivalentes. Escolhas distintas entre estas formas de medir distâncias podem gerar resultados diferentes no agrupamento.

No exemplo a seguir, vamos utilizar a distância euclidiana e a ligação entre os grupos da menor distância para realizar o agrupamento hierárquico.

Exemplo 3.1. *Considere o seguinte conjunto de pontos dados por:*

Variável	A	B	C	D	E
V_1	1	4	5	9	11

Tabela 1 – Dados - Exemplo 3.1.

Vamos calcular, primeiramente, a distância entre os pontos, utilizando a distância Euclidiana.

$$
\begin{aligned}
d(A,A) &= d(1,1) = \sqrt{(1-1)^2} = 0 = d(B,B) = ... = d(E,E);\\
d(A,B) &= d(1,4) = \sqrt{(4-1)^2} = 3 = d(B,A);\\
d(A,C) &= d(1,5) = \sqrt{(5-1)^2} = 4 = d(C,A);\\
d(A,D) &= d(1,9) = \sqrt{(9-1)^2} = 8 = d(D,A);\\
d(A,E) &= d(1,11) = \sqrt{(11-1)^2} = 10 = d(E,A);\\
d(B,C) &= d(4,5) = \sqrt{(5-4)^2} = 1 = d(C,B);\\
d(B,D) &= d(4,9) = \sqrt{(9-4)^2} = 5 = d(D,B);\\
d(B,E) &= d(4,11) = \sqrt{(11-4)^2} = 7 = d(E,B);\\
d(C,D) &= d(5,9) = \sqrt{(9-5)^2} = 4 = d(D,C);\\
d(C,E) &= d(5,11) = \sqrt{(11-5)^2} = 6 = d(E,C);\\
d(D,E) &= d(9,11) = \sqrt{(11-9)^2} = 2 = d(E,D).
\end{aligned}
$$

Assim obtemos a seguinte matriz de distâncias:

$$
D_1' = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array}
\begin{array}{c}
\begin{array}{ccccc} A & B & C & D & E \end{array} \\
\left[\begin{array}{ccccc}
0 & & & & \\
3 & 0 & & & \\
4 & 1 & 0 & & \\
8 & 5 & 4 & 0 & \\
10 & 7 & 6 & 2 & 0
\end{array}\right]
\end{array}.
$$

Como podemos ver, os elementos mais próximos, utilizando a distância euclidiana, são os elementos B e C. Realizamos, então, o agrupamento utilizando o método da "Menor Distância."

d_{BC} = 1, assim obtemos um novo grupo: (BC), excluímos as linhas e colunas correspondentes a B e C, comparamos as distâncias utilizando a menor distância e as novas distâncias são dadas por:

$$
\begin{aligned}
d_{(BC)A} &= min\{d_{BA}, d_{CA}\} = min\{3,4\} = 3;\\
d_{(BC)D} &= min\{d_{BD}, d_{CD}\} = min\{5,4\} = 4;\\
d_{(BC)E} &= min\{d_{BE}, d_{CE}\} = min\{7,6\} = 6.
\end{aligned}
$$

Assim obtemos a nova matriz:

$$D_2' = \begin{array}{c} \\ A \\ BC \\ D \\ E \end{array} \begin{array}{c} \begin{array}{cccc} A & BC & D & E \end{array} \\ \begin{bmatrix} 0 & & & \\ 3 & 0 & & \\ 8 & 4 & 0 & \\ 10 & 6 & 2 & 0 \end{bmatrix} \end{array}.$$

Temos agora que a menor distância é $d_{DE} = 2$, assim obtemos um novo grupo (DE).

Novas distâncias:

$$d_{(DE)A} = min\{d_{DA}, d_{EA}\} = min\{8, 10\} = 8;$$

$$d_{(DE)BC} = min\{d_{D(BC)}, d_{E(BC)}\} = min\{4, 6\} = 4.$$

Assim, obtemos a nova matriz:

$$D_3' = \begin{array}{c} \\ A \\ BC \\ DE \end{array} \begin{array}{c} \begin{array}{ccc} A & BC & DE \end{array} \\ \begin{bmatrix} 0 & & \\ 3 & 0 & \\ 8 & 4 & 0 \end{bmatrix} \end{array}.$$

A menor distância em D_3' é dada por $d_{A(BC)} = 3$, assim obtemos o grupo (ABC). Atualização das distâncias:

$$d_{(ABC)DE} = min\{d_{A(DE)}, d_{(BC)(DE)}\} = min\{8, 4\} = 4.$$

Assim obtemos a nova matriz:

$$D_4' = \begin{array}{c} \\ ABC \\ DE \end{array} \begin{array}{c} \begin{array}{cc} ABC & DE \end{array} \\ \begin{bmatrix} 0 & \\ 4 & 0 \end{bmatrix} \end{array}.$$

Assim, os grupos (ABC) e (DE), são agrupados em um único cluster e a distância mínima entre seus objetos é de 4 unidades.

Segundo Favero et. al. (2009), uma maneira de representar graficamente o processo de agrupamento hierárquico é por meio do dendrograma (do grego *dendron* = árvore e *grama* = lista ou registro), que mostra em cada etapa o esquema de aglomeração e a distância entre os grupos. De acordo com Quintal (2006), esse tipo de representação hierárquica foi desenvolvida primeiramente em biologia, aparecendo em vários formatos gráficos consoantes ao software que os produz, em suma é um gráfico bidimensional que ilustra as fusões ou divisões em cada nível da análise de agrupamento.

Em um dendrograma, a escala vertical indica o nível de similaridade e a escala horizontal indica os objetos, em uma ordem conveniente. A altura das linhas verticais que partem dos objetos corresponde ao nível em que os objetos são considerados semelhantes.

A figura abaixo representa o dendrograma do exemplo anterior, o qual foi realizado utilizando o software R (o *script* está disponível no apêndice A). Para fazer o esboço, "na mão" ou utilizando softwares como o Geogebra, basta observar as distâncias entre cada agrupamento, por exemplo, a menor distância na matriz D_1 foi 1 entre os objetos BC, logo podemos fazer a ligação entre eles no dendrograma com a altura 1 no eixo vertical.

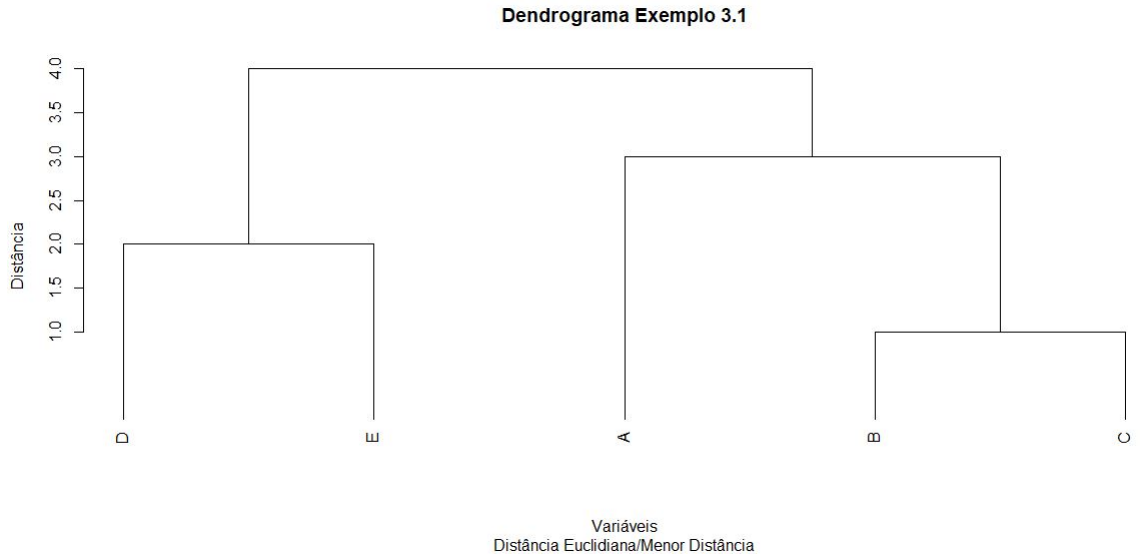


Figura 9 – Dendrograma Exemplo 3.1

A análise de um dendrograma permite estimar o número mais natural de clusters de um conjunto de dados. Para isso, pode-se escolher de acordo com a conveniência da pesquisa através de cortes horizontais no dendrograma conforme ilustrado na figura abaixo:

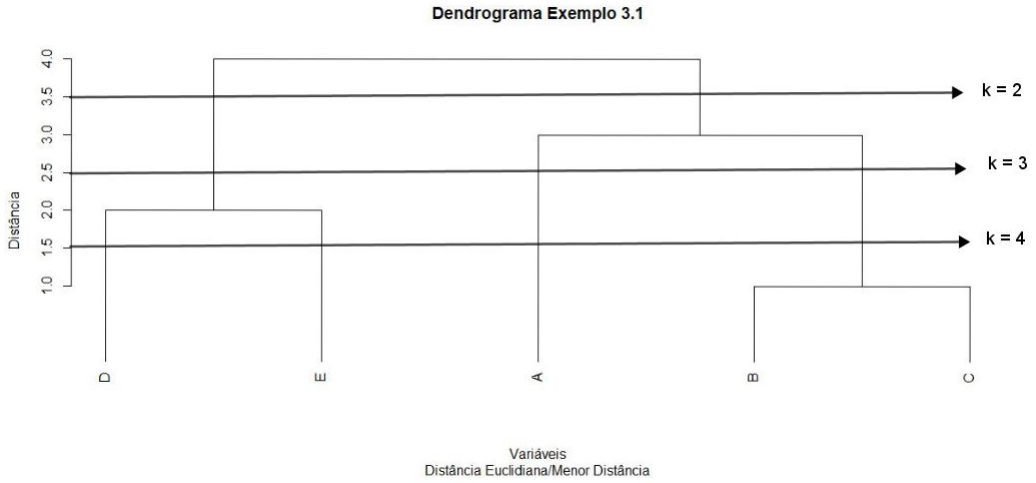


Figura 10 – Quantidade de grupos (k) no dendrograma Exemplo 3.1.

Note que a quantidade de grupos (k) coincide com a quantidade de interseções do corte com o dendrograma. Outra forma de verificar a quantidade de grupos é através da média das distâncias entre os agrupamentos.

A seguir, vamos apresentar um exemplo com duas variáveis, utilizando a distância do taxista e a ligação da maior distância.

Exemplo 3.2. *Considere as seguintes observações:*

Variáveis	A	B	C	D	E
V_1	3	4	4	2	6
V_2	2	5	6	7	6

Tabela 2 – Dados - Exemplo 3.2.

Vamos utilizar a métrica do taxista para calcular a distância entre cada uma das amostras, note que temos 5 amostras e duas variáveis observadas, assim temos um par ordenado para cada amostra, por exemplo, A é representada pelo par (3,2). Calculando as distâncias, temos:

$$
\begin{aligned}
d'(A,A) &= d'((3,2),(3,2)) = |3-3| + |2-2| = 0 = d'(B,B) = ... = d'(E,E);\\
d'(A,B) &= d'((3,2),(4,5)) = |3-4| + |2-5| = 4 = d'(B,A);\\
d'(A,C) &= d'((3,2),(4,6)) = |3-4| + |2-6| = 5 = d'(C,A);\\
d'(A,D) &= d'((3,2),(2,7)) = |3-2| + |2-7| = 6 = d'(D,A);\\
d'(A,E) &= d'((3,2),(6,6)) = |3-6| + |2-6| = 7 = d'(E,A);\\
d'(B,C) &= d'((4,5),(4,6)) = |4-4| + |5-6| = 1 = d'(C,B);\\
d'(B,D) &= d'((4,5),(2,7)) = |4-2| + |5-7| = 4 = d'(D,B);\\
d'(B,E) &= d'((4,5),(6,6)) = |4-6| + |5-6| = 3 = d'(D,B);\\
d'(C,D) &= d'((4,6),(2,7)) = |4-2| + |6-7| = 3 = d'(D,C);\\
d'(C,E) &= d'((4,6),(6,6)) = |4-6| + |6-6| = 2 = d'(D,B);\\
d'(D,E) &= d'((2,7),(6,6)) = |2-6| + |7-6| = 5 = d'(E,D).
\end{aligned}
$$

Assim a Matriz de distâncias fica da seguinte forma:

$$
D'_1 = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array}
\begin{array}{c}
\begin{array}{ccccc} A & B & C & D & E \end{array} \\
\left[\begin{array}{ccccc}
0 & & & & \\
4 & 0 & & & \\
5 & 1 & 0 & & \\
6 & 4 & 3 & 0 & \\
7 & 3 & 2 & 5 & 0
\end{array}\right]
\end{array}.
$$

Note que os elementos mais próximos, utilizando a distância do taxista, são os elementos B e C. Realizamos, então, o agrupamento utilizando o método da ligação de "Maior Distância".

Como $d'_{BC} = 1$*, assim obtemos um novo grupo: (BC), excluímos as linhas e colunas correspondentes a B e C, comparamos as distâncias utilizando a maior distância e as novas distâncias são dadas por:*

$$d_{(BC)A} = max\{d_{BA}, d_{CA}\} = max\{4, 5\} = 5;$$
$$d_{(BC)D} = max\{d_{BD}, d_{CD}\} = max\{4, 3\} = 4;$$
$$d_{(BC)E} = max\{d_{BE}, d_{CE}\} = max\{3, 2\} = 3.$$

Assim obtemos a nova matriz:

$$D'_2 = \begin{array}{c} \\ A \\ BC \\ D \\ E \end{array} \begin{array}{c} \begin{array}{cccc} A & BC & D & E \end{array} \\ \begin{bmatrix} 0 & & & \\ 5 & 0 & & \\ 6 & 4 & 0 & \\ 7 & 3 & 5 & 0 \end{bmatrix} \end{array}.$$

Temos agora que a menor distância é $d'_{(BC)E} = 3$*, assim obtemos um novo grupo (BCE). Novas distâncias:*

$$d_{(BCE)A} = max\{d_{(BC)A}, d_{EA}\} = max\{5, 7\} = 7;$$
$$d_{(BCE)D} = max\{d_{(BC)D}, d_{(BC)E}\} = max\{4, 3\} = 4.$$

Assim obtemos a nova matriz:

$$D'_3 = \begin{array}{c} \\ A \\ BCE \\ D \end{array} \begin{array}{c} \begin{array}{ccc} A & BCE & D \end{array} \\ \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 6 & 4 & 0 \end{bmatrix} \end{array}.$$

A menor distância em D'_3 *é dada por* $d'_{D(BCE)} = 4$*, assim obtemos o grupo (BCDE). Atualização das distâncias:*

$$d_{(BCDE)A} = max\{d_{A(BCE)}, d_{AD}\} = max\{7, 6\} = 7.$$

Assim obtemos a nova matriz:

$$D'_4 = \begin{array}{c} \\ A \\ BCDE \end{array} \begin{array}{c} \begin{array}{cc} A & BCDE \end{array} \\ \begin{bmatrix} 0 & \\ 7 & 0 \end{bmatrix} \end{array}.$$

Assim os grupos (A) e (BCDE), são agrupados em um único cluster e a distância máxima entre seus objetos igual a 7 unidades.

Segue abaixo o dendrograma referente a esse exemplo:

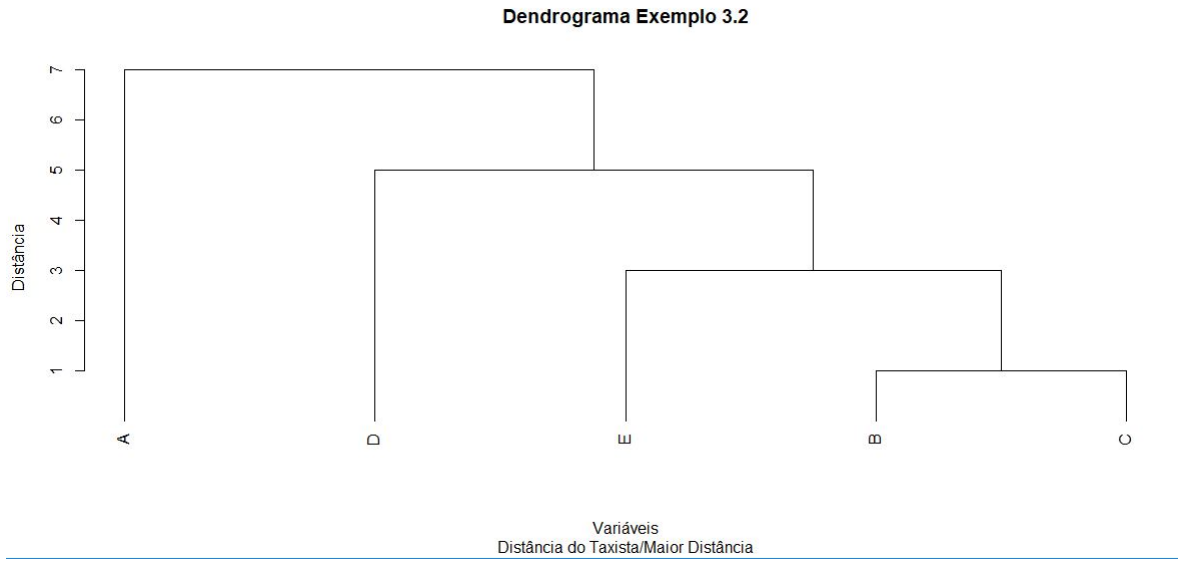


Figura 11 – Dendrograma Exemplo 3.2.

Capítulo 4

Agrupamento Não Hierárquico

Os procedimentos não hierárquicos são utilizados para agrupar indivíduos de acordo com um número de grupos predeterminado pelo pesquisador. De modo geral, foram desenvolvidos para agrupar elementos em k grupos, onde k é a quantidade de grupos definida previamente.

A ideia central é escolher uma partição inicial dos elementos, em seguida, alterar os membros dos grupos para obter a melhor partição. Note que, diferente dos métodos hierárquicos, os métodos não hierárquicos necessitam de informações iniciais a respeito da quantidade de grupos, essa quantidade de grupos pode ser obtida por procedimentos hierárquicos de forma exploratória.

De acordo com Favero et. al. (2009), os métodos hierárquicos não requerem o cálculo e o armazenamento da matriz de distâncias em cada etapa do processo, o que reduz o tempo computacional e possibilita sua aplicação em grandes bases de dados. Segundo o mesmo autor, a probabilidade de acontecerem classificações erradas é menor nos métodos não hierárquicos, mas, em contrapartida há uma dificuldade de se estabelecer um número de grupos.

Dos métodos não hierárquicos, o mais popular é o *k-means*, também chamado de *k-médias*. Esse método pode ser utilizado para o agrupamento de grandes grupos de observações, produzindo apenas uma solução para o número de grupos que deve ser definido pelo pesquisador, enquanto que o método hierárquico fornece uma série de soluções correspondentes, no qual podemos escolher uma quantidade favorável de grupos.

Segundo Quintal (2006), o algoritmo das k-médias pode ser descrito da seguinte maneira:

1) Escolha a quantidade K de grupos.

2) Calcule os k centros, esse é calculado pela média aritmética das coordenadas dos objetos que pertencem ao mesmo. A primeira divisão dos objetos para obter os centros

pode ser aleatória.

3) Associe cada ponto ao centro mais próximo.

4) Recalcule o centro de cada grupo.

5) Repita os passos 3 e 4 até que nenhum elemento mude de grupo.

Segue o exemplo para detalhar esse algoritmo.

Exemplo 4.1. *Dado o conjunto de 4 objetos (A,B,C,D) e duas variáveis (x_1,x_2), use o algoritmo k-Médias para identificar 2 clusters (k=2)*

Variáveis	A	B	C	D
x_1	2	5	1	8
x_2	0	2	4	4

Tabela 3 – Dados - Exemplo 4.1.

1) Para particionar os dados em dois grupos, vamos primeiramente particiona-los em dois grupos aleatórios para encontrar os centroides, AD e BC. Em seguida, vamos calcular os centroides; para este método, utilizamos a média aritmética simples.

Objeto	$\bar{X}_1$	$\bar{X}_2$
AD	$\frac{2+8}{2} = 5$	$\frac{0+4}{2} = 2$
BC	$\frac{5+1}{2} = 3$	$\frac{2+4}{2} = 3$

Assim, para AD temos como centroide o ponto (5,2), denotamos $C_{AD} = (5, 2)$, do mesmo modo que $C_{BC} = (3, 3)$.

2) Agora calculamos as respectivas distâncias de cada elemento aos centroides encontrados anteriormente. Cada elemento será associado ao centroide mais próximo, para isso, vamos utilizar a distância euclidiana quadrática.

$$d^2_{A(AD)} = (2-5)^2 + (0-2)^2 = 13;$$

$$d^2_{A(BC)} = (2-3)^2 + (0-3)^2 = 10.$$

Como $d^2_{A(BC)} < d^2_{A(AD)}$, mudamos o elemento A para o grupo BC, e recalculamos os centroides.

Objeto	$\bar{X}_1$	$\bar{X}_2$
D	8	4
ABC	$\frac{5+2+1}{3} = 2,667$	$\frac{0+4+2}{3} = 2$

Recalculando as distâncias dos objetos para o centroide dos grupos, obtemos:

$$d^2_{AD} = (2-8)^2 + (0-4)^2 = 52;$$
$$d^2_{A(ABC)} = (2-2,667)^2 + (0-2)^2 = 4,44;$$
$$d^2_{BD} = (5-8)^2 + (2-4)^2 = 13;$$
$$d^2_{B(ABC)} = (5-2,667)^2 + (2-2)^2 = 5,44;$$
$$d^2_{CD} = (1-8)^2 + (4-4)^2 = 49;$$
$$d^2_{C(ABC)} = (1-2,667)^2 + (4-2)^2 = 6,77.$$

Nenhuma realocação deve ser realizada, pois os objetos têm menor distância para os respectivos grupos aos quais eles pertencem. Desse modo, obtemos os grupos ABC e CD.

Uma representação prévia dos pontos antes do agrupamento pode dar uma pista, neste caso que são duas variáveis, de como serão os grupos. A representação final do agrupamento também pode ser realizada no plano cartesiano com os centroides e os grupos com cores distintas, conforme segue na imagem abaixo, que foi realizada no software Geogebra, no Apêndice A apresentamos o script para a realização no software R:

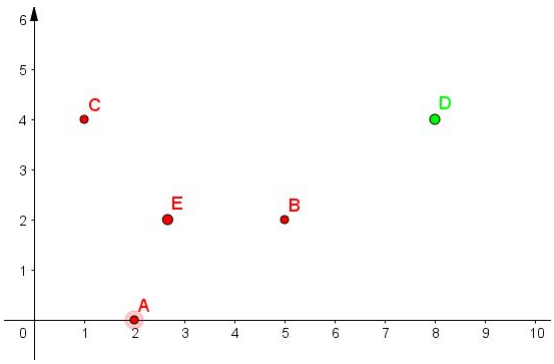


Figura 12 – Representação do algoritmo das k-médias.

A principal diferença entre os algoritmos hierárquicos e não hierárquicos é o uso da matriz de distâncias inicial, sendo uma das vantagens desse método que se torna mais

rápido e simples de realizar com uma quantidade grande de dados. De modo geral, métodos hierárquicos são preferidos quando serão analisadas várias alternativas de agrupamento e o tamanho da amostra é moderado (de 300 a 1000 objetos), enquanto que os métodos não-hierárquicos são preferidos quando o número de grupos é conhecido e tem a presença de *outliers*, isto é, são objetos que se diferenciam drasticamente de todos os outros, já que os métodos não-hierárquicos são menos influenciados pelos mesmos, vale lembrar que caso tenham *outliers* é interessante analisar os dados, para descartar algum erro de coleta. Os métodos de agrupamento não hierárquicos são usados também quando a pesquisa apresenta uma grande número de objetos a serem agrupados.

Capítulo 5

Proposta de Atividades

Vimos que os documentos norteadores enfatizam a importância do ensino de estatística, no entanto o tratamento estatístico durante todo o ensino básico é realizado apenas em uma variável. Destacamos ainda que o ensino de estatística deve proporcionar meios para que o aluno possa "Construir e interpretar tabelas e gráficos de frequências, com base em dados obtidos em pesquisas por amostras estatísticas, incluindo ou não o uso de softwares que inter-relacionem estatística, geometria e álgebra"(BNCC 2018, p.531), além de interpretar e comparar conjuntos de dados estatísticos por meio de diferentes tipos de gráficos.

Diante do que foi analisado nos documentos norteadores do processo de ensino e aprendizagem de matemática na Educação Básica, no presente capítulo apresentamos uma proposta de ensino, destinada aos alunos da educação básica, com o objetivo de compreender que também é possível analisar dados com várias variáveis através das técnicas de agrupamento. Nesta proposta são apresentadas tarefas que buscam proporcionar o entendimento de alguns conceitos fundamentais das técnicas de agrupamento, evidenciando a sua importância na vida cotidiana e na aplicação de outros conteúdos da disciplina de matemática, como o cálculo de distâncias, o cálculo de média, o esboço de gráficos, a manipulação de matrizes, entre outros.

5.1 Tarefa 1: características dos animais

Objetivo: compreender os princípios básicos das técnicas de agrupamento e sua utilização no cotidiano.

Série indicada: essa atividade pode ser utilizada a partir das séries iniciais do ensino fundamental.

Agrupe os animais abaixo de acordo com as suas características semelhantes:

Figura 13 – Imagem dos animais a serem agrupados.

a) Como você justifica a escolha dos grupos?

Encaminhamentos: o professor pode começar explicando sobre as classificações (agrupamentos) que realizamos em nosso cotidiano e, com isso, as decisões que devemos tomar para formar determinados grupos, isso é em geral o que se realiza nas técnicas de agrupamento.

Orientação de Resposta: uma possibilidade de resposta seria agrupá-los pelas classes do reino dos animais (mamíferos, aves, peixes e mamífero ovíparo), outra seria pensando em seu habitat (aquáticos e terrestres). Nessa questão, é interessante que cada aluno explique o motivo do seu agrupamento, elencando as características que foram analisadas por cada aluno.

Segue abaixo o exemplo de dois grupos segundo o habitat dos animais apresentados.

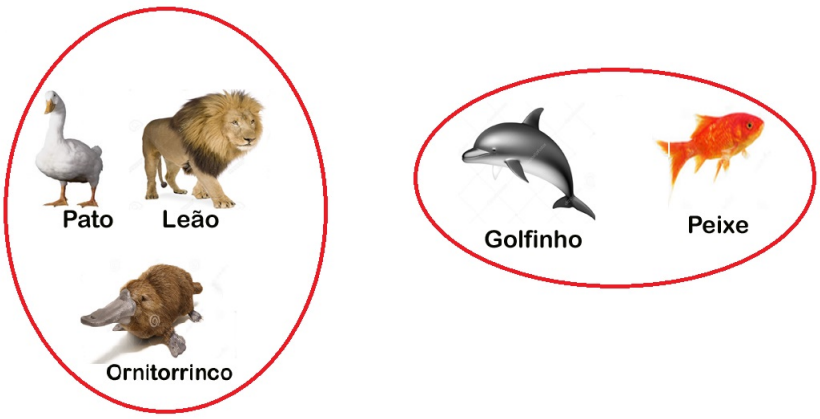


Figura 14 – Exemplo de agrupamento segundo o habitat.

b) É possível agrupá-los de outra forma? Converse com os seus colegas sobre como realizaram o agrupamento e as diferenças encontradas nos agrupamentos.

Encaminhamentos: para finalizar essa atividade é interessante o professor mencionar que as características de cada animal podem ser consideradas uma variável, sendo interessante utilizar a mesma ordem das características para tomarmos uma melhor decisão de agrupamento. Podemos considerar, por exemplo, as características: terrestre ou aquático.

Orientação de Resposta: nessa questão, o professor pode retomar alguns agrupamentos mencionados pelos alunos no item a).

5.2 Tarefa 2: grupos de supermercados

Objetivo: compreender os princípios básicos das técnicas de agrupamento não hierárquico e sua utilização no cotidiano.

Série indicada: essa atividade pode ser utilizada a partir do sétimo ano do ensino fundamental, após a noção de plano cartesiano e estatística que é trabalhada nessa série.

Uma pesquisa analisou o preço de itens da cesta básica (Produtos 1) e produtos de limpeza (Produtos 2) de mesma marca em cinco supermercados da cidade, conforme descrito na tabela abaixo:

	Produtos 1 (R$)	Produtos 2 (R$)
Mercado A	70	50
Mercado B	60	57
Mercado C	55	60
Mercado D	65	50
Mercado E	70	45

Tabela 4 – Dados Tarefa 2.

É possível separar esses supermercados em dois grupos diferentes de acordo com o preço desses produtos? Como podemos escolher o melhor mercado para as compras?

Encaminhamentos: Sim, podemos utilizar os preços para classificar os supermercados e, através dessa classificação, escolher o melhor mercado.

Como temos dados com duas variáveis, nessa tarefa o professor pode indicar a possibilidade de plotar os dados no eixo cartesiano, em seguida os alunos devem realizar os agrupamentos argumentando sobre os motivos da sua escolha. É importante ressaltar as possibilidades de utilização desse tipo de análise quando temos uma grande quantidade de dados. A figura abaixo indica os dados no plano cartesiano.

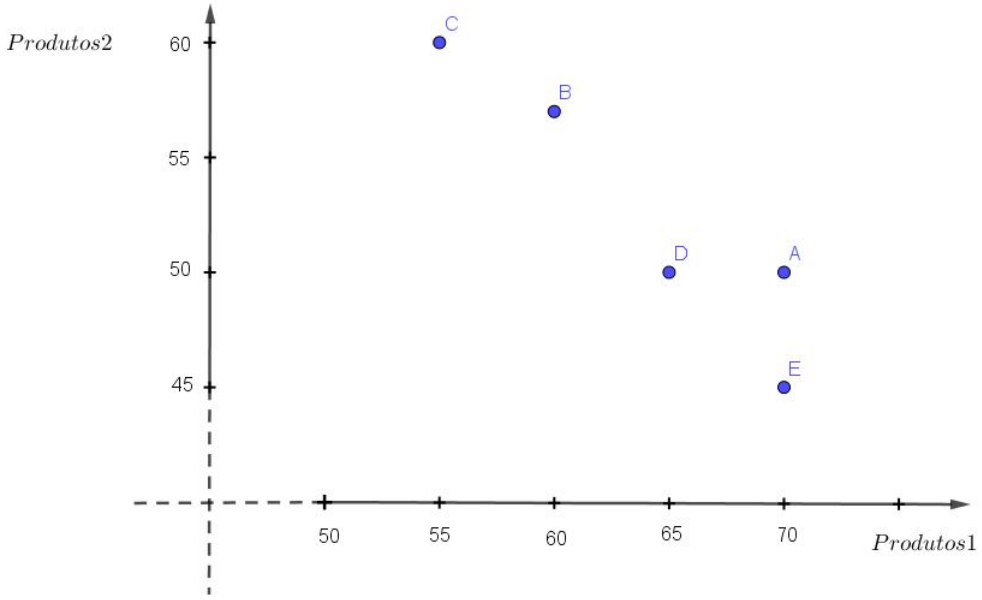


Figura 15 – Representação dos dados da Tarefa 2 no plano cartesiano.

Orientação de resposta: após esboçar os dados, é possível montar os grupos observando a distância entre os pontos no plano cartesiano. É interessante argumentar calculando as distâncias entre os pontos, conforme apresentamos no Método das k-médias e no capítulo 4.

Como os mercados ADE e CB aparentam estar mais próximos, vamos utilizar essa partição como inicial e encontrar os centroides.

1) Calcular os centroides.

Mercado	$\bar{X}_1$	$\bar{X}_2$
$C_1 = ADE$	$\frac{70+65+70}{3} = 68,3$	$\frac{50+50+45}{3} = 48,3$
$C_2 = BC$	$\frac{60+55}{2} = 57,5$	$\frac{57+60}{2} = 58,5$

Assim obtemos os seguintes centroides: $C_1 = (68,3; 48,3)$ e $C_2 = (57,5; 58,5)$.

2) Vamos calcular as respectivas distâncias dos objetos aos centroides, utilizando uma distância a sua escolha, nesse exemplo, utilizaremos a distância euclidiana:

$$d_{AC_1} = \sqrt{(70 - 68,3)^2 + (50 - 48,3)^2} = 2,4;$$
$$d_{AC_2} = \sqrt{(70 - 57,5)^2 + (50 - 58,5)^2} = 15,1.$$

Note que a menor distância é de A até o centroide ADE, logo A permanece neste grupo. A mesma verificação pode ser realizada com os demais objetos (BCDE) aos centroides calculados anteriormente, o que possibilita a representação de dois grupos como a seguir:

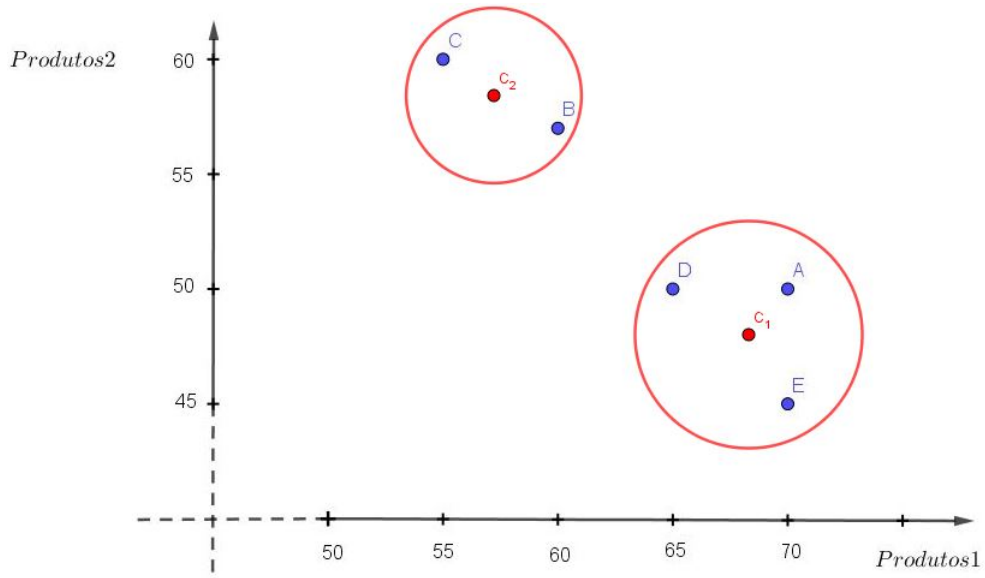


Figura 16 – Representação dos dados da Tarefa 2 no plano cartesiano.

Após esta análise, podemos calcular a média dos preços dos dois grupos, podemos concluir que os mercados agrupados no grupo 2 ou em torno de C_2 apresentam uma média de preços menor em relação ao grupo 1.

5.3 Tarefa 3: características dos planetas do sistema solar

Objetivo: compreender e aplicar o conceito de distância.

Série indicada: essa atividade pode ser utilizada a partir do sétimo ano do ensino fundamental.

A tabela abaixo representa dados de translação, diâmetro e massa de alguns dos planetas do sistema solar, normalizados a partir dos dados da Terra.

Planeta	Translação	Diâmetro	Massa
Mercúrio	0.241	0.383	0.055
Vênus	0.615	0.949	0.815
Terra	1	1	1
Marte	1.88	0.533	0.107

Tabela 5 – Dados Tarefa 3.

A partir desses dados, escolha uma distância de sua preferência e utilize a Matriz de Distâncias para calcular a semelhança entre esses planetas.

Encaminhamentos: para esta tarefa o professor deve inicialmente explicar o conceito de distância e as distâncias mais utilizadas presentes na unidade 2 deste trabalho. Lembrando que é necessário considerar os dados da tabela como pontos do $\mathbb{R}^3$, ou seja, os planetas serão representados pelos pontos ordenados:
Mercúrio = (0.241, 0.383, 0.055);
Vênus = (0.615, 0.949, 0.815);
Terra = (1, 1, 1);
Marte = (1.88, 0.533, 0.107).

Métodos de agrupamento que utilizam a Matriz de Distância para agrupar os elementos são chamados de métodos hierárquicos. Para facilitar, apresente a seguinte tabela para preenchimento das distâncias:

	Mercúrio	Vênus	Terra	Marte
Mercúrio				
Vênus				
Terra				
Marte				

Tabela 6 – Matriz de distâncias para preenchimento - Tarefa 3.

Nessa atividade, é interessante a comparação das diferentes medidas de distâncias, por exemplo, calculamos a seguir a distância de Terra a Marte com as distâncias Euclidiana e do Máximo.

Distância Euclidiana:

$$d(Terra, Marte) = d((1,1,1),(1.88,0.533,0.107)) =$$
$$\sqrt{(1-1.88)^2 + (1-0.533)^2 + (1-0.107)^2} =$$
$$\sqrt{(-0.88)^2 + (0.467)^2 + (0.893)^2} =$$
$$\sqrt{0.774 + 0.218 + 0.797} =$$
$$\sqrt{1.79} = 1.33.$$

Distância do Máximo

$$d''(Terra, Marte) = d''((1,1,1),(1.88, 0.533, 0.107)) =$$

$$max\{|1-1.88|, |1-0.533|, |1-0.107|\} =$$

$$max\{|-0.88|, |0.467|, |0.893|\} = 0.893.$$

Orientação de resposta: nessa atividade, cada aluno pode escolher uma distância e comparar com o resultado dos colegas, no exemplo abaixo, escolhemos a distância do Máximo.

$$d''(Mercúrio, Mercúrio) = d''(Vênus, Vênus) = ... = d^2(Marte, Marte) = 0.$$

$$d''(Mercúrio, Vênus) = d''((0.241, 0.383, 0.055),(0.615, 0.949, 0.815)) =$$

$$max\{|0.241-0.615|, |0.383-0.949|, |0.055-0.815|\} =$$

$$max\{|-0.374|, |-0.566|, |-0.76|\} = 0.76.$$

$$d''(Mercúrio, Terra) = d''((0.241, 0.383, 0.055),(1,1,1)) =$$

$$max\{|0.241-1|, |0.383-1|, |0.055-1|\} =$$

$$max\{|-0.759|, |-0.617|, |-0.945|\} = 0.945.$$

$$d''(Mercúrio, Marte) = d''((0.241, 0.383, 0.055),(1.88, 0.533, 0.107)) =$$

$$max\{|0.241-1.88|, |0.383-0.533|, |0.055-0.107|\} =$$

$$max\{|-1.639|, |-0.15|, |-0,052|\} = 1.639.$$

$$d''(Vênus, Terra) = d''((0.615, 0.949, 0.815),(1,1,1)) =$$

$$max\{|0.615-1|, |0.949-1|, |0.815-1|\} =$$

$$max\{|-0.385|, |-0.051|, |-0,185|\} = 0.385.$$

$$d''(Vênus, Marte) = d''((0.615, 0.949, 0.815),(1.88, 0.533, 0.107)) =$$

$$max\{|0.615-1.88|, |0.949-0.533|, |0.815-0.107|\} =$$

$$max\{|-1.245|, |-0.416|, |-0.708|\} = 1.245.$$

$$d''(Terra, Marte) = d''((1,1,1),(1.88, 0.533, 0.107)) =$$
$$max\{|1-1.88|, |1-0.533|, |1-0.107|\} =$$
$$max\{|-0.88|, |0.467|, |0,893|\} = 0.893.$$

Como a Matriz de distâncias é simétrica, ou seja,
$d''(Mercúrio, Vênus) = d''(Vênus, Mercúrio)$, obtemos a seguinte matriz:

$$D' = \begin{array}{c} \\ Mercúrio \\ Vênus \\ Terra \\ Marte \end{array} \begin{array}{c} \begin{array}{cccc} Mercúrio & Vênus & Terra & Marte \end{array} \\ \begin{bmatrix} 0 & & & \\ 0.76 & 0 & & \\ 0.945 & 0.385 & 0 & \\ 1.639 & 1.245 & 0.893 & 0 \end{bmatrix} \end{array}.$$

Conforme vimos no capítulo 3, quanto menor a distância entre dois objetos maior é a semelhança entre eles. Assim, de acordo com os dados apresentados (translação, diâmetro e massa), podemos afirmar que Terra e Vênus são os planetas mais semelhantes, em seguida, os mais semelhantes são Vênus e Mercúrio.

5.4 Tarefa 4: adubo, qual utilizar?

Objetivo: utilizar as técnicas de agrupamento hierárquico para encontrar a melhor solução para o problema.

Série indicada: essa atividade pode ser utilizada a partir do segundo ano do ensino médio, após o conteúdo de matrizes e o estatístico que é apresentado nessa série.

A tabela abaixo indica quais foram os resultados médios após a aplicação de quatro adubos em uma plantação de laranjas, as variáveis analisadas foram: qualidade do fruto e a quantidade de suco em uma escala de 1 a 5.

Adubo	Qualidade	Suco
A1	3	2
A2	4	3
A3	2	4
A4	2	5

Tabela 7 – Dados Tarefa 4.

Analisando os dados da tabela, determine qual seria a melhor opção de adubo para o agricultor utilizar, sabendo que os adubos A1 e A3 apresentam o menor custo, enquanto que o custo de A2 e A4 é o dobro do valor de custo dos adubos anteriores.

Encaminhamentos: nessa tarefa, o professor pode apresentar os passos do agrupamento hierárquico presentes no capítulo 3.

	A1	A2	A3	A4
A1				
A2				
A3				
A4				

Tabela 8 – Matriz de distâncias para preenchimento - Tarefa 4.

Orientação de resposta:

1) Iniciamos com 4 grupos e considerando os dados como pontos do $\mathbb{R}^2$:
A1 = (3,2);
A2 = (4,3);
A3 = (2,4);
A4 = (2,5).

Calculamos a Matriz de distâncias conforme a distância escolhida, na matriz abaixo, escolhemos a distância euclidiana quadrática.

Vamos calcular a distância entre os quatro objetos.

$$d^2_{A1,A1} = d^2_{A2,A2} = d^2_{A3,A3} = d^2_{A4,A4} = 0;$$
$$d^2_{A1,A2} = (3-4)^2 + (2-3)^2 = 2;$$
$$d^2_{A1,A3} = (3-2)^2 + (2-4)^2 = 5;$$
$$d^2_{A1,A4} = (3-2)^2 + (2-5)^2 = 10;$$
$$d^2_{A2,A3} = (4-2)^2 + (3-4)^2 = 5;$$
$$d^2_{A2,A4} = (4-2)^2 + (3-5)^2 = 8;$$
$$d^2_{A3,A4} = (2-2)^2 + (4-5)^2 = 1.$$

$$D'_1 = \begin{array}{c} \\ A1 \\ A2 \\ A3 \\ A4 \end{array} \begin{array}{c} \begin{array}{cccc} A1 & A2 & A3 & A4 \end{array} \\ \begin{bmatrix} 0 & & & \\ 2 & 0 & & \\ 5 & 5 & 0 & \\ 10 & 8 & 1 & 0 \end{bmatrix} \end{array}.$$

2) Identifique o menor elemento da matriz de distâncias para encontrar o par de grupos mais similares.

Note que a menor distância é dada por $d^2_{A3,A4}$.

3) Reunimos os dois grupos que apresentam a menor distância.

Realizamos então o agrupamento de A3 e A4 utilizando o método da "Menor Distância", assim obtemos um novo grupo: (A3A4), excluímos as linhas e colunas correspondentes a A3 e A4, comparamos as distâncias utilizando a menor distância, as novas distâncias são dadas por:

$$d_{(A3A4)A1} = min\{d_{A3A1}, d_{A4A1}\} = min\{5, 10\} = 5;$$

$$d_{(A3A4)A2} = min\{d_{A3A2}, d_{A4A2}\} = min\{5, 8\} = 5.$$

Assim obtemos a nova matriz de distâncias:

$$D'_2 = \begin{array}{c} \\ A1 \\ A2 \\ A3A4 \end{array} \begin{array}{c} \begin{array}{cccc} A1 & A2 & A3A4 \end{array} \\ \begin{bmatrix} 0 & & \\ 2 & 0 & \\ 5 & 5 & 0 \end{bmatrix} \end{array}.$$

A menor distância em D'_2 é entre A1 e A2, assim obtemos um novo grupo (A1A2), excluímos as linhas e colunas correspondentes e comparamos novamente utilizando a menor distância, as novas distâncias são dadas por:

$$d_{(A1A2)A3A4} = min\{d_{A1(A3A4)}, d_{A2(A3A4)}\} = min\{5, 5\} = 5.$$

Assim os grupos (A1A2) e (A3A4), são agrupados num único cluster, cuja distância mínima entre seus objetos é de 5 unidades.

Como os adubos A3 e A4 apresentam melhores resultados gerais nas variáveis analisadas (qualidade do fruto e suco) e estão próximos no agrupamento, podemos optar pelo adubo de menor custo, que nesse exemplo, é o adubo A3.

5.5 Tarefa 5: distância entre países

Objetivo: utilizar o agrupamento hierárquico e a sua representação (dendrograma) para resolver um problema.

Série indicada: essa atividade pode ser utilizada a partir do sétimo ano do ensino fundamental.

Os dados da Tabela abaixo representam, de acordo com o banco de dados da ONU (2002) os índices de expectativa de vida, educação e renda (PIB) de 6 países. Quanto maior o valor do índice, melhor é a qualidade do país.

Países	Expectativa de vida	Educação	PIB
Reino Unido	0,88	0,99	0,91
Canadá	0,9	0,98	0,94
França	0,89	0,97	0,92
Paraguai	0,75	0,83	0,63
Brasil	0,71	0,83	0,72
Egito	0,7	0,62	0,6

Tabela 9 – Dados Tarefa 5.

Como podemos agrupar esses países baseando-nos nesses dados?

Encaminhamentos: para essa tarefa, podemos utilizar o agrupamento hierárquico com uma distância previamente escolhida.

Orientação de resposta: vamos considerar os dados como pontos do $\mathbb{R}^3$. Logo, temos:
Reino Unido: $R = (0.88, 0.99, 0.91)$;
Canadá: $C = (0.9, 0.98, 0.94)$;
França: $F = (0.89, 0.97, 0.92)$;
Paraguai: $P = (0.75, 0.83, 0.63)$;
Brasil: $B = (0.71, 0.83, 0.72)$;
Egito: $E = (0.7, 0.62, 0.6)$.

Vamos utilizar a distância do Taxista para realizar o agrupamento:

$$d'(R,R) = d'(C,C) = \cdots = d'(E,E) = 0;$$

$$d'(R,C) = |0.88 - 0.9| + |0.99 - 0.98| + |0.91 - 0.94| = 0.06;$$

$$d'(R,F) = |0.88 - 0.89| + |0.99 - 0.97| + |0.91 - 0.92| = 0.04;$$

$$d'(R,P) = |0.88 - 0.75| + |0.99 - 0.83| + |0.91 - 0.63| = 0.57;$$

$$d'(R,B) = |0.88 - 0.71| + |0.99 - 0.83| + |0.91 - 0.72| = 0.52;$$

$$d'(R,E) = |0.88 - 0.7| + |0.99 - 0.62| + |0.91 - 0.6| = 0.86;$$

$$d'(C,F) = |0.9 - 0.89| + |0.98 - 0.97| + |0.94 - 0.92| = 0.04;$$

$$d'(C,P) = |0.9 - 0.75| + |0.98 - 0.83| + |0.94 - 0.63| = 0.61;$$

$$d'(C,B) = |0.9 - 0.71| + |0.98 - 0.83| + |0.94 - 0.72| = 0.56;$$

$$d'(C,E) = |0.9 - 0.7| + |0.98 - 0.62| + |0.94 - 0.6| = 0.90;$$

$$d'(F,P) = |0.89 - 0.75| + |0.97 - 0.83| + |0.92 - 0.63| = 0.57;$$

$$d'(F,B) = |0.89 - 0.71| + |0.97 - 0.83| + |0.92 - 0.72| = 0.52;$$

$$d'(F,E) = |0.89 - 0.7| + |0.97 - 0.62| + |0.92 - 0.6| = 0.86;$$

$$d'(P,B) = |0.75 - 0.71| + |0.83 - 0.71| + |0.63 - 0.72| = 0.13;$$
$$d'(P,E) = |0.75 - 0.7| + |0.83 - 0.62| + |0.63 - 0.6| = 0.29;$$
$$d'(B,E) = |0.71 - 0.7| + |0.83 - 0.62| + |0.72 - 0.6| = 0.34.$$

Assim obtemos a matriz de distâncias:

$$D'_1 = \begin{array}{c} \\ R \\ C \\ F \\ P \\ B \\ E \end{array} \begin{array}{c} \begin{array}{cccccc} R & C & F & P & B & E \end{array} \\ \begin{bmatrix} 0 & & & & & \\ 0.06 & 0 & & & & \\ 0.04 & 0.04 & 0 & & & \\ 0.57 & 0.61 & 0.57 & 0 & & \\ 0.52 & 0.56 & 0.52 & 0.13 & 0 & \\ 0.86 & 0.90 & 0.86 & 0.29 & 0.34 & 0 \end{bmatrix} \end{array}.$$

Após calcular a matriz de distâncias, escolha o método de agrupamento (menor distância, maior distância ou distância média), conforme indicado no capítulo 3, em seguida faça o esboço do dendrograma para a melhor visualização dos grupos. É válido salientar que se duas distâncias iguais na matriz de distâncias, mesmo sendo a menor, não interfere no agrupamento, como na matriz acima nas distâncias de RF e CF, basta fazer um agrupamento por vez. Abaixo temos o dendrograma utilizando o método da menor distância.

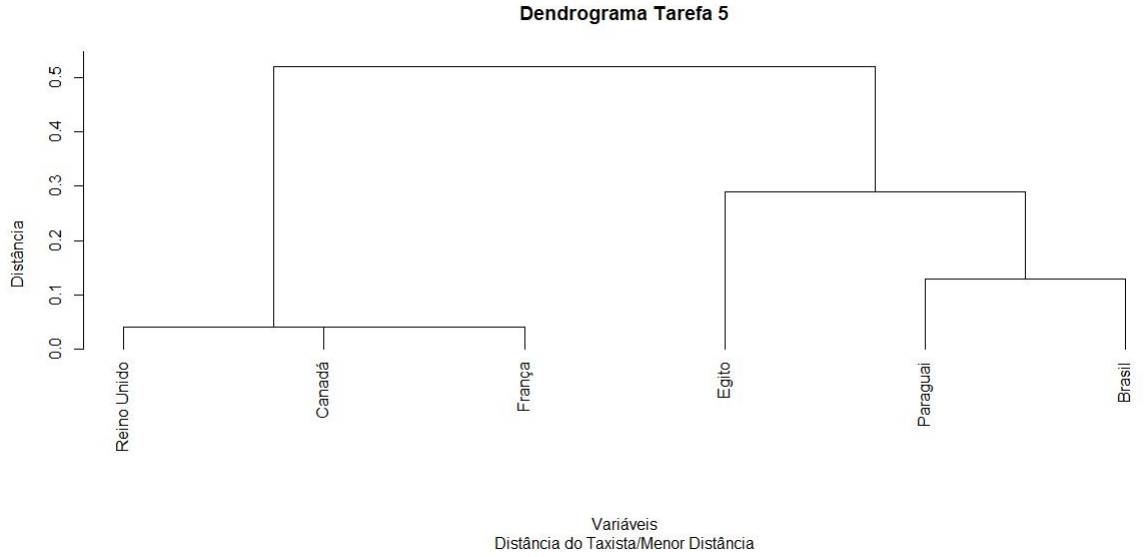


Figura 17 – Dendrograma Tarefa 5

Logo, utilizando a distância do taxista e o método da menor distância para o agrupamento podemos observar dois grupos de países: Reino Unido, Canadá e França, em um grupo, Egito, Paraguai e Brasil em outro grupo.

5.6 Tarefa 6: agrupamento utilizando dados da turma

Objetivo: perceber a utilização das técnicas de agrupamento em dados do nosso cotidiano.

Nessa tarefa, o professor pode coletar dados da turma ou da escola para realizar os agrupamentos, utilizando as técnicas mencionadas nos capítulos 3 e 4. Segue abaixo um exemplo com dados fictícios de idade e peso de alunos utilizando o método das k-médias.

Alunos	Idade	Peso
Aluno A	12	35
Aluno B	13	40
Aluno C	12	42
Aluno D	14	45
Aluno E	13	48
Aluno F	12	40

Tabela 10 – Tabela de dados - Tarefa 6.

Encaminhamentos: nesta atividade, o professor pode coletar dados com a turma, ou solicitar a coleta aos alunos para discussão em sala, utilizando tanto o método hierárquico quanto o não hierárquico para realizar os agrupamentos.

Orientação de resposta: vamos considerar os dados como pontos do $\mathbb{R}^2$, em seguida, podemos plotar os pontos no plano cartesiano, escolher uma partição conveniente e realizar o agrupamento pelo método das k-médias, conforme visto na tarefa 2 e no capítulo 4. Segue abaixo a representação dos dados no plano cartesiano.

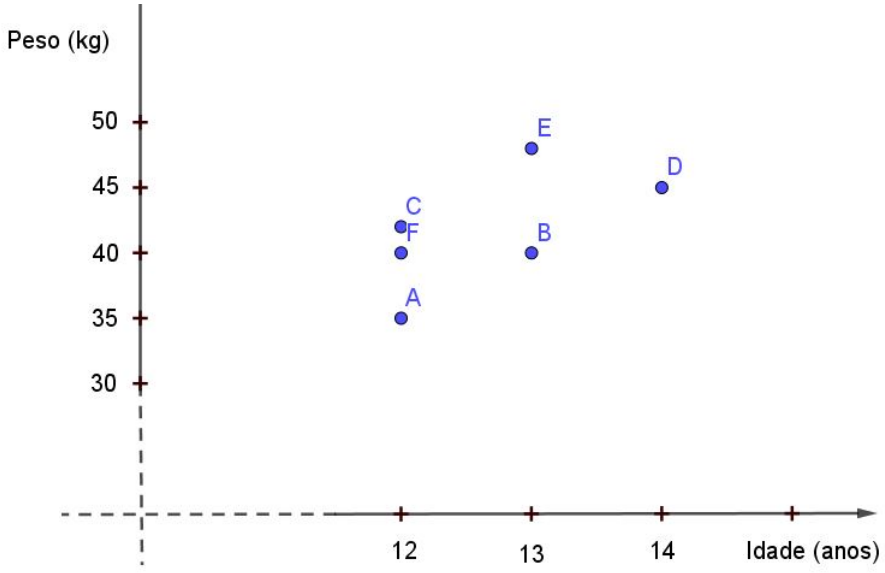


Figura 18 – Representação dos dados da Tarefa 6 no plano cartesiano

Capítulo 6

Considerações Finais

Percebemos, com esse trabalho, que as análises estatísticas são separadas em univariadas e multivaridadas, as técnicas de agrupamento fazem parte da análise multivariada e têm como característica agrupar ou classificar objetos conforme as suas semelhanças. Mesmo sendo reconhecida a importância do ensino da estatística nos documentos norteadores, este conteúdo é pouco explorado mediante a sua potencialidade.

Vale dizer que as classificações e agrupamentos estão presentes em nosso cotidiano e também nas diversas áreas da ciência, sendo amplamente utilizados em pesquisas das mais diversas áreas, desse modo, é importante uma abordagem das técnicas de agrupamento no ensino básico.

Para compreender as técnicas, precisamos primeiramente estudar o conceito de distância ou métrica, que foi apresentado no capítulo 2. As técnicas de agrupamento são divididas em agrupamento hierárquico e agrupamento não hierárquico. Os agrupamentos hierárquicos, em suma, consistem em separar objetos através de suas características em comum, apresentam como vantagem a representação gráfica através dos dendrogramas, que possibilitam a divisão em grupos conforme a conveniência da pesquisa. Já os agrupamentos não hierárquicos são utilizados para agrupar indivíduos de acordo com um número de grupos predeterminado pelo pesquisador e apresentam vantagem quando a quantidade de dados é grande, possibilitando a escolha a priori da quantidade de grupos e é menos afetado por *outliers*.

As tarefas propostas no capítulo 5 foram elaboradas com a intenção de mostrar que é possível abordar o conceito das técnicas de agrupamento, mesmo para séries iniciais ainda que sem estrutura matemática, mas desenvolvendo o sentido crítico e sistemático, mostrando que é possível uma outra análise estatística além da univariada. Outras atividades podem ser elaboradas visando também a interdisciplinaridade entre as disciplinas do ensino básico, fica como sugestão o uso das atividades em oficinas ou no decorrer do ensino regular, após o conteúdo estatístico da referida série.

Por fim esperamos que esse trabalho sirva de apoio para professores que tenham interesse de aprofundar os conhecimentos estatísticos de seus alunos, proporcionando o entendimento da análise multivariada de dados, em especial das técnicas de agrupamento, a estes futuros pesquisadores.

Referências

BRASIL, M. E. - Parâmetros Curriculares Nacionais, Brasília: MEC/SEMTEC, 4v., 1998.

BRASIL, M. E.- Base Nacional Comum Curricular Ensino Médio, Brasília: MEC/2017. Disponível em: http://basenacionalcomum.mec.gov.br/abase. Acesso em: 1 de outubro de 2018.

FAVERO, L. P. L; et al.- Análise de dados: modelagem multivariada para tomada de decisões, Rio de Janeiro, 2009.

FIGUEIREDO FILHO, D. B; et al - Classificando regimes políticos utilizando análise de conglomerados. Revista Opinião Pública, p. 109 - 128, 2012. Disponível em: http://www.scielo.br/pdf/op/v18n1/v18n1a06.pdf. Acesso em 4 de janeiro de 2019.

HAIR J. F; et al - Análise multivariada de dados, Porto Alegre: Bookman, 2009. Disponível em: https://dlscrib.com/download/hair-j-f-an-aacute-lise-multivariada-de-dados-6-ordf-edi-ccedil-atilde-o-pdf_58e6753ddc0d603035da97f8_pdf. Acesso em: 10 de outubro de 2018.

MOITA NETO J. M. - Uma visão didática-metodológica de estatística multivariada, Revista Filosofia da Ciência, p. 1-13, 2004. Disponível em: http://www.pucrs.br/ciencias/viali/especializa/realizadas/ceea/multivariada/textos/Moita_Neto.pdf. Acesso em: 4 de outubro de 2018.

LIMA, E. L - Espaços Métricos. Rio de Janeiro: IMPA Projeto Euclides, 1977.

LINDEN, R. - Técnicas de agrupamento, Revista de Sistemas de Informação da FSMA, v. 4, p. 18-36, 2009. Disponível em: http://www.fsma.edu.br/si/edicao4/FSMA SI 2009 2 Tutorial.pdf. Acesso em: 4 de outubro de 2018.

SARTORIO, S. D.- Aplicações de Técnicas de Análise Multivariada em Experimentos Agropecuários Utilizando o Software R, Piracicaba: USP 2008. Disponível em: https://www.teses.usp.br/teses/disponiveis/11/11134/tde-06082008-172655/publico/simone.pdf. Acesso em: 10 de fevereiro de 2019.

SILVA JÚNIOR, E. F, Uma Análise Multivariada do sucesso ou fracasso em Matemática dos alunos do 8 ano do Ensino Fundamental, Juazeiro: UNIVASF 2014. Disponível em: https://sca.profmat-sbm.org.br/sca_v2/get_tcc3.php?id=406. Acesso em: 10 de outubro de 2018.

QUINTAL, G. M. d. C. C, Análise de clusters aplicada ao Sucesso/Insucesso em Matemática. Funchal: Universidade da Madeira 2006. Disponível em: https://digituma.uma.pt/handle/10400.13/224. Acesso em: 10 de fevereiro de 2019.

VICINI,L.;SOUZA,A.M, Análise Multivariada da Teoria à Prática, Santa Maria: UFSM, CCNE, 2005.

VITAL, V. C. V, Introdução ao uso do software R para as Ciências Biológicas 2015. Disponível em: https://cantinhodor.files.wordpress.com/2015/03/introduc3a7c3a3o-ao-software-r-para-biologia-marcos-vital-ufal-marc3a7o-2015.pdf. Acesso em 15 de maio de 2019.

Apêndices

APÊNDICE A

Análise de Agrupamentos utilizando o Software R

O R é um software estatístico livre, ou seja, não tem custos e seu código fonte está acessível para qualquer usuário. Ele pode ser obtido em seu site oficial: http://www.r-project.org/. O Rstudio é uma interface do R mais "amigável", pois apresenta diversas funcionalidades aparentes, pode ser obtido no site: http://www.rstudio.com/.

Esse software apresenta diversos manuais, assim como *scripts* (são comandos utilizados para a análise) disponíveis para as mais diversas áreas de pesquisa. Dentre suas peculiaridades, podemos citar que ele distingue letras maiúsculas e minúsculas, as linhas seguidas de "#"não são consideradas como comando (são utilizadas para explicações ou lembretes), para executar os comandos, utilize "ctrl + enter". A seguir apresentamos os dois *scripts* utilizados no decorrer deste trabalho, estes podem ser copiados e colados diretamente no console do R.

Script 1: Agrupamento hierárquico e o dendrograma do exemplo 3.1

```
# quantidade de linhas e colunas. 1:5
# para limpar a memória do R de dados utilizados anteriormente.
rm(list = ls())
exemplo3.1 <- data.frame(Dados=c("A","B","C","D","E"),
V1=c(1,4,5,9,11), row.names = 1)
# para visualizar a tabela de dados no console.
exemplo3.1
# para calcular a matriz de distâncias utilizando a distância Euclidiana.
```

d<-dist(exemplo3.1 ,method="euclidean")

para visualizar a matriz de distâncias.

d

para fazer os agrupamentos dos objetos utilizando a menor distância (single).

fit<-hclust(d,method="single")

fit

construir dendrograma.

plot(fit,main="Dendrograma Exemplo 1", xlab="Variáveis",

ylab ="Distância", sub="Distância Euclidiana/Menor Distância", cex=1, hang=-1)

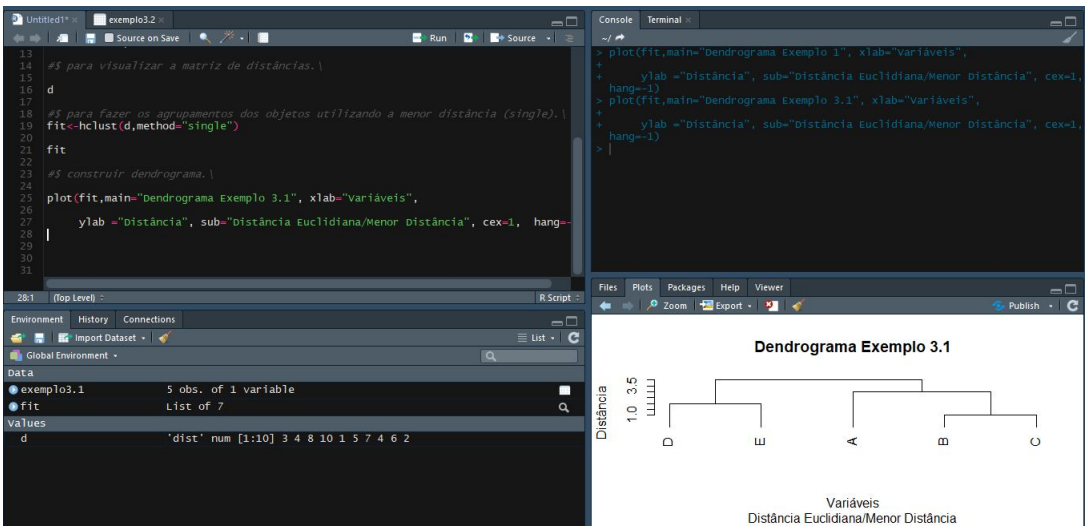


Figura 19 – Imagem do Software RStudio para o exemplo 3.1.

Script 2: Agrupamento não hierárquico, método de k-médias, do exemplo 4.1.

rm(list = ls())

exemplo4.1 <- data.frame(Dados=c("A","B","C","D"), V1=c(2,5,1,8), V2 =c(0,2,4,4), row.names = 1)

exemplo4.1

clust <- kmeans(exemplo4.1, 2)

clust

clust[]

construir a representação do k-médias

plot(exemplo4.1, col = clust$cluster) points(clust$centers, col = 1:2, main="K-médias

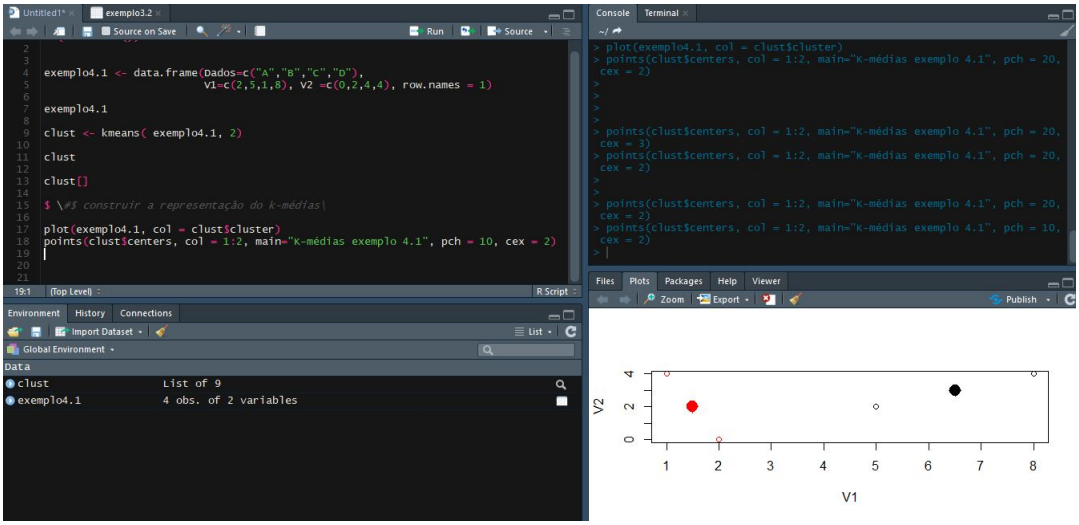


Figura 20 – Imagem do Software RStudio para o exemplo 4.1.

exemplo 4.1", pch = 20, cex = 2)