UNIVERSIDADE FEDERAL DO TOCANTINS
CAMPUS DE ARAGUAÍNA
CURSO DE LICENCIATURA EM MATEMÁTICA

MARIA CRISTINA CORDEIRO SOUSA

UMA ANÁLISE DO ALGORITMO *K-MEANS* COMO INTRODUÇÃO AO APRENDIZADO DE MÁQUINAS

ARAGUAÍNA - TO
2019

MARIA CRISTINA CORDEIRO SOUSA

UMA ANÁLISE DO ALGORITMO *K-MEANS* COMO INTRODUÇÃO AO APRENDIZADO DE MÁQUINAS

Monografia apresentada ao curso de Licenciatura em Matemática da Universidade Federal do Tocantins, como requisito parcial para a obtenção de título de Licenciado em Matemática.
Orientador: Prof. Dr. Alvaro Julio Yucra Hancco.

ARAGUAÍNA - TO
2019

MARIA CRISTINA CORDEIRO SOUSA

UMA ANÁLISE DO ALGORITMO *K-MEANS* COMO INTRODUÇÃO AO APRENDIZADO DE MÁQUINAS

Monografia apresentada ao curso de Licenciatura em Matemática da Universidade Federal do Tocantins, como requisito parcial para a obtenção de título de Licenciado em Matemática.
Orientador: Prof. Dr. Alvaro Julio Yucra Hancco.

Aprovada em: 11 / 12 / 2019.

BANCA EXAMINADORA

Prof. Dr. Alvaro Julio Yucra Hancco (orientador)

Prof. Dr. José Carlos de Oliveira Junior

Prof. Me. André Luiz Ortiz da Silva

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da Universidade Federal do Tocantins

S725a Sousa, Maria Cristina Cordeiro .

Uma análise do algoritmo K-means como introdução ao Aprendizado de Máquinas . / Maria Cristina Cordeiro Sousa. – Araguaína, TO, 2019.

74 f.

Monografia Graduação - Universidade Federal do Tocantins – Câmpus Universitário de Araguaína - Curso de Matemática, 2019.

Orientador: Alvaro Julio Yucra Hancco

1. Otimização. 2. K-means. 3. Clustering. 4. Aprendizado de Máquinas. I. Título

CDD 510

Elaborado pelo sistema de geração automática de ficha catalográfica da UFT com os dados fornecidos pelo(a) autor(a).

Dedico este trabalho à minha família e a todos aqueles que amam matemática.

AGRADECIMENTOS

Agradeço primeiramente a Deus por me conceder saúde, coragem, persistência e sabedoria para conduzir meus planos durante toda a minha vida, por todo o discernimento e por ter colocado as pessoas certas no meu caminho. É com a permissão d'Ele que esse sonho está se tornando realidade.

À minha mãe, Maria Dalvany, devo-lhe tudo. A ela e ao meu pai Edilson, agradeço por terem me educado, por todo apoio e incentivo em busca de um futuro melhor. À minha irmã, Raquel, por todos os momentos que me incentivou e acreditou na minha capacidade, que me acompanhou e me deu forças para continuar. Aos meus tios Edilton e Creuza por toda a preocupação, credibilidade, pelas ajudas financeiras e por todas as demonstrações de carinho.

Aos meus colegas de curso, que de uma forma ou outra contribuíram para que eu chegasse até aqui, pelos momentos bons e ruins que passamos juntos, pelos corujões da vida e pelo apoio quando precisei. Aos amigos que consegui nestes 3 anos, meu muito obrigada pelos grandes momentos juntos, consideração, diversão, apoio e carinho. Ao grupo Euklidea, que vai ficar eternizado no meu coração, de cada momento juntos, sonhos almejados e cada evento do Seminário de Coisas Legais que organizamos. Ao meu grande amigo Marcio, por todo apoio e noites em claro me acompanhando, por todas as vezes que me fez acreditar ainda mais na minha capacidade, e que tudo valeria a pena.

Aos grandes professores que tive o prazer de ser aluna durante o curso, destaco o Dr. José Carlos de Oliveira Junior, Dr. Alvaro Julio Yucra Hancco, Dra. Samara Leandro, Dr. Sinval de Oliveira e Dr. Matheus Lobo, os quais contribuíram bastante para com a pessoa que sou hoje, e com a profissional que serei daqui a alguns dias. Pelo incentivo na continuação dos estudos e o apoio nas realizações acadêmicas.

Um agradecimento mais que especial ao meu orientador Dr. Alvaro Julio Yucra Hancco, pelo compartilhamento de conhecimento para comigo, por todo apoio, encorajamento do início ao fim na construção deste trabalho, por todo tempo disposto, por todos os momentos em que se preocupou e me fez acreditar que conseguiria, teve paciência e compreensão. Além de um grande professor, és também um ser humano excepcional.

As demais pessoas que torcem pelo meu sucesso, que mandaram energias positivas e que contribuíram de alguma forma, e ainda aos projetos Residência Pedagógica e Estendendo o Conhecimento, meu muito obrigada!

O futuro a Deus pertence.

Autor Desconhecido

RESUMO

Este trabalho tem como objetivo analisar a convergência do método *K-means*, um algoritmo de aprendizado não supervisionado que agrupa n dados em k-*clusters*. Neste sentido, apresentamos algumas das vantagens e desvantagens do método *K-means*, comparando o agrupamento original e a clusterização feita pelo algoritmo. Também, apresentamos a aplicação do algoritmo em dois conjuntos de dados: o câncer de mama e diabetes, analisando a clusterização feita pelo *K-means* assim como os padrões e regularidades presentes nos *clusters*. Dessa forma, buscamos apresentar um estudo introdutório da teoria do Aprendizado de Máquina, que busca fazer com que as máquinas realizem tarefas sem que sejam instruídas o tempo todo, partindo apenas de algumas instruções iniciais. Especificamente, procuramos compreender algumas de suas definições e características que permitirão identificar o tipo de aprendizado estudado.

Palavras-chave: Otimização. *K-means*. *Clustering*. Aprendizado de Máquina.

ABSTRACT

This work aims to analyze the convergence of the K-means method, an unsupervised learning algorithm that groups n data into k-clusters. In this sense, we presented some of the advantages and disadvantages of the K-means method, comparing the original clustering and the clustering done by the algorithm. Also, we presented the application of the algorithm in two data sets: breast cancer and diabetes, analyzing the clustering done by K-means as well as the patterns and regularities present in the clusters. In this way, we seek to present an introductory study of Machine Learning theory, which seeks to make machines perform tasks without being instructed all the time, starting only from some initial instructions. Specifically, we seek to understand some of its definitions and characteristics that will allow identifying the type of learning studied.

Keywords: Optimization. K-means. Clustering. Machine Learning.

Lista de Figuras

Sumário

Capítulo 1

Introdução

Os grandes avanços tecnológicos e também no que diz respeito a aplicativos de busca, detecção e armazenamento de informações, vídeos e imagens, produzem enormes quantidades de dados diariamente.A predominância no uso de dispositivos de identificação por rádio frequência ou *transponders* tem duas facetas, por um lado são mecanismos baratos e de pequeno porte, mas por outro necessita de milhões de sensores e estes transmitem dados regularmente, câmeras de vigilância, *e-mails*, páginas da Web geram novos dados a cada segundo. Por estes e outros meios, acredita-se que o universo digital tenha consumido mais de 280 exabytes(EB) no ano de 2007, dez vezes este tamanho em 2011, e projeta-se que cheguemos a 40 zettabytes(ZB) em 2020, o que equivale a $4 \cdot 10^{22}$ bytes.

Devido ao grande volume e diversidade de dados, são necessárias cada vez mais "ferramentas" e meios para se analisar, compreender, processar e resumir estes dados. Existem duas técnicas de análise de dados, a saber, exploratória ou descritiva e inferencial ou preditiva, na primeira significa que não temos hipóteses sobre nossos dados, o objetivo é entender as características deles, já no segundo, queremos confirmar as hipóteses e suposições do nosso conjunto de dados, e fazer previsões a partir disso. Em uma forma de analisar ou outra, o que buscamos mesmo é reconhecimento de padrões para construirmos modelos e fazer previsões de comportamento dos dados. Essas tarefas são chamadas de aprendizagem, e são distinguidas em supervisionada, que trabalha com problemas de classificação em que os dados são rotulados com categorias conhecidas e não supervisionada que envolve problemas de agrupamento. Neste trabalho, consideramos apenas dados diversos e que não conhecemos suas características comuns.

Na aprendizagem não supervisionada, o objetivo é descobrir o padrão do agrupamento natural de um conjunto de pontos, objetos, população ou de qualquer outro conjunto de dados; esta análise chama-se *clusterização* ou análise de *cluster*. Assim como em outras análises, no agrupamento de dados também se faz necessário o desenvolvimento de algoritmos que descubra natureza dos dados, que realize a seguinte tarefa: dado n objetos, encontre k grupos com

determinada medida de semelhança, entre os objetos do mesmo grupo que seja alto grau de semelhança, e baixo grau de semelhança entre os objetos de grupos diferentes, ou seja, os pontos que estiverem próximos são semelhantes entre si, e são diferentes de outros pontos contidos em outro grupo. Dizemos que quanto mais compactados os pontos de um grupo estiverem e mais isolados os grupos forem, mais ideal será nosso *cluster*. Diante da produção de tantos dados, não se tem dúvida da importância e utilização dos agrupamentos, desde a segmentação de imagem em problemas da visão computacional, passando por agrupamento de clientes para um marketing eficiente, até o estudo de dados do genoma na biologia. São três os propósitos de se agrupar para se: obter informações e características importantes sobre os dados, identificar semelhanças, ou como método de organizar e resumir os dados, os quais chamamos de estrutura subjacente, classificação natural e compactação, respectivamente. A metodologia de agrupamento ou clusterização tem sido usada e desenvolvida por taxinomistas, psicólogos, engenheiros, biólogos, cientistas sociais e outros, mas há registros de que o agrupamento de dados tenha sido comentado pela primeira vez em um artigo de 1954 que tratava de dados antropológicos.

Dentre os algoritmos de *clustering* um dos pioneiros e mais conhecido é o *K-means*, que embora tenha mais de 50 anos de surgimento ainda é um dos algoritmos mais procurados dada a sua simplicidade na implementação com certo grau de sucesso em suas aplicações.

Estruturamos essa monografia em 6 capítulos, dispostos da seguinte maneira:

No Capítulo 2, apresentamos algumas noções básicas, definições e proposições de análise convexa, otimização e norma, tais como, conjuntos, poliedros e funções convexas, soluções básicas e básicas admissíveis, métrica euclidiana, quadrado da distância euclidiana, entre outros, que servirá de base para compreensão deste trabalho.

No Capítulo 3, abordamos algumas definições, aplicações, características e classificações que serão importantes para introduzir a teoria do Aprendizado de Máquinas, que será aplicada ao método do qual dissertaremos, *K-means*, um algoritmo de aprendizagem não supervisionada de *clustering*.

No Capítulo 4, explanamos os principais lemas, proposições, propriedades e teoremas que garantem a convergência do algoritmo *K-means* para uma solução ótima. Apresentamos também alguns exemplos com o funcionamento do algoritmo, destacando algumas vantagens e desvantagens do método e para quais tipos de dados o algoritmo se torna mais eficiente.

No Capítulo 5, apresentamos duas aplicações do método *K-means* nas quais analisamos os dados do câncer de mama e diabetes que foram clusterizados a partir da implementação do algoritmo no programa MATLAB.

Recomenda-se para maior compreensão deste trabalho, que o leitor tenha algum conhecimento prévio de Álgebra Linear, Análise Convexa e Otimização.

Capítulo 2

Noções Básicas

Neste capítulo, abordamos algumas definições e resultados de Análise Convexa, de Otimização e Norma, tais como, conjuntos convexos, poliedros convexos, solução ótima, solução básica e básica admissível, métrica euclidiana, entre outros, que servirão de base para compreensão da análise de convergência do algoritmo *K-means*. O desenvolvimento teórico deste capítulo está baseado nas referências [10, 12].

2.1 Análise Convexa

Nesta seção, apresentamos os conceitos e resultados elementares da Análise Convexa, necessários para compreensão deste trabalho.

Definição 2.1 (*Combinação Afim)*. Sejam $x_i \in \mathbb{R}^n$ e $\alpha_i \in \mathbb{R}$ com $i = 1, 2, \ldots, n$. Uma combinação afim de $x_1, x_2, \ldots, x_n \in \mathbb{R}^n$ é uma combinação linear dos pontos x_i,

$$\sum_{i=1}^{n} \alpha_i x_i = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_n x_n,$$

em que $\sum_{i=1}^{n} \alpha_i = 1$.

Exemplo 2.2. Sejam x_1, x_2 e x_3 quaisquer elementos de $\mathbb{R}^n$. Então

$$0.9x_1 - 0.2x_2 + 0.3x_3$$

é uma combinação afim de x_1, x_2, x_3.

Definição 2.3 (*Combinação Convexa)*. Dados $x_i \in \mathbb{R}^n$ e $\alpha_i \in [0, 1]$ com $i = 1, 2, \ldots, n$. Tais que $\sum_{i=1}^{n} \alpha_i = 1$, o ponto $\sum_{i=1}^{n} \alpha_i x_i$ é denominado combinação convexa dos pontos $x_i \in \mathbb{R}^n$ com parâmetros α_i.

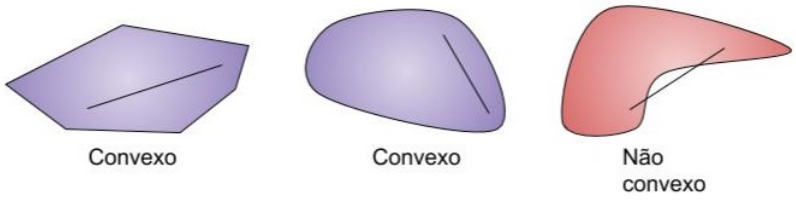
Exemplo 2.4. Sejam x_1, x_2 e x_3 quaisquer elementos de $\mathbb{R}^n$. Então

$$0.7x_1 + 0.2x_2 + 0.1x_3,$$

é uma combinação convexa de x_1, x_2, x_3.

Definição 2.5 (*Conjunto Convexo)***.** Consideremos o espaço vetorial euclidiano $\mathbb{R}^n$. Um conjunto $C \subset \mathbb{R}^n$ é convexo se o segmento de reta entre dois pontos quaisquer do conjunto estiver contido em C, isto é, para quaisquer $x_1, x_2 \in C$, a combinação convexa dada por $\alpha x_1 + (1-\alpha)x_2 \in C$, com $0 \leqslant \alpha \leqslant 1$.

Figura 2.1: Conjunto Convexo e não Convexo



Fonte: Arquivo Pessoal

Teorema 2.6. *Um conjunto $C \subset \mathbb{R}^n$ é convexo se, e somente se, para quaisquer $m \in \mathbb{N}$ pontos $x^j \in C$ e coeficientes $\alpha_j \in [0,1]$, $j = 1, 2, \ldots, m$, tais que $\sum_{j=1}^{m} \alpha_j = 1$, a combinação convexa $\sum_{j=1}^{m} \alpha_j x^j$ pertença a C.*

Demonstração: Vamos supor que $C \subset \mathbb{R}^n$ seja um conjunto convexo, então temos que provar que para quaisquer $m \in \mathbb{N}$, $x^j \in C$ e $\alpha_j \in [0,1]$ tais que $\sum_{j=1}^{m} \alpha_j = 1$, vale que $x = \sum_{j=1}^{m} \alpha_j x^j \in C$.

Por indução,

(i) Se $m = 1$, então $\alpha_1 = 1$ e $x = 1 \cdot x^1 \in C$.

(ii) Suponhamos que vale para $m = n$, e mostremos que vale para $m = n + 1$.

Temos

$$\sum_{j=1}^{n+1} \alpha_j = 1,$$

então

$$1 - \sum_{j=1}^{n} \alpha_j = \alpha_{n+1}.$$

Analisando, temos

- Se $\alpha_{n+1} = 1$ então $\alpha_j = 0$, para todo $j = 1, 2, \ldots, m$.

  Logo,

$$x = \sum_{j=1}^{m} 0 \cdot x^j + 1 \cdot x^{n+1} = x^{n+1} \in C.$$

  Agora,

- Se $\alpha_{n+1} \in [0, 1)$, temos $(1 - \alpha_{n+1}) > 0$.

  Logo,

$$\begin{aligned} x = \sum_{j=1}^{n+1} \alpha_j x^j &= \sum_{j=1}^{n} \alpha_j x^j + \alpha_{n+1}\, x^{n+1} \\ &= (1 - \alpha_{n+1}) \sum_{j=1}^{n} \frac{\alpha_j}{(1 - \alpha_{n+1})} x^j + \alpha_{n+1}\, x^{n+1}. \end{aligned} \tag{2.1}$$

  Então, basta que tomemos

$$z = \sum_{j=1}^{n+1} \beta_j x^j \text{ onde, } \beta_j = \frac{\alpha_j}{(1 - \alpha_{n+1})} \leq 1,\ j = 1, 2, \ldots, m.$$

  Como $\sum_{j=1}^{n+1} \alpha_j = 1$, segue que $1 - \alpha_{n+1} = \sum_{j=1}^{n} \alpha_j$,

então

$$\begin{aligned} \sum_{j=1}^{n} \beta_j &= \frac{1}{1 - \alpha_{n+1}} \cdot \sum_{j=1}^{n} \alpha_j \\ &= \frac{1}{(1 - \alpha_{n+1})} (1 - \alpha_{n+1}) \\ &= 1. \end{aligned}$$

  Logo, por hipótese de indução $z \in C$.

  Substituindo z em (2.1), obtemos

$$x = (1 - \alpha_{n+1}) z + \alpha_{n+1} x^{n+1},$$

que garante pela convexidade de C, que $x = \sum_{j=1}^{n+1} \alpha_j x^j \in C$, como queríamos.

Na implicação contrária, suponhamos que $x = \sum_{j=1}^{m} \alpha_j x^j \in C$, para todo $m \in \mathbb{N}$ com

$$x^j \in C \text{ e } \alpha_j \in [0,1] \text{ tais que } \sum_{j=1}^{m} \alpha_j = 1.$$

Então, particularizando para $m = 2$, temos

$$\begin{aligned} \sum_{j=1}^{2} \alpha_j x^j &= \alpha_1 x^1 + \alpha_2 x^2 \in C, \text{ com } x^1, x^2 \in C \text{ e } \alpha_1, \alpha_2 \in [0,1], \text{ tais que} \\ \alpha_1 + \alpha_2 &= 1. \end{aligned}$$

Assim, $\alpha_1 = 1 - \alpha_2$, então,

$$(1-\alpha_2)x^1 + \alpha_2 x^2 = \alpha_1 x^1 + \alpha_2 x^2 \in C.$$

Portanto, pela Definição 2.5, C é um conjunto convexo. □

Propriedade 1. *A interseção finita de conjuntos convexos é convexa.*

Demonstração: Consideremos S_j, $j = 1, 2, \ldots, n$, conjuntos convexos. Seja $S = \bigcap_{j=1}^{n} S_j$ e considere $x_1, x_2 \in S$, pela definição de S, temos $x_1, x_2 \in S_j$, para todo $j = 1, 2, \ldots, n$. Como cada S_j é convexo, vale que

$$\alpha x_1 + (1-\alpha)x_2 \in S_j, \text{ para todo } \alpha \in [0,1].$$

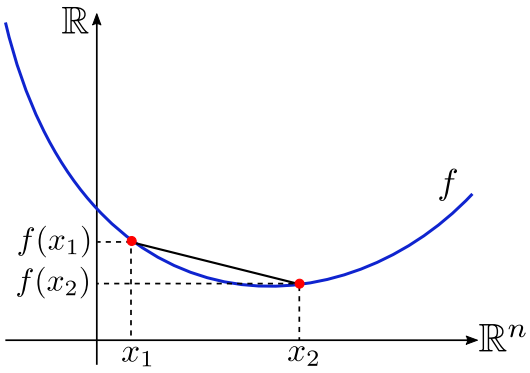
Portanto, $\alpha x_1 + (1-\alpha)x_2 \in S$, para todo $\alpha \in [0,1]$, que significa que S é convexo. □

Definição 2.7 (*Função Convexa*)**.** Considere um conjunto convexo C. Uma função $f : C \to \mathbb{R}$ é denominada convexa, se para todo $x_1, x_2 \in C$ e todo $\alpha \in [0,1]$

$$f(\alpha x_1 + (1-\alpha)x_2) \leqslant \alpha f(x_1) + (1-\alpha) f(x_2),$$

ou seja, o segmento de reta que une os pontos $(x_1,\ f(x_1))$ e $(x_2,\ f(x_2))$ está sempre acima ou sobre o gráfico da função.

Figura 2.2: Função Convexa



Fonte: Arquivo Pessoal

Exemplo 2.8. Seja $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^2$ e $\alpha \in [0, 1]$. Mostraremos que $f(x) = x^2$ é convexa.

Resolução:

Consideremos $x, y \in \mathbb{R}$ quaisquer. Segue que,

$$\begin{aligned} 0 &\leqslant (y-x)^2 \\ 0 &\leqslant y^2 - 2xy + x^2 \\ 2xy - x^2 &\leqslant y^2 \end{aligned}$$

como $\alpha \in [0, 1]$, então $\alpha(1-\alpha) \geqslant 0$. Assim,

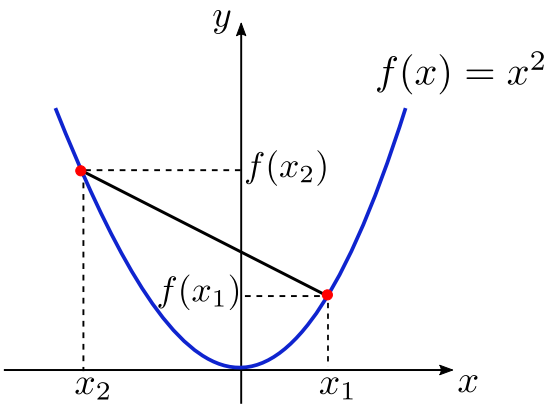
$$\begin{aligned} 2x\alpha(1-\alpha)y - \alpha(1-\alpha)x^2 &\leqslant (1-\alpha)y^2\alpha \\ 2x\alpha(1-\alpha)y + (\alpha^2-\alpha)x^2 &\leqslant (1-\alpha)y^2\alpha \\ 2x\alpha(1-\alpha)y + (\alpha^2-\alpha)x^2 &\leqslant (1-\alpha)y^2\left[1-(1-\alpha)\right] \\ 2x\alpha(1-\alpha)y + (\alpha^2-\alpha)x^2 &\leqslant (1-\alpha)y^2 - (1-\alpha)^2y^2 \\ (1-\alpha)^2y^2 + 2x\alpha(1-\alpha)y + (\alpha^2-\alpha)x^2 &\leqslant (1-\alpha)y^2 \\ (1-\alpha)^2y^2 + 2x\alpha(1-\alpha)y + \alpha^2x^2 &\leqslant (1-\alpha)y^2 + \alpha x^2 \\ \alpha^2x^2 + 2x\alpha(1-\alpha)y + (1-\alpha)^2y^2 &\leqslant \alpha x^2 + (1-\alpha)y^2 \\ \big(\alpha x + (1-\alpha)y\big)^2 &\leqslant \alpha x^2 + (1-\alpha)y^2 \\ f\big(\alpha x + (1-\alpha)y\big)^2 &\leqslant \alpha f(x) + (1-\alpha)f(y). \end{aligned}$$

Logo, $f(x) = x^2$ é convexa.

Definição 2.9. Considere um conjunto convexo C. Uma função $f : C \to \mathbb{R}$ é estritamente convexa, se $\forall x_1, x_2 \in C$ e $\forall \alpha \in [0, 1]$

$$f(\alpha x_1 + (1-\alpha)x_2) < \alpha f(x_1) + (1-\alpha)f(x_2).$$

Figura 2.3: Função $f(x) = x^2$



Fonte: Arquivo Pessoal

Exemplo 2.10. Seja $f : (0, \infty) \to \mathbb{R}$, $f(x) = \dfrac{1}{x}$ e $\alpha \in [0, 1]$. Vamos mostrar que $f(x)$ é estritamente convexa.

Resolução:

Sabemos que $0 \leqslant \alpha \leqslant 1 \;\Rightarrow\; 0 \leqslant \alpha - \alpha^2$. Sejam $x, y \in \mathbb{R}^+ \setminus \{0\}$ quaisquer.
Supondo $x < y \;\Rightarrow\; (x-y)^2 > 0$, temos

$$\begin{aligned} xy &< (\alpha - \alpha^2)(x-y)^2 + xy \\ xy &< \alpha^2(y-x)(x-y) + \alpha x(x-y) - \alpha x(x-y) + xy \\ xy &< [\alpha(y-x) + x]\,[\alpha(x-y) + y] \end{aligned}$$

$$\frac{1}{\alpha(x-y)+y} < \frac{\alpha(y-x)+x}{xy}$$

$$f(\alpha(x-y)+y) < \alpha\left(\frac{y-x}{xy}\right) + \frac{1}{y}$$

$$f(\alpha(x-y)+y) < \left(\frac{1}{x} - \frac{1}{y}\right) + \frac{1}{y}$$

$$\begin{aligned} f(\alpha(x-y)+y) &< \alpha\big(f(x) - f(y)\big) + f(y) \\ f(\alpha x + (1-\alpha)y) &< \alpha f(x) + (1-\alpha) f(y). \end{aligned}$$

Como $\alpha \in [0, 1]$ é qualquer, logo, $f(x) = \dfrac{1}{x}$ é estritamente convexa.

Definição 2.11 (*Função Côncava*)**.** Considere um conjunto convexo C. Uma função $f : C \to \mathbb{R}$ é côncava, se para todo $x_1, x_2 \in C$ e todo $\alpha \in [0, 1]$

$$f(\alpha x_1 + (1-\alpha)x_2) \geqslant \alpha f(x_1) + (1-\alpha) f(x_2),$$

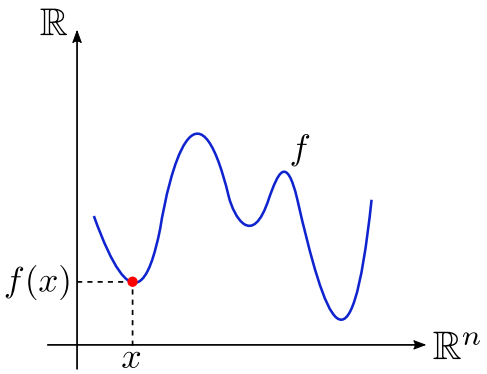
ou vale simplesmente que, f é côncava se $-f$ for convexa.

Definição 2.12. Uma função $f : \mathbb{R}^n \to \mathbb{R}$ é linear se, para todo $\alpha_1, \alpha_2 \in \mathbb{R}$ e para x_1, x_2 pertencentes ao domínio da função, vale que

$$f(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 f(x_1) + \alpha_2 f(x_2).$$

Definição 2.13. Dada uma função $f : \mathbb{R}^n \to \mathbb{R}$. O ponto x é um *mínimo local* de f, se $f(x) \leqslant f(x_i)$ para todo x_i em alguma vizinhança de x.

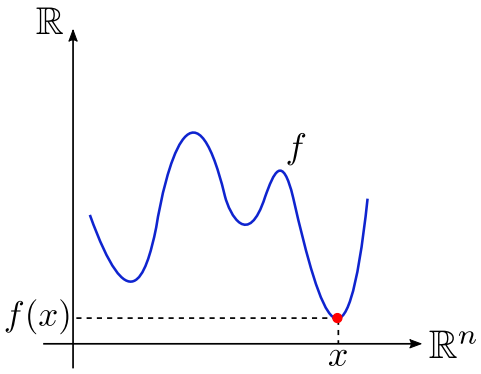
Figura 2.4: Mínimo Local



Fonte: Arquivo Pessoal

Definição 2.14. Seja $f : \mathbb{R}^n \to \mathbb{R}$. O ponto x é o *mínimo global* de f se $f(x) \leqslant f(x_i)$ para todo $x_i \in \mathbb{R}^n$.

Figura 2.5: Mínimo Global



Fonte: Arquivo Pessoal

Definição 2.15. Um poliedro $S \subset \mathbb{R}^n$ é o conjunto solução de um número finito de igualdades e desigualdades lineares,

$$\begin{cases} Ax \geqslant b \\ x \geqslant 0, \end{cases}$$

em que, $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ e $x \in \mathbb{R}^n$, com cada coordenada $x_i \geqslant 0$, $i = 1, 2, \ldots, n$.

Proposição 1. Todo poliedro é um conjunto convexo.

Demonstração: Consideramos dois elementos x_1, x_2 de um poliedro descrito na forma $Ax \geqslant b$, com $x \in \mathbb{R}^n$, A uma matriz de ordem $(m \times n)$ e b um vetor em $\mathbb{R}^m$. Então $Ax_1 \geqslant b$ e $Ax_2 \geqslant b$. Assim, para todo $\alpha \in [0, 1]$, temos

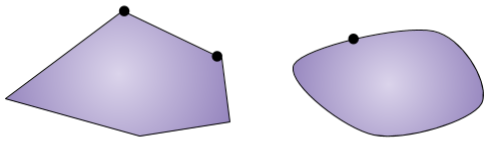
$$\begin{aligned} A\left(\alpha x_1 + (1-\alpha)x_2\right) &= \alpha Ax_1 + (1-\alpha)Ax_2 \\ &\geqslant \alpha b + (1-\alpha)b \\ &= b. \end{aligned}$$

Portanto, $\alpha x_1 + (1-\alpha)x_2$ também pertence ao poliedro, que implica que todo poliedro é convexo. □

Definição 2.16 (*Ponto Extremo)*. Seja S um poliedro. Um vetor $x \in S$ é um ponto extremo de S se podemos encontrar dois vetores $x_1, x_2 \in S$, diferentes de x, e um escalar $\alpha \in [0, 1]$ tais que $x = \alpha x_1 + (1-\alpha)x_2$.

Exemplo 2.17. Ilustramos alguns dos pontos extremos nos conjuntos a seguir:

Figura 2.6: Pontos Extremos



Fonte: Arquivo Pessoal

2.2 Otimização

Nesta seção, apresentamos alguns conceitos sobre problemas de otimização e existência de soluções. Também apresentamos algumas definições importantes para compreensão do trabalho. Todas as definições e exemplos aqui abordadas, servem tanto para problemas de minimização quanto para problemas de maximização, uma vez que minimizar uma função é equivalente a maximizar sua simétrica, ou seja, $\min f = -\max g$, em que $f = -g$.

Um problema de otimização apresenta-se da seguinte forma

$$\min \qquad f(x_1, x_2, \ldots, x_n) = c_1x_1 + c_2x_2 + \ldots + c_nx_n$$

s.a [1]

$$\begin{array}{ccc} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n & \{\leqslant ; \geqslant ; =\} & b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n & \{\leqslant ; \geqslant ; =\} & b_2 \\ \vdots & & \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n & \{\leqslant ; \geqslant ; =\} & b_m \\ x_1, x_2, \ldots, x_n & \{\leqslant ; \geqslant\} & 0. \end{array} \tag{2.2}$$

E admite-se que um problema de otimização possa ter restrições de variáveis não negativas $(x_i \geqslant 0)$, não positivas $(x_i \leqslant 0)$ ou ainda variáveis livres $(x_i \in \mathbb{R})$ e consiste em determinar uma solução ou um conjunto de soluções para uma determinada função ou conjunto de funções. São do tipo maximização e/ou minimização.

De (2.2), a função linear de minimização

$$f(x_1, x_2, \ldots, x_n) = c_1x_1 + c_2x_2 + \ldots + c_nx_n,$$

é chamada de função objetivo, da qual se pretende encontrar a solução ótima, definimos a seguir:

Definição 2.18 (*Função Objetivo).* Diz respeito à função da qual se deseja minimizar ou maximizar, que está sujeita a uma série de restrições que poderão ser de três tipos $\{\leqslant ; \geqslant ; =\}$.

[1] sujeito a

Definição 2.19 (*Solução Ótima)*. A solução que satisfaz as restrições do problema e que dá o maior valor possível à função objetivo, quando o problema é de maximização, ou aquela que dá o menor valor, se o problema é de minimização.

Observação 2.20. O conceito de solução ótima é específico do problema que se deseja otimizar. Um problema de otimização pode ter uma única solução ou um conjunto de soluções ou ainda não haver solução que satisfaça todas as restrições. A solução ótima da função objetivo de f relaciona-se facilmente com g (simétrica da função f).

Para facilitar a resolução e encontrar a solução ótima, é necessário que o problema esteja apresentado na sua forma mais simples, ou seja, quando todas as suas restrições são de igualdade e suas variáveis são não negativas. Para tal processo, usamos novas variáveis para normalizar as restrições nas desigualdades do tipo "$\leqslant$"e do tipo "$\geqslant$", respectivamente, e as definimos a seguir.

Definição 2.21 (*Variável de Folga)*. Consiste em uma nova variável $x_s \geqslant 0$, que se adiciona em restrições com desigualdades não positivas, esta nova variável aparece com coeficiente zero na função objetivo e, somando na equação correspondente, ou seja,

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \leqslant b_1 \quad \longrightarrow \quad & a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n + x_s = b_1 \\ & x_s \geqslant 0. \end{aligned}$$

Definição 2.22 (*Variável de Excesso)*. No caso de restrições com desigualdades não negativas, deve-se acrescentar também uma nova variável chamada de variável de excesso $x_s \geqslant 0$. Esta nova variável também aparece com coeficiente zero na função objetivo e, subtraindo na equação correspondente, ou seja, se tivermos

$$\begin{aligned} a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \geqslant b_2 \quad \longrightarrow \quad & a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n - x_s = b_2 \\ & x_s \geqslant 0. \end{aligned}$$

Este processo é importante para encontrarmos as soluções básicas e as básicas admissíveis de um problema de otimização, uma vez que encontrar uma solução ótima para o problema não é tão simples. Isto motiva o estudo da seguinte subseção.

2.2.1 Solução Básica: Solução Admissível

Consideremos a seguir, um problema de minimização qualquer, cujo as variáveis são não negativas,

$$\begin{aligned} \min \quad & f \\ s.a \quad & Ax = b \\ & x \geqslant 0, \end{aligned}$$

em que $Ax = b$ é o conjunto de restrições de um problema, onde A tem m equações e n variáveis, com $m < n$. Separamos a matriz A em submatrizes B e N. Se a matriz B do sistema for inversível, a solução é bem determinada, caso contrário, B será formada pelas m-colunas linearmente independentes da matriz A, logo possui inversa, e N estará formada pelas colunas restantes. Destacamos que as variáveis associadas a essas submatrizes serão fixas, ou seja,

$$Ax = b \quad \Leftrightarrow \quad Bx_B + Nx_N = b. \tag{2.3}$$

Logo,

$$\begin{aligned} Ax = b \quad &\Leftrightarrow \quad [B \mid N] \cdot \begin{bmatrix} x_B \\ x_N \end{bmatrix} = b \\ &\Leftrightarrow \quad Bx_B + Nx_N = b. \end{aligned}$$

Apenas aplicando a inversa e isolando x_B, temos

$$x_B = B^{-1}b - B^{-1}Nx_N,$$

nessa expressão de x_B é conhecida como solução geral do sistema.

Através dessa análise, conseguimos introduzir os seguintes termos:

- Uma *partição básica* de A, ou seja, $A = [B \mid N]$, em que a matriz $B_{m\times m}$ é uma matriz básica de A e possui inversa, e $N_{m\times(n-m)}$ é a matriz não básica composta pelas colunas restantes de A.
- As variáveis x_B são chamadas de *variáveis básicas*, ou seja, são as m variáveis que compõem a solução básica do problema.
- E também x_N, chamadas de*variáveis não básicas*, que são as variáveis que não compõem a solução básica do problema, e valem, obrigatoriamente, zero.

Considerando a partição básica $A = [B \mid N]$ da matriz A, uma solução é dita básica quando:

$$\begin{cases} x_B &= B^{-1}b \\ x_N &= 0 \end{cases} \tag{2.4}$$

Observação 2.23. Solução básica existe se, e somente se, as colunas das restrições de igualdade correspondentes às m variáveis básicas são linearmente independentes, isto é, formam uma base.

Nesse sentido, se o problema de otimização possui uma solução básica, definida por (2.4), então, é possível estudar um tipo especial de solução básica, que definimos a seguir.

Definição 2.24 (*Solução Básica Admissível*)**.** É uma solução básica em que todas as variáveis básicas são também não-negativas, ou seja, $x_B = B^{-1}b \geqslant 0$.

Se alguma variável da solução básica for não positiva, então a solução é dita *Solução Básica não Admissível*.

Exemplo 2.25. Dadas as restrições de um problema de otimização qualquer,

$$\begin{aligned} x_1 + x_2 &\leqslant 9 \\ x_1 - x_2 &\leqslant 3 \\ 5x_1 - x_2 &\geqslant 4 \\ x_1, x_2 &\geqslant 0\,, \end{aligned} \tag{2.5}$$

vamos encontrar as soluções básica e básica admissível desse problema.

Resolução:

Primeiramente, fazemos o processo de padronização das variáveis, usando variáveis de folga e de excesso como nas Definições 2.21 e 2.22,

$$\begin{aligned} x_1 + x_2 + x_3 &= 9 \\ x_1 - x_2 + x_4 &= 3 \\ 5x_1 - x_2 - x_5 &= 4 \\ x_1, x_2, x_3, x_4, x_5 &\geqslant 0. \end{aligned} \tag{2.6}$$

Agora, as restrições do problema, se tornam em um sistema do tipo $Ax = b$, ou seja, podemos escrevê-lo

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 \\ 5 & -1 & 0 & 0 & -1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 9 \\ 3 \\ 4 \end{bmatrix}$$

onde,

$$A = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 \\ 5 & -1 & 0 & 0 & -1 \end{bmatrix},\ x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} \text{ e } b = \begin{bmatrix} 9 \\ 3 \\ 4 \end{bmatrix}.$$

Agora podemos obter uma partição básica de A, e as restrições básicas e não básicas

associadas a ela,

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ 5 & -1 & -1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_5 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 9 \\ 3 \\ 4 \end{bmatrix},$$

em que,

$$B = \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ 5 & -1 & -1 \end{bmatrix}, \; x_B = \begin{bmatrix} x_1 \\ x_2 \\ x_5 \end{bmatrix}, \; N = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \; x_N = \begin{bmatrix} x_3 \\ x_4 \end{bmatrix} \text{ e } b = \begin{bmatrix} 9 \\ 3 \\ 4 \end{bmatrix}.$$

Daqui, para encontrarmos as soluções básicas de (2.6), basta que encontremos a matriz B^{-1}, que fazendo as devidas substituições e operações

$$B^{-1} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & -\frac{1}{2} & 0 \\ 2 & 3 & -1 \end{bmatrix}.$$

De (2.4),

$$x_B = \begin{bmatrix} x_1 \\ x_2 \\ x_5 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & -\frac{1}{2} & 0 \\ 2 & 3 & -1 \end{bmatrix} \cdot \begin{bmatrix} 9 \\ 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 6 \\ 3 \\ 23 \end{bmatrix},$$

$$x_N = \begin{bmatrix} x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Logo, x_B é solução básica, e como cada uma de suas componentes é não negativa, ela se torna uma solução básica admissível do Problema de Otimização, dado em (2.5). Considerando que a solução de (2.6) permite encontrar a solução de (2.5).

2.3 Distância

Nesta seção, apresentamos algumas noções básicas de distâncias e propriedades determinadas por uma métrica. Especificamente, vamos estudar a métrica induzida por uma norma.

2.3.1 Norma

Dado $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, definimos a *Norma Euclidiana* da seguinte forma:

Definição 2.26. A norma euclidiana associa um número real a cada vetor $x \in \mathbb{R}^n$, e podemos

interpretar geometricamente $\|x\|_2$ como o comprimento do vetor x.

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Teorema 2.27 (Desigualdade de Cauchy-Schwarz). *Para cada $x, y \in \mathbb{R}^n$, temos*

$$|x \cdot y| \leqslant |x| \cdot |y|,$$

com igualdade se, e somente se $x = \alpha y$, em que $\alpha \in \mathbb{R}$ (ou trivialmente $x = 0$ ou $y = 0$).

Demonstração: A demonstração desse Teorema pode ser encontrada em [12], p. 6. □

A norma euclidiana satisfaz as seguintes propriedades, para todo $x, y \in \mathbb{R}^n$ e $\alpha \in \mathbb{R}$,

(i) *Positiva:* $\|x\|_2 \geqslant 0$, $\|x\|_2 = 0 \Leftrightarrow x = 0$.
Por definição,

$$\begin{aligned} \|x\|_2 = \sqrt{x \cdot x} \geqslant 0 \text{ e } \|x\|_2 = 0 &\Leftrightarrow 0 = \|x\|_2^2 = x \cdot x \\ &\Leftrightarrow x = 0. \end{aligned}$$

(ii) *Homogênea:* $\|\alpha x\|_2 = |\alpha| \cdot \|x\|_2$.
Uma vez que,

$$\begin{aligned} \|\alpha x\|_2 = \sqrt{\alpha x \cdot \alpha x} &= \sqrt{\alpha^2 x \cdot x} \\ &= |\alpha| \cdot \sqrt{x \cdot x} \\ &= |\alpha| \cdot \|x\|_2. \end{aligned}$$

(iii) Desigualdade Triangular: $\|x + y\|_2 \leqslant \|x\|_2 + \|y\|_2$.
Analisemos que

$$\|x + y\|_2^2 = (x + y) \cdot (x + y) = \|x\|_2^2 + 2x \cdot y + \|y\|_2^2$$

Pelo Teorema 2.27,

$$\begin{aligned} \|x\|_2^2 + 2\|xy\|_2 + \|y\|_2^2 &\leqslant \|x\|_2^2 + 2\|x\|_2\|y\|_2 + \|y\|_2^2 \\ &= (\|x\|_2 + \|y\|)_2^2, \end{aligned}$$

logo, $\|x + y\|_2 \leqslant \|x\|_2 + \|y\|_2$.

Exemplo 2.28. Seja o vetor $y = (9, 3, 5, 5, 1, 7, -1, 4) \in \mathbb{R}^8$, vamos calcular a norma euclidiana $\|y\|_2$.

Resolução:

$$\begin{aligned}\|y\|_2 &= \sqrt{9^2+3^2+5^2+5^2+1^2+7^2+(-1)^2+4^2} \\ &= \sqrt{81+9+25+25+1+49+1+16} \\ &= \sqrt{196+11} \\ &\leqslant \sqrt{196}+\sqrt{11} \\ &= 14+\sqrt{11}.\end{aligned}$$

A Norma Euclidiana não é a única norma possível em $\mathbb{R}^n$, também temos a Norma da Soma e Norma do Máximo, que definimos a seguir.

Definição 2.29 (*Norma da Soma)*. Dado um vetor $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ a norma da soma do vetor x é dada por:

$$\|x\|_s = \sum_{j=1}^{n} |x_j|.$$

Exemplo 2.30. Considere o vetor $z = (-8, 3, -5, 5, 1, -3, -1, 2, -4) \in \mathbb{R}^9$, devemos calcular a distância da soma $\|z\|_s$.

Resolução:

$$\begin{aligned}\|z\|_s &= |-8|+|3|+|-5|+|5|+|1|+|-3|+|-1|+|2|+|-4| \\ &= 8+3+5+5+1+3+1+2+4 \\ &= 32.\end{aligned}$$

Definição 2.31 (*Norma do Máximo)*. Dado um vetor $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ a norma do máximo do vetor x é dada por:

$$\|x\|_m = \max_j\{|x_j|\}$$

Exemplo 2.32. Dado o vetor $w = (3, 7, -2, 6, -9, 12, 3, 1, -1) \in \mathbb{R}^9$, calcular a norma do máximo $\|w\|_m$.

Resolução:

$$\begin{aligned}\|w\|_m &= \max\{|3|,|7|,|-2|,|6|,|-9|,|12|,|3|,|1|,|-1|\} \\ &= \max\{3,7,2,6,9,12,3,1,1\} \\ &= 12.\end{aligned}$$

Observação 2.33. As Normas da Soma e do Máximo, satisfazem as mesmas propriedades da Norma Euclidiana, isto é, as propriedades positivas, homogênea e a desigualdade triangular.

Definição 2.34. Um espaço vetorial que tenha uma norma chama-se *espaço vetorial normado*, ou simplesmente *espaço normado*.

2.3.2 Métrica

Uma métrica é uma função que permite medir distâncias. Se $(W, \| \ \|)$ é um espaço normado, então a norma $\| \ \|$ induz uma métrica em W.

Para todo $u, v \in W$,

$$d_{\| \ \|}(u, v) = \|u - v\|.$$

A norma euclidiana, induz em $\mathbb{R}^n$ a *métrica euclidiana*.

Definição 2.35 (*Métrica Euclidiana*)**.** É uma função $d_2 : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, em que para todo $x = (x_1, x_2, \ldots, x_n), y = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$, temos

$$d_2(x, y) = \|x - y\|_2 = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2}.$$

Esta métrica satisfaz as seguintes propriedades:

(i) *Positiva:* $x, y \in \mathbb{R}^n$, $d_2(x, y) \geqslant 0$ e $d_2(x, y) = 0 \Leftrightarrow x = y$.

Pois,

$$\|x\| = 0 \Leftrightarrow x = 0.$$

(ii) *Simétrica:* $x, y \in \mathbb{R}^n$, $d_2(x, y) = d_2(y, x)$

Já que, se $\alpha \in \mathbb{R}$, temos $\|\alpha x\|_2 = |\alpha| \cdot \|x\|_2$, assim $\| - x\|_2 = \|x\|_2$.

Logo,

$$\begin{aligned} d_2(x, y) &= \|x - y\|_2 \\ &= \| - (x - y)\|_2 \\ &= d_2(y, x). \end{aligned}$$

(iii) *Desigualdade Triangular:* $x, y \in \mathbb{R}^n$, $d_2(x, z) \leqslant d_2(x, y) + d_2(y, z)$.

Uma vez que

$$\begin{aligned} d_2(x, z) &= \|x - z\|_2 \\ &= \|x - y + y - z\|_2 \\ &\leqslant \|x - y\|_2 + \|y - z\|_2 \\ &= d_2(x, y) + d_2(y, z) \end{aligned}$$

Dessa forma, qualquer métrica induzida por uma norma satisfaz as propriedades acima, e isso é fácil de mostrar, para tal definimos as métricas da soma e do máximo.

Definição 2.36 (*Métrica da Soma*)**.** Sejam os vetores $x = (x_1, x_2, \ldots, x_n), y = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$, vale que

$$d_s(x, y) = \|x - y\|_s = |x_1 - y_1| + |x_2 - y_2| + \ldots + |x_n - y_n|.$$

Definição 2.37 (*Métrica do Máximo).* Consideremos $x = (x_1, x_2, \ldots, x_n), y = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$, temos

$$d_m(x, y) = \|x - y\|_m = \max_j\{|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_n - y_n|\}.$$

Percebemos que se mudarmos a métrica mudam-se as distâncias, vejamos o próximo exemplo.

Exemplo 2.38. Consideremos a distância da origem até o ponto $x \in \mathbb{R}^6$, onde $x = (3, 5, 4, 1, 2, -3) \in \mathbb{R}^6$, calcular d_2, d_s e d_m.

Resolução:

- Na métrica euclidiana:
$$\begin{aligned} d_2(x, 0) = \|x\|_2 &= \sqrt{3^2 + 5^2 + 4^2 + 1^2 + 2^2 + (-3)^2} \\ &= 8. \end{aligned}$$

- Na métrica da soma:
$$\begin{aligned} d_s(x, 0) = \|x\|_s &= |3| + |5| + |4| + |1| + |2| + |-3| \\ &= 18. \end{aligned}$$

- Na métrica do máximo:
$$\begin{aligned} d_m(x, 0) = \|x\|_m &= \max_j\{|3|, |5|, |4|, |1|, |2|, |-3|\} \\ &= 5. \end{aligned}$$

Teorema 2.39 (Teorema da Equivalência). *Sejam d_2, d_s e d_m, respectivamente, as métricas euclidiana, da soma e do máximo em $\mathbb{R}^n$. Então, para cada vetor $x, y \in \mathbb{R}^n$, temos*

$$d_m(x, y) \leqslant d_2(x, y) \leqslant d_s(x, y) \leqslant n \cdot d_m(x, y).$$

Demonstração: Dados dois vetores $x = (x_1, x_2, \ldots, x_n),\ y = (y_1, y_2, \ldots, y_n) \in \mathbb{R}^n$, temos que

$$\begin{aligned} d_m(x, y) &= \max_j\{|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_n - y_n|\} \\ &\leqslant \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2} \\ &= d_2(x, y). \end{aligned}$$

Daqui,

$$d_2(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2},$$

Pela desigualdade triangular, resulta

$$\begin{aligned} d_2(x, y) &\leqslant |x_1 - y_1| + |x_2 - y_2| + \ldots + |x_n - y_n| \\ &= d_s(x, y), \end{aligned}$$

elevando os dois membros da desigualdade ao quadrado, temos

$$(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2 \leqslant (|x_1 - y_1| + |x_2 - y_2| + \ldots + |x_n - y_n|)^2$$

ou seja, todo o membro também é positivo.

Da última desigualdade, segue que

$$\begin{aligned} d_s(x,y) &= |x_1 - y_1| + |x_2 - y_2| + \ldots + |x_n - y_n| \\ &\leqslant n \cdot \max_j \{|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_n - y_n|\} \\ &= n \cdot d_m(x,y). \end{aligned}$$

□

Exemplo 2.40. Considerando ainda o Exemplo 2.38, percebemos que o vetor $x = (3,5,4,1,2,-3) \in \mathbb{R}^6$ verifica o teorema anterior, pois

$$d_2(x,0) = 8 \quad d_s(x,0) = 18 \quad d_m(x,0) = 5$$

$$\begin{array}{ccccccc} d_m(x,0) & < & d_2(x,0) & < & d_s(x,0) & < n \cdot d_m(x,0) \\ 5 & < & 8 & < & 18 & < 6 \cdot 5 = 30. \end{array}$$

No desenvolvimento deste trabalho, para simplificação de cálculos e operações, usaremos o quadrado da distância euclidiana, que também verifica as propriedades de métrica.

Definição 2.41. *(Quadrado da Distância Euclidiana):*

$$d(x,y) = \|x - y\|_2^2 = \sum_{j=1}^{n} (x_j - y_j)^2.$$

Exemplo 2.42. Dados os vetores $x = (1,5,10,-3,-1), y = (0,5,7,-5,-6) \in \mathbb{R}^5$, caculamos o quadrado da distância euclidiana entre os vetores x e y.

Resolução:

$$\begin{aligned} d(x,y) &= (1-0)^2 + (5-5)^2 + (10-7)^2 + (-3-(-5))^2 + (-1-(-6))^2 \\ &= (1-0)^2 + (5-5)^2 + (10-7)^2 + (-3+5)^2 + (-1+6)^2 \\ &= (1)^2 + (0)^2 + (3)^2 + (2)^2 + (5)^2 \\ &= 1 + 0 + 9 + 4 + 25 \\ &= 39. \end{aligned}$$

Capítulo 3

Introdução ao Aprendizado de Máquinas

Neste capítulo, iremos apresentar uma noção básica de Aprendizado de Máquinas(AM) e alguns conceitos e definições importantes. O objetivo desse não é abordar um estudo teórico da área, com análise de processos de aprendizagens, discussões dos anseios da sociedade para com ela, e os significativos avanços nas pesquisas, e sim de apenas contextualizar o leitor e responder algumas perguntas das quais fazemos ao nos depararmos com tal conceito, como por exemplo: O que seria mesmo o (AM)? Quais suas características? Onde são aplicados? O que podemos concluir de alguns de seus algoritmos? Entre outras. Tentaremos responder essas perguntas no decorrer deste capítulo. A elaboração deste capítulo foi baseado nas referências [1, 13]

3.1 Algumas Definições

Definição 3.1 (*Algoritmo)***.** Uma sequência com números finitos de instruções que permitem chegar à solução de um determinado problema é denominada algoritmo.

Definição 3.2 (*Banco de Dados)***.** São conjuntos organizados de dados que se relacionam e formam uma informação, que estão relacionados entre si, por exemplo, registros sobre pessoas, lugares ou coisas.

Definição 3.3 (*Cluster)***.** Faz referência a um agrupamento de dados, aglomeração, também denominado instâncias.

Definição 3.4 (*Mineiração)***.** Consiste na procura de um modelo simples e de grande importância a partir do processamento de uma enorme quantidade de dados, tornando-os em um modelo simples de grande utilidade.

Definição 3.5 (*Clustering)***.** Técnica de mineração de dados multivariados, que os agrupa em *clusters.*

Definição 3.6 (*Reconhecimento de Padrões*)**.** Visa analisar agrupamentos de dados baseados em informações anteriores, buscando padrões e regularidades nos dados *clusterizados*.

Definição 3.7 (*Extração de Conhecimento*)**.** Processo que a máquina faz ao aprender uma regra a partir de um conjunto de dados, sendo regra um modelo simples que caracteriza os dados.

Definição 3.8 (*Modelo Preditivo*)**.** É uma função que, aplicada a um conjunto de dados, consegue identificar padrões ocultos e fazer previsões do que poderá ocorrer com dados futuros.

Definição 3.9 (*Modelo Descritivo*)**.** Consiste encontrar grupos de dados que compartilhem a mesma característica.

Definição 3.10 (*Algoritmos hierárquicos*)**.** Consistem em iniciar o problema com todos os dados em um único *cluster*, e divide repetidamente cada *cluster* em grupos menores, impondo uma hierarquia de dados.

Definição 3.11 (*Algoritmos Particionais*)**.** Nos algoritmos particionais todos os *clusters* são encontrados ao mesmo tempo com os dados já particionados.

3.2 Uma ideia sobre Aprendizado de Máquinas

Figura 3.1: Aprendizado de Máquinas

Fonte: Depositphots(2019) [1]

[1]Disponível em: https://br.depositphotos.com/186535438/stock-illustration-neural-network-deep-learning-artificial.html;. Acesso em dez. 2019.

A crescente utilização de meios eletrônicos em nossas vidas tem gerado um crescimento exponencial na produção de dados, além de *Websites* que rastreiam tudo que seus usuários pesquisam, registros de localização e velocidades são feitas pelos celulares o tempo todo. Empresas de grande e pequeno porte, sejam públicas ou privadas, trabalham também com sistema de gerenciamento de banco de dados, ou seja, diariamente são produzidos uma enorme quantidade de dados, desde o código de barras dos seus produtos até os sensores colocados no estabelecimento, sobre seus clientes, empregados, produtos, e assim muitos dados são produzidos, ficando inviável que estes dados sejam analisados por seres humanos, de modo que, tais atividades estão ficando cada vez mais automatizadas e, por isso, necessitam de ferramentas eficientes que auxiliem na análise dessa grande quantidade de dados produzidos e que os tornem em algo útil.

Por tais motivos, vemos progressivamente o avanço da ciência de dados e da computação em nossas vidas, que se preocupa em armazenar e processar enormes quantidades de dados, e ainda na inferência e em ter algoritmos eficientes para resolver problemas de otimização. No entanto, para realizar algumas tarefas não existe um algoritmo que as descreva, apesar de ter muitos dados que as exemplifique. Para isso, almejava-se que um computador/máquina ao rodar vários dados, conseguisse extrair automaticamente um algoritmo para tais tarefas, acreditando que embora não pudesse descrever o processo, daria para identificar certas regularidades, construir uma aproximação útil e usá-las para fazer previsões.

Pesquisadores de Inteligência Artificial, cientistas de dados e da computação, e outros, trabalharam por muito tempo para o desenvolvimento de uma área hoje denominada Aprendizado de Máquinas, que fosse um sistema capaz de identificar *e-mails* de *spam*, de prever quem são os possíveis compradores para determinado produto, de resolver problemas de visão e reconhecimento de fala. Por exemplo, para nós, seres humanos, reconhecer o rosto das pessoas pessoalmente ou em fotos, de perfil ou de frente, perto ou um pouco mais distante, é uma tarefa fácil e que não precisa de muitos esforços, fazemos de maneira inconsciente, de modo que não conseguimos explicar, por isso é impossível que um programa saiba fazer tal atividade por meio de um algoritmo, uma vez que o rosto de uma pessoa é particular dela, tem uma estrutura, possui olhos, boca, nariz com formas e localizações específicas do rosto de cada um.

Mas, analisando amostras de imagens do rosto de uma pessoa, um sistema de Aprendizado de Máquinas, consegue capturar um padrão para essa pessoa, e reconhecê-la em outras imagens. Além disso, o sistema consegue, mesmo quando exposto a um grande volume de dados, processa-los e construir um modelo simples, é o que chamamos de mineração de dados. O (AM) é uma ramificação da Inteligência Artificial, que vinha sendo almejada a alguns anos, nasceu da curiosidade de pesquisadores da área de que seria possível as máquinas aprenderem a partir de reconhecimento de padrões de dados, porque até então as máquinas só faziam atividades das quais elas eram instruídas constantemente. Além de resolver problemas de banco de dados, é capaz de se adaptar ao ambiente, pois o sistema aprende com processamentos anteri-

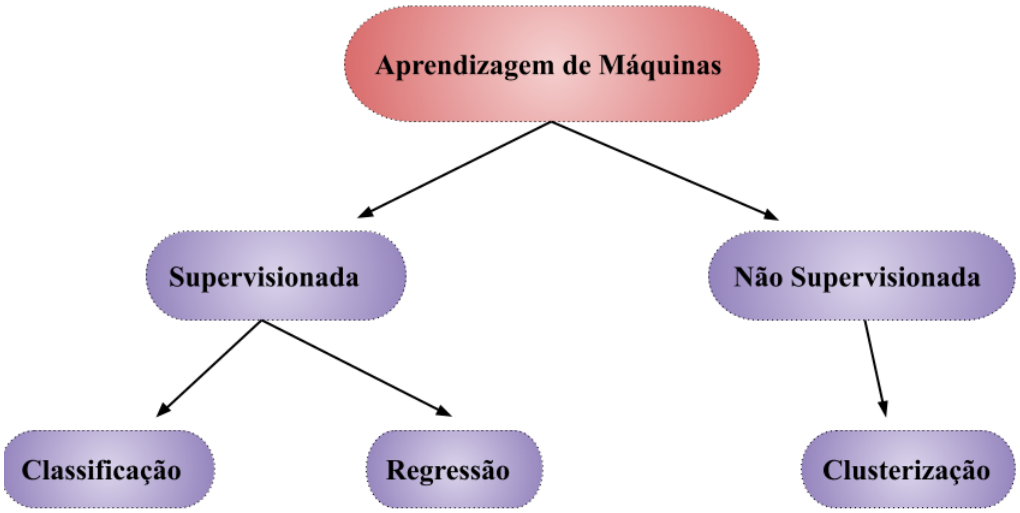
ores e se adaptam livremente quando expostos a novos dados, capazes de produzir resultados confiáveis e sem repetições. O objetivo principal ao se desenvolver um sistema do tipo é fazer com que a máquina seja capaz de adquirir conhecimento, aprenda com dados anteriores com o mínimo de intervenção humana, quase de forma automática.

Nós seres humanos estamos em constante construção e evolução do conhecimento que são adquiridos. Por exemplo, quando aprendemos a melhor estratégia em uma situação, nós guardamos isso na memória e a utilizamos quando nos deparamos com esta mesma situação. No Aprendizado de Máquinas o processo de aprendizagem se dá de forma parecida, pois o programa precisa apenas de algumas instruções iniciais, a partir disso geram seu próprio conhecimento conforme os resultados que for obtendo. Porém, diferente do que a ciência cognitiva e a neurociência em nós seres humanos explica, segundo Ethem(2004, p. 4): " o objetivo do Aprendizado de Máquinas não é entender os processos subjacentes a aprendizagem em humanos e animais, mas construir sistemas úteis, como em qualquer domínio da engenharia".

O uso aprendizagem de máquina já abrange vários campos do nosso cotidiano e são utilizadas para solucionar os mais diversos problemas, é utilizada também para construir sistemas de *streaming* como os da *Netflix* e *Spotify* que recomendam filmes e músicas aos seus usuários com base no seu histórico de acesso e seus favoritos, ou quando o *Google Maps* sugere rotas alternativas para um mesmo local, com ela é possível fazer identificação de mudanças climáticas, categorizar páginas da *Web* conforme o gênero, marcar mensagens de *e-mail* como *spam* ou até mesmo fazer deduções sobre a bolsa de valores, reconhecimento de voz, predição de taxa de curas de pacientes com diferentes doenças, detecção de uso fraudulento de cartões de crédito, condução de automóveis de forma autônoma em rodovias, diagnóstico de câncer por meio da análise de dados de expressão gênica. São diversos os métodos de (AM), que são utilizados em múltiplas tarefas, cada um com suas especificidades, organizadas de acordo com diferentes critérios, um deles é quanto a sua categoria de aprendizado, como mostra a seção seguinte.

3.3 Tipos de Aprendizagem

Figura 3.2: Tipos de Aprendizagem



Fonte: Arquivo Pessoal

3.3.1 Aprendizagem Supervisionada

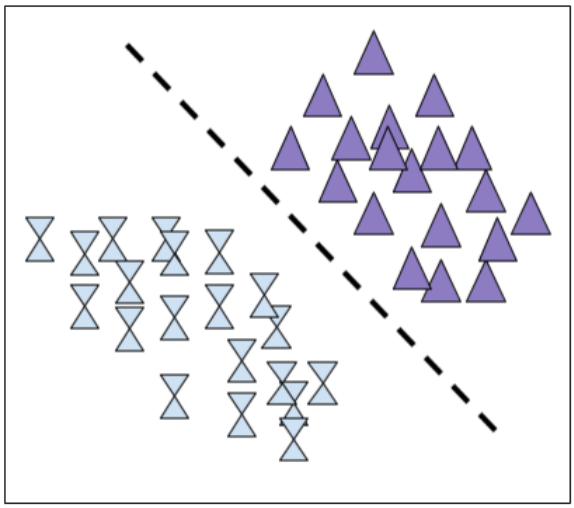
No Aprendizado Supervisionado temos problemas onde há uma entrada e uma saída bem definida, tendo ideia que existe uma relação entre eles. Ou seja, temos os dados com quais queremos analisar e temos o que procurar, a tarefa é aprender o mapeamento desse percurso. Para tal, há uma fase denominada treinamento, em que alguns dados são inseridos no sistema, e o objetivo dele é encontrar parâmetros que se ajuste a dados desconhecidos do conjunto de teste, ou seja, esse tipo de aprendizado tem como finalidade prever o resultado dado um conjunto de amostras de treinamento, junto com seus rótulos de treinamento. Se subdivide em método de classificação e regressão.

Classificação

Os problemas de classificação são problemas em que os dados da entrada já são previamente constituídos, e a tarefa de um algoritmo classificador é atribuir esses dados em determinadas classes. Por exemplo, importante que um banco seja capaz de prever o risco em empréstimos, se tem grande probabilidade de ter um *default* por parte do cliente, mas também não pode incomodar seu cliente, ou até mesmo deixá-lo constrangido com perguntas sobre sua situação financeira. Em situações como essa, o banco reúne um conjunto de dados de cada cliente, usando seu histórico de empréstimos anteriores e outros atributos que definem um cliente

e seu risco. Um sistema de (AM) ajusta um modelo baseado nos dados, que o banco usará para classificar seus clientes automaticamente, em clientes de alto e baixo risco. Este seria um exemplo de problema de classificação, que define clientes de alto e baixo risco para um sistema bancário.

Figura 3.3: Classificação



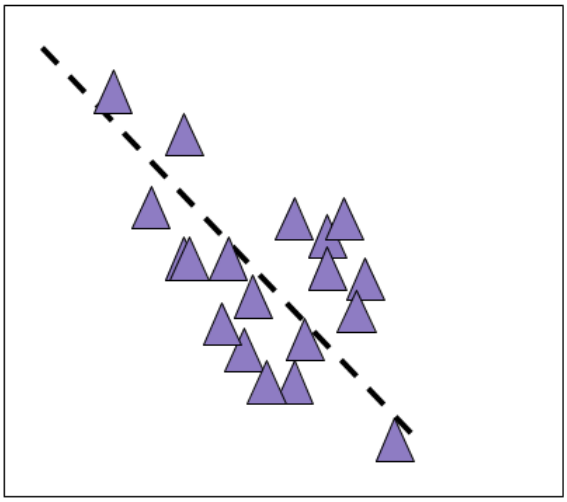
Fonte: Arquivo Pessoal

Após o treinamento do sistema com um determinado conjunto de dados é possível que ele aprenda uma regra de classificação, tendo essa regra, o principal objetivo é fazer previsões corretas sobre dados novos, via uma função discreta.

Regressão

Temos uma situação em que queremos comprar um carro usado, o que fazemos na maioria das vezes é entrar em um site ou aplicativo para pesquisar a tabela de preço para aquele modelo de carro em que estou procurando. O que não sabemos é que esses aplicativos em que inserimos dados para pesquisar o preço do carro são programados a partir de um sistema de (AM), em que a entrada do programa são os atributos do carro, e a saída é exatamente o preço tabelado para aquele modelo, que pode variar entre um preço x e um preço y. Esses problemas em que a saída é um número e os dados de saída e entrada que se relacionam via uma função contínua, são chamados de problemas de regressão.

Figura 3.4: Regressão



Fonte: Arquivo Pessoal

3.3.2 Aprendizagem não Supervisionada

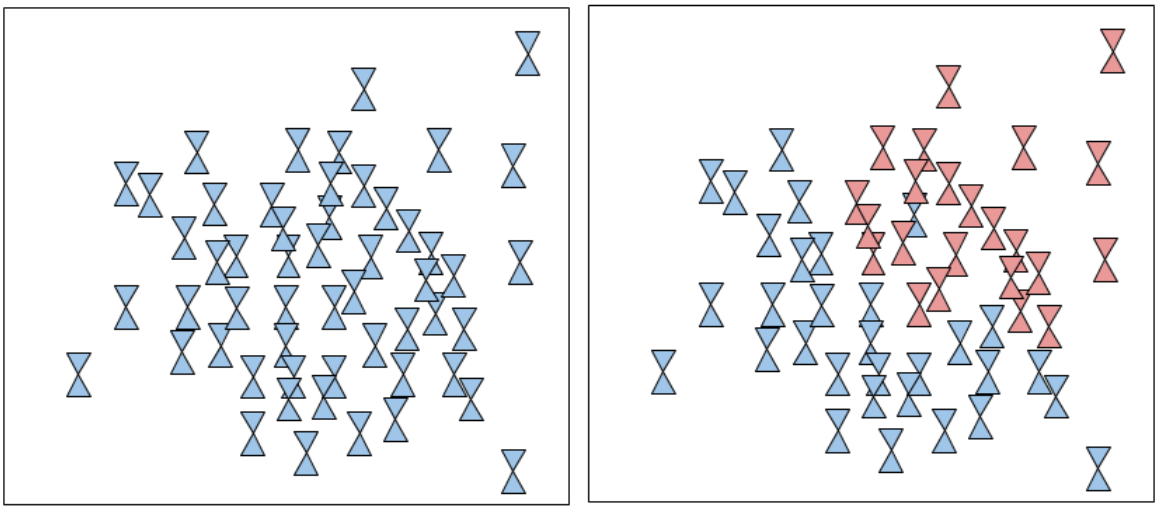
Diferentemente da Aprendizagem Supervisionada em que temos entradas e saídas definidas por um supervisor, aqui na Aprendizagem não Supervisionada como o nome sugere, não teremos um supervisor para definir a saída do nosso modelo, conheceremos apenas os dados de entrada. O objetivo principal aqui é justamente encontrar os padrões, regularidades ou estruturas existentes no nosso conjunto inicial de dados , ou seja na entrada, e assim analisar a partir do funcionamento do programa, o que aconteceu e não aconteceu com os nossos dados. Nesse tipo de aprendizado não se tem ideia (ou pouca) do resultado, não há correções, pois toda a tarefa é realizada segundo semelhanças que a máquina entende.

Clustering

Na aprendizagem não supervisionada existem certos padrões nos dados de entrada, alguns mais frequentes que outros. O que a máquina faz é agrupar esses dados que aparentemente têm algo em comum, essa tarefa consiste em reunir em um *cluster*, dados que são semelhantes entre si, mas que cada *cluster* seja diferente um do outro. Os seus algoritmos são usados para explorar padrões ou criar grupos. A clusterização ou *clustering* é a análise dos *clusters* (dados agrupados), analisando o grau de semelhanças ou diferenças entre eles e seus elementos, que seria uma função distância, em que quanto menor a distância mais semelhantes são os dados analisados. Por exemplo, se nosso objetivo é agrupar relatórios de notícias, então teremos *clusters* de notícias relacionadas a esporte, política, artes, etc, e dentro deles teremos notícias que se aproximem mais de cada eixo. O conceito de clusterização muda de acordo com a área em que está inserido, pode ser chamado de análise Q, tipologia,

agrupamento e taxonomia. Segundo Anil(2009, p. 3) a clusterização e seus algoritmos foram bastante estudados por vários autores, tais como, Han e Kamber(2000) e Bishop(2006), que classificam esses algoritmos em hierárquicos e particionais.

Figura 3.5: *Clustering*



Fonte: Arquivo Pessoal

Dos algoritmos hierárquicos os mais conhecidos são o link único e o link completo, já os de partição, o mais popular é o *K-means*, este tem uma história rica e diversa, e embora já tenha mais de 50 anos de surgimento ainda é um dos algoritmos mais usados para *clustering*, por sua facilidade de implementação, simplicidade, eficiência e sucesso nas suas aplicações, por isto, nos próximos capítulos serão abordados o algoritmo *K-means* e algumas de suas aplicações.

Capítulo 4

Método de *K-means*

O *K-means* é um algoritmo de aprendizagem não supervisionado e de clusterização, utilizado para particionar dados em k agrupamentos distintos. Ele agrupa dados que compartilham características importantes e parecidas. De modo empírico, uma boa solução para o processo de clusterização é aquela em que os dados do grupo sejam mais semelhantes entre si, do que comparados com outro grupo.

Neste capítulo, abordamos as principais definições, propriedades, teoremas, lemas e proposições para compreendermos matematicamente como funciona, e como realmente o algoritmo *K-means* converge para uma solução ótima. Também apresentaremos o funcionamento do algoritmo, mostrando alguns exemplos e destacaremos algumas vantagens e desvantagens. A implementação do algoritmo é possível em vários *softwares*, tais como, C++, Python, entre outros. Mas neste trabalho, usamos o *software* MATLAB versão $R2019b$, este que é um *software* voltado para cálculo numérico e destinado a fazer cálculos com matrizes, onde seus problemas e solução são escritos na forma matemática, e nos permitem ver como diferentes algoritmos funcionam, executando iterações até que os resultados desejados sejam obtidos. Os conteúdos deste capítulo se encontram baseados nas referências [1, 8].

4.1 Definições Iniciais

Pela enorme quantidade e diversidade de dados produzidos diariamente, se faz necessário cada vez mais a criação de ferramentas que facilitem e permitam a mineração desses dados. Um dos métodos que nos auxiliam na mineração de dados multivariados é o agrupamento ou *clustering*, que consiste em minimizar ou maximizar a função objetivo definida em partes, homogenizando os dados dentro do *cluster* e heterogeneizando os *clusters* uns dos outros. Os algoritmos de clusterização se dividem em hierárquicos ou particionais. Neste trabalho, tratamos somete do algoritmo *K-means*, classificado como algoritmo particional. Apresentamos uma análise de sua convergência.

Analisemos agora alguns conceitos e notações. Consideremos um conjunto de dados com n pontos, definido:

$$P = \{x_i\}_{i=1}^{n} ,$$

um agrupamento, particiona esse conjunto P em k subconjuntos

$$C = \{C_1, C_2, ..., C_k\},$$

cada C_i é denominado um *cluster* do conjunto. E ainda para cada C_i associaremos um y_i nomeado centroide, que além de representar o *cluster*, será comparado com todos os outros elementos do mesmo. Existem outras formas de obtê-lo, mas neste trabalho, o determinaremos através da média de todos os pontos pertencentes ao *cluster*,

$$y_i = \frac{1}{n_i} \sum_{x_j \in C_i} x_j, \tag{4.1}$$

onde n_i será o número de elementos de C_i.

Para medir a dispersão dos elementos de um *cluster*, e verificar quão próximos eles estão, compararando a distância dos pontos ao centroide correspondente, usaremos a função de semelhança,

$$F_s(P) = \sum_{i=1}^{k} \sum_{x_j \in C_i} D(x_j, y_i)$$

em que D é a distância euclidiana. Em outras palavras, F_s define a soma de todas as distância entre cada elemento e o centroide do seu *cluster*, isto é, medirá quão bem o centroide representa seu grupo.

4.2 O Algoritmo *K-means*

O *K-means* é caracterizado como um algoritmo guloso, que em suas iterações escolhe o objeto que lhe parece mais "apetitoso", que chama mais sua atenção, torna o objeto parte da solução do problema, entretanto não analisa as consequências de suas escolhas. Em termos mais específicos o algoritmo procura minimizar F_s, convergindo para a solução local, a mais viável, não se preocupando se esta é a solução ótima do problema. O algoritmo consiste nos seguintes passos, ao "encarar" um conjunto de dados.

- Distribui todos os pontos do conjunto P de forma aleatória em k *clusters*.
- Calcula através de (4.1), o centroide de cada *cluster* C_i.
- Associa cada ponto $x_j \in P$ a um *cluster* C_{i^*}, do centroide y_{i^*} mais próximo ao ponto, ou

seja,

$$i^* = \underset{i=1,2,...,k}{\operatorname{argmin}} \|x_j - y_i\|_2^2 \,,$$

significa que um ponto x_j qualquer será agrupado a C_i, quando este ponto possuir a menor distância ao centroide desse *cluster*, comparado com a distância aos outros centroides do resto de *clusters*.

- Após o passo anterior, muitos pontos terão mudado de grupo, por isso faz-se necessário a atualização dos centroides de cada *cluster*, assim, repete-se o 2º passo, onde encontraremos um novo centroide y_i para o *cluster* C_i.

- Os dois últimos passos serão repetidos de forma iterativa, até que os respectivos centroides não mudem mais ou satisfaçam a precisão estabelecida, então esta iteração será o mínimo local do nosso problema.

- E assim como em outros processos iterativos, neste também tem o teste de parada, que é analisado através da soma das diferenças dos centroides da iteração atual pela anterior,

$$\sum_{i=1}^{k} \|y_i^t - y_i^{t-1}\|_2^2 \leqslant \epsilon \,,$$

y_i^t representa o centroide da iteração atual, $\epsilon > 0$ é a precisão determinada e o limite da convergência. Abaixo representa-se o pseudo-código do algoritmo *K-means*.

Algoritmo 1: K-means

Dados: P conjunto dos pontos, k número de clusters, ϵ erro

início
  iteração $\leftarrow 0$
  Inicializar os $y_i, i = \{1, 2, ..., k\}$ (Com pontos aleatórios de P)
  repita
    $C_i \leftarrow \emptyset, \quad \forall j = 1, 2, \ldots k$
    para $x_j \in P$ **faça**
      $i^* \leftarrow \text{argmin}\{\|x_j - y_i^{\textit{iteração}}\|^2\}$
      $C_{i^*} = C_{i^*} \cup \{x_j\}$
    fim
    para $i = 1$ *até* k **faça**
      $y_i^{\textit{iteração}} = \dfrac{1}{|C_i|} \displaystyle\sum_{x_j \in C_i} x_j$
    fim
  até $\displaystyle\sum_{i=1}^{k} \|y_i^{\textit{iteração}} - y_i^{\textit{iteração}-1}\| < \epsilon$;
fim

4.3 Condições de Convergência do Algoritmo

O *K-means* pode ser visto também como um algoritmo que resolve problemas de otimização, em que usamos a $F_s(P)$ para medir a representatividade dos centroides de cada grupo, e desta forma, deduziremos nosso problema a partir dela, que será a nossa função objetivo. Para tal, consideramos Y uma matriz linha dos centroides do conjunto P,

$$Y = [y_1, y_2, ..., y_k]_{d\times k}\,,$$

onde y_i é o centroide do *cluster* C_i, e também W uma matriz real, cujo seu elemento w_{ij} representa o *cluster* i e o ponto j. Em que se $w_{ij} = 1$ significa que o ponto x_j pertence ao *cluster* C_i, ou seja,

$$w_{ij} = 1 \Rightarrow x_j \in C_i,$$

por outro lado se,

$$w_{ij} = 0 \Rightarrow x_j \notin C_i.$$

Denotaremos a matriz das entradas w_{ij} por:

$$W = [w_{ij}]_{k\times n}\,.$$

Assim, consideremos o problema de otimização para o algoritmo *K-means*,

$$\begin{aligned}(O) : min \quad & f(W,Y) = \sum_{i=1}^{k}\sum_{j=1}^{n} w_{ij} D(x_j, y_i) \\ & s.a \sum_{i=1}^{k} w_{ij} = 1, j = 1, 2, ..., n\end{aligned} \tag{4.2}$$

para $i = 1, 2, ..., k$ e $j = 1, 2, ...n$, teremos a variação do elemento w_{ij}.

No problema (4.2) D mede a distância euclidiana do elemento x_j ao centroide y_i, e a definimos por:

$$D(x_j, y_i) = \|x_j - y_i\|_2^2.$$

Fixando a linha i da matriz W, temos

$$W_i = [w_{i1}, w_{i2}, ..., w_{in}]\,,$$

e sobre influência do *cluster* i, a função f, assume a seguinte notação:

$$f_i(W_i, Y_i) = \sum_{j=1}^{n} w_{ij} D(x_j, y_i). \tag{4.3}$$

Ou seja, o centroide y_i também será denotado por Y_i. Por consequência de (4.3), escrevemos a função objetivo do problema (O) de (4.2) como sendo,

$$f(W, Y) = \sum_{i=1}^{k} f_i(W_i, Y_i).$$

Por questões de simplificações usaremos a notação $f_{\hat{W}}(Y)$ para a função $f(\hat{W}, Y)$, quando a matriz $\hat{W}$ for fixada. Para continuar a análise de convergência do algoritmo *cluster* a partir do problema (O), a seguir, apresentamos alguns resultados importantes.

Propriedade 2. *Se a função $D(x, Y_i)$ que mede a distância de qualquer elemento aos centroides Y_i for estritamente convexa, variando no conjunto Y, então $f_{\hat{W}}(Y)$ também é uma função estritamente convexa.*

Demonstração: Consideremos duas matrizes de centroides Y^1 e Y^2 de ordem $d \times k$ e um escalar $\alpha \in [0, 1]$. Por hipótese temos que a função $D(x, Y_i)$, que mede a distância dos pontos x_j aos centroides Y_i^1 e Y_i^2, é estritamente convexa. O que implica

$$D(x_j, \alpha Y_i^1 + (1-\alpha) Y_i^2) \leqslant \alpha D(x_j, Y_i^1) + (1-\alpha) D(x_j, Y_i^2). \tag{4.4}$$

Usando a definição da função objetivo, sabemos que,

$$f_{\hat{W}}(\alpha Y^1 + (1-\alpha) Y^2) = \sum_{i=1}^{k} \sum_{j=1}^{n} w_{ij} D(x_j, \alpha Y_i^1 + (1-\alpha) Y_i^2), \tag{4.5}$$

desse modo, por (4.4)

$$\sum_{i=1}^{k} \sum_{j=1}^{n} w_{ij} D(x_j, \alpha Y_i^1 + (1-\alpha) Y_i^2) < \sum_{i=1}^{k} \sum_{j=1}^{n} w_{ij} \left[\alpha D(x_j, Y_i^1) + (1-\alpha) D(x_j, Y_i^2)\right] \tag{4.6}$$

pela propriedade de somatório,

$$\sum_{i=1}^{k} \sum_{j=1}^{n} w_{ij} \left[\alpha D(x_j, Y_i^1) + (1-\alpha) D(x_j, Y_i^2)\right] =$$
$$\sum_{i=1}^{k} \sum_{j=1}^{n} w_{ij} \left[\alpha D(x_j, Y_i^1)\right] + \sum_{i=1}^{k} \sum_{j=1}^{n} w_{ij} \left[(1-\alpha) D(x_j, Y_i^2)\right].$$

Agora, pela propriedade de multiplicação por um escalar,

$$\alpha \sum_{i=1}^{k}\sum_{j=1}^{n} w_{ij}D(x_j, Y_i^1) + (1-\alpha)\sum_{i=1}^{k}\sum_{j=1}^{n} w_{ij}D(x_j, Y_i^2),$$

mas,

$$\sum_{i=1}^{k}\sum_{j=1}^{n} w_{ij}D(x_j, Y_i^1) = f_{\hat{W}}(Y^1).$$

Portanto, por (4.5) e (4.6) concluímos que

$$f_{\hat{W}}(\alpha Y^1 + (1-\alpha)Y^2) < \alpha f_{\hat{W}}(Y^1) + (1-\alpha) f_{\hat{W}}(Y^2),$$

garantindo que a função $f_{\hat{W}}(Y)$, seja estritamente convexa. □

Precisamos agora provar que de fato a função $D(x, Y_i) = \|x - Y_i\|_2^2$ é estritamente convexa, para continuarmos com a análise da convergência do algoritmo.

Lema 4.1. *A função $D(x, Y_i) = \|x - Y_i\|_2^2$ é estritamente convexa.*

Demonstração: Sejam Y_1 e Y_2 dois elementos quaisquer e $\alpha \in [0, 1]$, então pela definição da função D, temos:

$$D\left[x, (\alpha Y_1 + (1-\alpha)Y_2)\right] = \|x - (\alpha Y_1 - (1-\alpha)Y_2)\|_2^2.$$

Daqui,

$$\begin{aligned} \|x - (\alpha Y_1 - (1-\alpha)Y_2)\|_2^2 &= \|\alpha x - \alpha x + x - (\alpha(Y_1 - (1-\alpha)Y_2\|_2^2 \\ &= \|\alpha x + (1-\alpha)x - \alpha Y_1 - (1-\alpha)Y_2\|_2^2 \end{aligned}$$

colocando α e $(1-\alpha)$ em evidência, resulta

$$\|x - (\alpha Y_1 - (1-\alpha)Y_2\|_2^2 = \|\alpha(x - Y_1) + (1-\alpha)(x - Y_2)\|_2^2.$$

Pela desigualdade triangular, temos

$$\|\alpha(x - Y_1) + (1-\alpha)(x - Y_2)\|_2^2 < \|\alpha(x - Y_1)\|_2^2 + \|(1-\alpha)(x - Y_2)\|_2^2,$$

e como

$$\|\alpha(x - Y_1)\|_2^2 + \|(1-\alpha)(x - Y_2)\|_2^2 = \alpha D(x, Y_1) + (1-\alpha)D(x, Y_2)$$

obtemos,

$$\|\alpha(x - Y_1) + (1 - \alpha)(x - Y_2)\|_2^2 \quad < \quad \alpha D(x, Y_1) + (1 - \alpha)D(x, Y_2),$$

concluindo que $D(x, Y_i)$ é estritamente convexa, com $0 \leqslant \alpha \leqslant 1$. □

Observação 4.2. Do Lema 4.1 concluímos que a função semelhança F_s do *K-means* é estritamente convexa.

Propriedade 3. *Dados $\hat{Y} \in \mathbb{R}^{d\times k}$ e a função $f_{\hat{Y}}(W) = f(W, \hat{Y})$. Então $f_{\hat{Y}}(W)$ é linear.*

Demonstração: Sejam W^1 e W^2 matrizes de ordem $k \times n$, e $\alpha_1, \alpha_2 \in \mathbb{R}$. Por hipótese e do problema (4.2) sabemos que

$$f_{\hat{Y}}(\alpha_1 W^1 + \alpha_2 W^2) = \sum_{i=1}^{k}\sum_{j=1}^{n}(\alpha_1 w_{ij}^1 + \alpha_2 w_{ij}^2)D(x_j, \hat{y}_i),$$

aplicando a propriedade distributiva, resulta

$$f_{\hat{Y}}(\alpha_1 W^1 + \alpha_2 W^2) \quad = \quad \sum_{i=1}^{k}\sum_{j=1}^{n}\left[\alpha_1 w_{ij}^1 D(x_j, \hat{y}_i) + \alpha_2 w_{ij}^2 D(x_j, \hat{y}_i)\right],$$

agora, pela propriedade de somatório,

$$f_{\hat{Y}}(\alpha_1 W^1 + \alpha_2 W^2) \quad = \quad \sum_{i=1}^{k}\sum_{j=1}^{n}\alpha_1 w_{ij}^1 D(x_j, \hat{y}_i) + \sum_{i=1}^{k}\sum_{j=1}^{n}\alpha_2 w_{ij}^2 D(x_j, \hat{y}_i)$$

Logo, por (4.2)

$$\sum_{i=1}^{k}\sum_{j=1}^{n}\alpha_1 w_{ij}^1 D(x_j, \hat{Y}_i) = \alpha_1 f_{\hat{Y}}(W^1),$$

consequentemente,

$$f_{\hat{Y}}(\alpha_1 W^1 + \alpha_2 W^2) = f_{\hat{Y}}(\alpha_1 W^1) + f_{\hat{Y}}(\alpha_2 W^2).$$

Portanto, $f_{\hat{Y}}(W)$ é linear. □

Definiremos a partir do problema (O), uma função reduzida F, e provaremos que esta é côncava.

Definição 4.3. No problema (O), tínhamos a função objetivo completa, com suas restrições, a forma reduzida da mesma função $f(W)$ será dada por:

$$F(W) = \min\left\{f_Y(W) : Y \in \mathbb{R}^{d\times k}\right\},$$

em que W continua sendo uma matriz real de dimensão $k \times n$.

Lema 1. *A função objetivo F reduzida é côncava.*

Demonstração: Dadas as matrizes $W^1, W^2 \in \mathbb{R}^{k\times n}$ e um escalar $\alpha \in [0,1]$. Mostraremos que a desigualdade a seguir acontece,

$$F(\alpha W^1 + (1-\alpha)W^2) \geqslant \alpha F(W^1) + (1-\alpha)F(W^2), \forall \alpha \in [0,1]\,.$$

De fato, se

$$F(\alpha W^1 + (1-\alpha)W^2) = \min\left\{f_Y(\alpha W^1 + (1-\alpha)W^2 : Y \in \mathbb{R}^{d\times k}\right\}$$

pela Propriedade 3, temos a linearidade de f_Y,

$$\min\left\{f_Y(\alpha W^1 + (1-\alpha)W^2 : Y \in \mathbb{R}^{d\times k}\right\} = \min\left\{\alpha f_Y(W^1) + (1-\alpha) f_Y(W^2) : Y \in \mathbb{R}^{d\times k}\right\},$$

retirando os escalares,

$$F(\alpha W^1 + (1-\alpha)W^2) \geqslant \alpha\ \min\left\{f_Y(W^1) : Y \in \mathbb{R}^{d\times k}\right\} + (1-\alpha)\ \min\left\{f_Y(W^2) : Y \in \mathbb{R}^{d\times k}\right\}$$

desta forma, como

$$\alpha\ \min\left\{f_Y(W^1) : Y \in \mathbb{R}^{d\times k}\right\} = \alpha F(W^1)$$

obtemos,

$$F(\alpha W^1 + (1-\alpha)W^2) \geqslant \alpha F(W^1) + (1-\alpha)F(W^2).$$

Portanto, F é côncava. □

Definimos em seguida o conjunto solução do nosso problema como um produto cartesiano de poliedros ($S = S_1 \times S_2 \times ... \times S_n$), e iremos demonstrar que esse é um poliedro convexo.

Teorema 4.4. *O conjunto S é um poliedro convexo.*

Demonstração: Se W_j é uma matriz qualquer pertencente ao poliedro S_j, então podemos escrever equivalentemente o seguinte sistema:

$$S_j = \begin{cases} AW_j = b_j \\ W_j \geqslant 0, \end{cases}$$

em que $A = [1...1]$ uma matriz linha composta por k entradas, W_j é um vetor coluna de W em que a j-ésima coluna está fixada, e b_j = [1].

i) Primeiramente, devemos analisar se $S = S_1 \times S_2 \dots S_n$ é realmente um poliedro. De fato, pois $W \in S \Leftrightarrow \bar{A}x = b$, e $W_j \geqslant 0$ para $j = 1, 2, \dots, n$, ou seja

$$\begin{array}{ccccccc} AW_1 & 0 & \cdots & 0 & = & b_1 & \text{, com } W_1 \geq 0 \\ 0 & AW_2 & \cdots & 0 & = & b_2 & \text{, com } W_2 \geq 0 \\ \vdots & \vdots & \ddots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & AW_n & = & b_n & \text{, com } W_n \geq 0\,. \end{array}$$

Sendo, equivalente a

$$\bar{A} = \begin{bmatrix} A & 0 & \cdots & 0 \\ 0 & A & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A \end{bmatrix}_{n\times(nk)} \cdot \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_n \end{bmatrix}_{(nk)\times 1} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}_{n\times 1}$$

onde,

$$\bar{A} = \begin{bmatrix} 1\cdots 1 & 0\cdots 0 & \cdots & 0\cdots 0 \\ 0\cdots 0 & 1\cdots 1 & \cdots & 0\cdots 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0\cdots 0 & 0\cdots 0 & \cdots & 1\cdots 1 \end{bmatrix},\ x = \begin{bmatrix} W1 \\ W_2 \\ \vdots \\ W_n \end{bmatrix},\ b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}.$$

Dessa forma, W verifica

$$\begin{cases} \bar{A}x = b \\ x \geqslant 0, \end{cases}$$

logo, $W \in S$. Portanto, S é um poliedro.

ii) Agora, devemos mostrar que S é convexo, para tal fim, provamos que cada $S_j \in S$ com $j = \{1, 2, \dots, n\}$ é convexo. De fato, dados W_j^1 e W_j^2 dois elementos de S_j, então $\sum_{i=1}^{k} w_{ij}^1 = 1$ e $w_{ij}^1 \geqslant 0$, o mesmo vale para w_{ij}^2. Queremos provar que o segmento de reta que liga W_j^1 e W_j^2 pertence a S_j, ou seja,

$$\alpha W_j^1 + (1-\alpha) W_j^2 \in S_j.$$

Realizando os cálculos, obtemos

$$\begin{aligned}\sum_{i=1}^{k} \alpha w_{ij}^1 + (1-\alpha) w_{ij}^2 &= \alpha \sum_{i=1}^{k} w_{ij}^1 + (1-\alpha) \sum_{i=1}^{k} w_{ij}^2 \\ &= \alpha + 1 - \alpha \\ &= 1.\end{aligned}$$

Como $\alpha w_{ij}^1 + (1-\alpha) w_{ij}^2 \geqslant 0$ verifica-se que o segmento está contida em S_j, logo é convexo. Dado que S é produto cartesiano de poliedros convexos, portanto é convexo também.

□

Teorema 4.5. *Um ponto W é um ponto extremo de S se, e somente se, W é uma solução admissível de (O).*

Demonstração: Consideremos W um ponto extremo de S, e como $W \in S$ então $\sum_{i=1}^{k} w_{ij} = 1$ e $w_{ij} \geqslant 0$, queremos então provar que $w_{ij} \in \{0,1\}$. Como W é ponto extremo de S temos que satisfaz o sistema da forma $\bar{A}x = b$, sendo que podemos decompor $\bar{A} = [B \mid N]$ em que B é uma matriz formada pelas colunas linearmente independentes da matriz $\bar{A}$, portanto possui inversa e está associada as k variáveis básicas de W, e N contém todas as outras colunas de $\bar{A}$, que estão associadas às variáveis não básicas(x_N), que assumem o valor 0. Assim, obtemos:

$$\begin{aligned}\bar{A}x = b &\Leftrightarrow \bar{A} = [B \mid N] \cdot \begin{bmatrix} x_B \\ x_N \end{bmatrix} = b \\ &\Leftrightarrow Bx_B + Nx_N = b \\ &\Leftrightarrow x_B = b = \mathbb{1}.\end{aligned}$$

Notemos que, todas as variáveis básicas assumem o valor 1, o que prova que $w_{ij} \in \{0,1\}$. Portanto, W é uma solução admissível de (O).

Na implicação contrária, consideramos W uma solução admissível de (O), então $w_{ij} \in \{0,1\}$ e $\sum_{i=1}^{k} w_{ij} = 1$ para todo $j = \{1,2,\ldots,n\}$, logo $W \in S$. Com isso, precisamos provar apenas que W é um ponto estremo de S. Para isso vamos considerar uma submatriz B inversível de W formada pelas colunas onde os elementos w_{ij} valem 1. A matriz B só pode ser a matriz identidade, pois as variáveis w_{ij} com $j = \{1,2,\ldots,n\}$ tem apenas uma entrada igual a 1 e as restantes valem 0, porque $\sum_{i=1}^{k} w_{ij} = 1$ e $w_{ij} \in \{0,1\}$. Daqui, temos que B será uma base, portanto será também solução básica de (O), que implica em W ser um ponto extremo de S. □

Observação 4.6. Como o problema (O) é um problema discreto em que necessita-se análise de cada ponto, os algoritmos tornam-se mais rigorosos. Logo é interessante reduzir o problema substituindo as restrições $w_{ij} \in \{0, 1\}$ por simplesmente $w_{ij} \geqslant 0$.

Definição 4.7. O problema reduzido (RO) de (O) é definido da seguinte forma:

$$\begin{aligned} (RO) : min \quad & f(W), \\ s.a \quad & W \in S. \end{aligned}$$

Propriedade 4. *Se F é uma função côncava num poliedro convexo S, então F atinge o mínimo num extremo de S.*

Demonstração: Seja x um minimizador de F e suponha que x não é um ponto extremo de S.

Ou seja, podemos escrever x como uma combinação linear de pontos extremos de S, isto é, $x = \sum_{i=1}^{n} \alpha_i x_i$ tal que $\sum_{i=1}^{n} \alpha_i = 1$, com x_i ponto extremo de S e $\alpha_i \geqslant 0$.
Como F é côncava, então temos:

$$\begin{aligned} F\left(\sum_{i=1}^{n} \alpha_i x_i\right) &\geqslant \sum_{i=1}^{n} \alpha_i F(x_i) \\ F(x) &\geqslant \sum_{i=1}^{n} \alpha_i F(x_i) \\ F(x) - \sum_{i=1}^{n} \alpha_i F(x_i) &\geqslant 0. \end{aligned}$$

Sendo $\sum_{i=1}^{n} \alpha_i = 1$, obtemos,

$$\begin{aligned} \sum_{i=1}^{n} \alpha_i F(x) - \sum_{i=1}^{n} \alpha_i F(x_i) &\geqslant 0 \\ \sum_{i=1}^{n} \alpha_i \left(F(x) - F(x_i)\right) &\geqslant 0. \end{aligned} \tag{4.7}$$

Sendo $F(x)$ um mínimo, significa que,

$$F(x) \leqslant F(x_i), \text{ para todo } i = 1, 2, ..., n.$$

ou seja,

$$F(x) - F(x_i) \leqslant 0, \text{ para todo } i = 1, 2, ..., n. \tag{4.8}$$

Assim, por (4.7) e (4.8) resulta,

$$F(x) = F(x_i), \text{ para todo } i = 1, 2, ..., n.$$

Portanto, F também atinge o mínimo nos pontos extremos x_i. □

Daqui, obtemos o seguinte lema:

Lema 4.8. *O problema (RO) e (O) são equivalentes.*

Demonstração: De fato,

(i) Se W é uma solução admissível de (RO), temos que F atinge o mínimo em W. Pela Propriedade 4, temos que W é um ponto extremos de S. Assim, pelo Teorema 4.5, W é uma solução admissível de (O).

(ii) Se W é uma solução admissível de (O), temos que F também atinge o mínimo em W. Por outro lado, como W é uma solução admissível de (O), então $w_{ij} \in \{0, 1\}$ e $\sum_{i=1}^{k} w_{ij} = 1$ $\forall j = 1, 2, ..., n$. Daqui, $W \in S$, e pelo Teorema 4.5 W é um ponto extremo de S.

Portanto, W é uma solução admissível de (RO).

□

Observação 4.9. A solução direta do problema (RO) não é fácil de encontrar, por esta razão, introduzindo o conceito de *solução ótima parcial*.

Definição 4.10. Uma solução ótima parcial do problema (O) é uma solução (W^*, Y^*) que satisfaz as seguintes condições:

$$\begin{array}{llll} \textit{(i)} & f_{Y^*}(W^*) & \leqslant & f_{Y^*}(W), \quad \text{para todo } W \in S. \\ \textit{(ii)} & f_{W^*}(Y^*) & \leqslant & f_{W^*}(Y), \quad \text{para todo } Y \in \mathbb{R}^{d\times k}. \end{array}$$

Para encontrar uma solução ótima parcial, (W^*, Y^*) do problema (O), minimizamos alternadamente a função f em W e Y, por meio de dois tipos de problemas de minimização:

- O problema $P_{\hat{Y}}$ pretende minimizar $f_{\hat{Y}}(W)$ tal que

$$W \in S, \text{ dado um } \hat{Y} \in \mathbb{R}^{d\times k}.$$

Este é responsável por atribuir os pontos aos *clusters*.

- O problema $P_{\hat{W}}$ pretende minimizar $f_{\hat{W}}(Y)$ em que

$$Y \in S, \text{ dado um } \hat{W} \in \mathbb{R}^{d\times k}.$$

Em $P_{\hat{W}}$ atualizam-se os centroides.

Observação 4.11. A solução para o $P_{\hat{Y}}$ é simples, pois $D(x_j, y_i) = \|x_j - y_i\|_2^2$ como se verifica no Lema a seguir.

Lema 4.12. *Seja $\hat{Y} \in \mathbb{R}^{d\times k}$ fixo, define-se $\tilde{W} \in \mathbb{R}^{k\times n}$ da seguinte maneira: para cada $j \in \{1, 2, ..., n\}$, temos*

$$\tilde{w}_{ij} = \begin{cases} 1, & \text{se} \quad \|x_j - \hat{y}_i\|_2^2 \leqslant \|x_j - \hat{y}_l\|_2^2, \text{ para todo } l \in \{1, 2, ..., k\} \\ 0, & \text{se caso contrário.} \end{cases}$$

Então $\tilde{W}$ é uma solução ótima do problema $P_{\hat{Y}}$.

Demonstração: Inicialmente provamos que $\tilde{W} \in S$. Da definição de $\tilde{w}_{ij}$, temos $\tilde{w}_{ij} \in \{0, 1\}$, logo $\tilde{w}_{ij} \geqslant 0$.

Notemos que para cada $j \in \{1, 2, ..., n\}$, existe apenas um $i^* \in \{1, 2, ..., k\}$, em que $\tilde{w}_{i^*j} = 1$, ou seja, existe um índice i^* onde a distância $\|x_j - \hat{y}_{i^*}\|_2^2$ é mínima. Para os restantes índices $i \neq i^*$ em $\{1, 2, ..., k\}$, temos que $\hat{w}_{ij} = 0$.

Daqui,

$$\sum_{i=1}^{k} \hat{w}_{ij} = 1.$$

Portanto, $\tilde{W} \in S$.

Agora, provaremos que $f_{\hat{Y}}(\tilde{W}) \leqslant f_{\hat{Y}}(W)$, para todo $W \in \mathbb{R}^{k\times n}$. Para isso, consideremos arbitrariamente um índice $j \in \{1, 2, ..., n\}$ e, seja $i^* \in \{1, 2, ..., k\}$ tal que,

$$\tilde{w}_{i^*j} = 1, \quad \text{se } \|x_j - \hat{y}_{i^*}\|_2^2 \leqslant \|x_j - \hat{y}_l\|_2^2, \quad \text{para todo } l \in \{1, 2, ..., k\}\,.$$

Para todo $(w_{ij}) = W \in S$, temos

$$w_{lj}\|x_j - \hat{y}_{i^*}\|_2^2 \leqslant w_{lj}\|x_j - \hat{y}_l\|_2^2, \text{ para todo } l \in \{1, 2, ..., k\}\,,$$

obteremos aqui k desigualdades. Somando elas membro a membro, resulta em:

$$\sum_{l=1}^{k} w_{lj}\|x_j - \hat{y}_{i^*}\|_2^2 \leqslant \sum_{l=1}^{k} w_{lj}\|x_j - \hat{y}_l\|_2^2,$$

e ainda como $\sum_{l=1}^{k} w_{lj} = 1$, temos que

$$\|x_j - \hat{y}_{i^*}\|_2^2 \leqslant \sum_{l=1}^{k} w_{lj}\|x_j - \hat{y}_l\|_2^2. \tag{4.9}$$

Vamos analisar agora somente o termo $\|x_j - \hat{y}_{i^*}\|_2^2$ da equação (4.9)

$$\|x_j - \hat{y}_{i^*}\|_2^2 \;=\; \tilde{w}_{i^*j}\|x_j - \hat{y}_{i^*}\|_2^2,$$

devido a $\tilde{w}_{i^*j} = 1$, e como $\hat{w}_{lj} = 0$ para todo $l \neq i^*$, obtemos

$$\|x_j - \hat{y}_{i^*}\|_2^2 = \tilde{w}_{i^*j}\|x_j - \hat{y}_{i^*}\|_2^2 + \sum_{l\neq i^*} \tilde{w}_{ij}\|x_j - \hat{y}_l\|_2^2$$

$$\|x_j - \hat{y}_{i^*}\|_2^2 = \sum_{l=1}^{k} \tilde{w}_{lj}\|x_j - \hat{y}_l\|_2^2 \,. \tag{4.10}$$

Substituindo (4.10) em (4.9), resulta

$$\sum_{l=1}^{k} \tilde{w}_{lj}\|x_j - \hat{y}_l\|_2^2 \;\leqslant\; \sum_{l=1}^{k} w_{lj}\|x_j - \hat{y}_l\|_2^2 \,,$$

aplicamos o somatório em j em ambos os lados dessa desigualdade, e encontramos

$$\sum_{j=1}^{n}\sum_{l=1}^{k} \tilde{w}_{lj}\|x_j - \hat{y}_l\|_2^2 \;\leqslant\; \sum_{j=1}^{n}\sum_{l=1}^{k} w_{lj}\|x_j - \hat{y}_l\|_2^2 \,.$$

Permutando os somatórios, temos

$$\sum_{i=1}^{k}\sum_{j=1}^{n} \tilde{w}_{ij}\|x_j - \hat{y}_l\|_2^2 \;\leqslant\; \sum_{l=1}^{k}\sum_{j=1}^{n} w_{lj}\|x_j - \hat{y}_l\|_2^2 \,,$$

daqui, pela definição de f,

$$f_{\hat{Y}}(\tilde{W}) \;\leqslant\; f_{\hat{Y}}(W).$$

Portanto, $\tilde{W}$ é uma solução ótima de $P_{\hat{Y}}$. □

Observação 4.13. A solução para o problema P_W não é tão simples de se obter. Mas, no método *K-means* através da função de semelhança, $D(x, y) = \|x - y\|_2^2$, é possível deduzir uma expressão direta para a solução ótima de $P_{\hat{W}}$. Como vemos no próximo lema.

Lema 4.14. *Para um $\hat{W} \in \mathbb{R}^{n\times k}$ fixo, definiremos $Y \in \mathbb{R}^{d\times k}$ da seguinte forma:*

$$Y = [y_1, y_2, \ldots, y_k] \,,\ \textit{onde}\ y_i = \frac{\displaystyle\sum_{j=1}^{n} w_{ij}x_j}{\displaystyle\sum_{j=1}^{n} w_{ij}} \,,\ i = 1, 2, \ldots, k.$$

Então Y é uma solução ótima de $P_{\hat{W}}$.

Demonstração: Consideremos um $\hat{W} \in \mathbb{R}^{k\times n}$ fixo. Notemos que, pela Propriedade 2, a função objetivo definida por

$$f_{\hat{W}}(Y) = \sum_{i=1}^{k}\sum_{j=1}^{n} \hat{w}_{ij}\|x_j - y_i\|_2^2 \,,$$

é uma função estritamente convexa. Logo, se encontrarmos um $Y \in \mathbb{R}^{d\times k}$ em que $f_{\hat{W}}$ atinge o mínimo, ele é único.

Segue que, para encontrar a solução de $P_{\hat{W}}$, precisamos achar o zero da derivada da função $f_{\hat{W}}$, em relação a y_i, $i \in \{1, 2, \ldots, k\}$.

Derivando $f_{\hat{W}}$, temos

$$\begin{aligned}
\nabla_{y_i} f_{\hat{W}}(Y) &= \nabla_{y_i} \sum_{i=1}^{k}\sum_{j=1}^{n} \hat{w}_{ij}\|x_j - y_i\|_2^2 \\
&= \sum_{j=1}^{n} \hat{w}_{ij}(-2(x_j - y_i))\nabla_{y_i} y_i \\
&= \sum_{j=1}^{n} \hat{w}_{ij}(-2(x_j - y_i)).
\end{aligned}$$

Igualando a zero para encontrar os pontos críticos, resulta

$$\begin{aligned}
\nabla_{y_i} f_{\hat{W}}(Y) &= 0 \\
-2\sum_{j=1}^{n} \hat{w}_{ij}(x_j - y_i) &= 0 \\
\sum_{j=1}^{n} \hat{w}_{ij}x_j - \sum_{j=1}^{n} \hat{w}_{ij}y_i &= 0 \\
y_j \sum_{j=1}^{n} \hat{w}_{ij} &= \sum_{j=1}^{n} \hat{w}_{ij}x_j \\
y_j &= \frac{\sum_{j=1}^{n} \hat{w}_{ij}x_j}{\sum_{j=1}^{n} \hat{w}_{ij}}, \; i = 1, 2, \ldots, k
\end{aligned}$$

Portanto, $Y = [y_1, y_2 \ldots, y_k]$ é solução de $P_{\hat{W}}$. □

As soluções de $\tilde{W} \in \mathbb{R}^{d\times k}$ dos problemas $P_{\hat{Y}}$ e $P_{\hat{W}}$, respectivamente, satisfazem as condições de Karush-Kuhn-Tucker (KKT) para o problema (RO), que são as condições necessárias de otimalidade de primeira ordem. Por esse motivo são candidatos a uma solução ótima parcial de (O), como estabelece o seguinte Teorema 4.15.

Para não fugir do escopo deste trabalho, a análise de Karush-Kuhn-Tucker (KKT) pode

ser encontrada em [8] p. 13.

Teorema 4.15. *(W^*, Y^*) é um ponto KKT do problema reduzido (RO) se, e somente se, (W^*, Y^*) for uma solução ótima parcial do problema (O).*

A seguir apresentamos dois lemas importantes que ajudarão a provar a convergência do algoritmo *K-means*.

Observação 4.16. Consideremos a seguinte notação:

- W^t são as atribuições dos pontos em relação à iteração t.
- Y^t são os centroides na iteração t.

Lema 4.17. *Para quaisquer iterações t se $Y^{t-1} \neq W^t$, então $f(W^{t-1}, Y^t) < f(W^{t-2}, Y^{t-1})$*

Demonstração: Considerando a iteração t temos:
(i) Para o problema P_Y:

W^{t-1} é solução do problema $P_{Y^{t-1}}$, isto é

$$f_{Y^{t-1}}(W^{t-1}) \leqslant f_{Y^{t-1}}(W), \text{ para todo } W \in S.$$

Logo,

$$f_{Y^{t-1}}(W^{t-1}) \leqslant f_{Y^{t-1}}(W^{t-2}),$$

equivalentemente,

$$f(W^{t-1}, Y^{t-1}) \leqslant f(W^{t-2}, Y^{t-1}) \tag{4.11}$$

(ii) Para o problema P_W:

Y^t é solução do problema $P_{W^{t-1}}$, ou seja,

$$f_{W^{t-1}}(Y^t) \leqslant f_{W^{t-1}}(Y), \text{ para todo } Y \in \mathbb{R}^{d\times k}.$$

Em particular, vale que

$$f_{W^{t-1}}(Y^t) \leqslant f_{W^{t-1}}(Y^{t-1}),$$

logo,

$$f(W^{t-1}, Y^t) \leqslant f(W^{t-1}, Y^{t-1}). \tag{4.12}$$

Assim, de (4.11) e (4.12), segue que

$$f(W^{t-1}, Y^t) \leqslant f(W^{t-2}, Y^{t-1}).$$

Se $Y^{t-1} \neq Y^t$, temos:

$$f(W^{t-1}, Y^t) \quad < \quad f(W^{t-2}, Y^{t-1}).$$

Como $f_{W^{t-1}}$ é estritamente convexa, concluímos que o mínimo é único. □

Observação 4.18. Através do lema anterior provamos que a sequência (x_t) de termos $x_t = f(W^{t-1}, Y^t)$ é decrescente.

Lema 4.19. *Suponhamos que o algoritmo K-means realizou $t-$iterações. Então $W^0, W^1, \ldots, W^t$ são distintos entre si.*

Demonstração: Suponhamos por absurdo que, existem duas iterações t_1 e t_2 antes da iteração t tal que, $W^{t_1} = W^{t_2}$, vamos supor sem perda de generalidade que $t_1 < t_2 < t$.
Conforme as iterações, obteremos:

- Y^{t_1+1} é solução do problema $P_{W^{t_1}}$.
- Y^{t_2+1} é solução do problema $P_{W^{t_2}}$.

Dado que $W^{t_1} = W^{t_2}$, e pela unicidade da solução do problema, $P_{W^{t_1}} = P_{W^{t_2}}$, resulta em:

$$Y^{t_1+1} = Y^{t_2+1},$$

por consequência,

$$f(W^{t_1}, Y^{t_1+1}) \quad = \quad f(W^{t_2}, Y^{t_2+1}).$$

Como a sequência (x_t) de termos $f(W^{t-1}, Y^t)$ é decrescente, teríamos que

$$Y^{t_1+1} = Y^{t_1+2} = \cdots = Y^{t_2} = Y^{t_2+1},$$

Logo, o algoritmo *K-means* deveria parar na iteração $t_1 + 1$ e não na iteração t, lembremos $t_1 + 1 < t$, resultando em uma contradição. Portanto, todos os W^j, com $j = 0, 1, \cdots, t$ são distintos entre si. □

Teorema 4.20. *O algoritmo K-means converge para uma solução ótima parcial do problema (O) num número finito de iterações.*

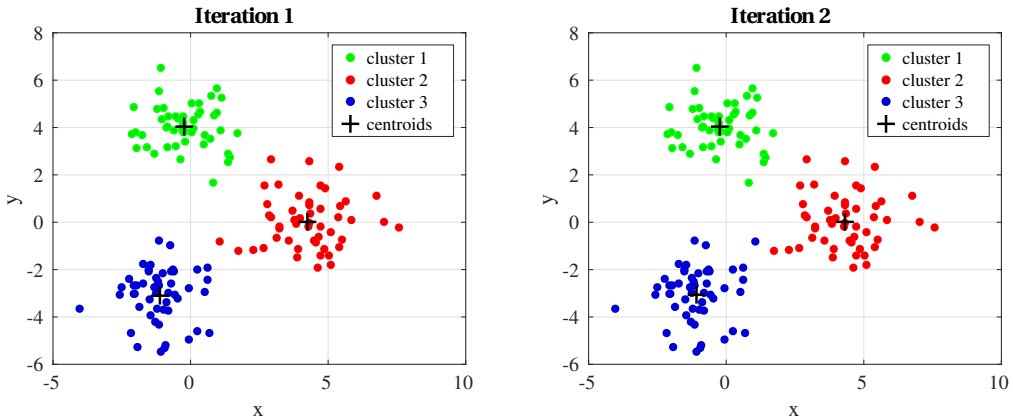
Demonstração: Do Lema 4.19, seque que todos os W^t encontrados pelo algoritmo *K-means* são distintos, e conforme o Teorema 4.5 correspondem a pontos extremos do poliedro S, sendo eles uma quantidade finita. Dessa forma, concluímos que o algoritmo *K-means*, precisa de no máximo um número finito de iterações para convergir a uma solução. □

4.4 Vantagens e Desvantagens do Método *K-means*

O algoritmo *K-means* tem como vantagens [5], por exemplo, ser relativamente escalável e eficiente para grandes conjuntos de dados. O método frequentemente termina num local ótimo. Entretanto, este método só pode ser aplicado quando a média (centroide) de um *cluster* pode ser definido (ver a Figura 4.1, e mais detalhes da base de dados no Apêndice A). Isto pode não ser o caso em algumas aplicações, que utilizam dados com atributos categóricos (nominais) envolvidos. A abordagem por *K-means* é sensível à partição inicial, gerada pela escolha aleatória dos centroides. A técnica *K-means* necessita que o número k de *clusters* seja informado com antecedência. Além disso, ele é sensível a ruídos, visto que um pequeno número de tais dados pode influenciar, substancialmente, o valor médio das coordenadas.

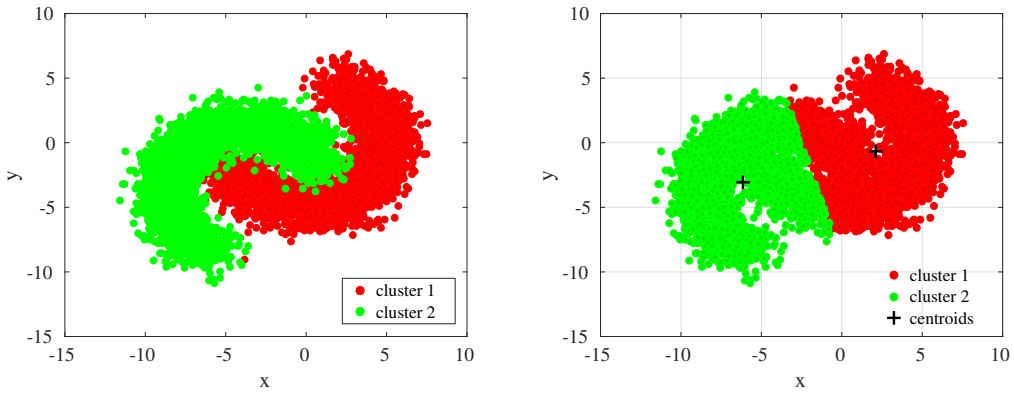
O algoritmo *K-means* também possui certas deficiências [5], por exemplo, o resultado depende muito do palpite inicial dos centroides. Somente atributos numéricos são abordados. O algoritmo não é adequado para descobrir *clusters* com formas não convexas (ver a Figura 4.2, e mais detalhes da base de dados no Apêndice B). Pelo fato deste método gerar *clusters* com figuras circulares, este problema é conhecido como problema da superposição de classes. Também, o algoritmo tem problemas quando os *clusters* são de formas não globulares (ver a Figura 4.3 e Figura 4.4 e mais detalhes nos Apêndices C e D, respectivamente), ou quando os *clusters* têm densidades diferentes (ver a Figura 4.5 e mais detalhes no Apêndice E).

Figura 4.1: Conjunto de Dados: Dados Sintéticos.
Agrupamento usando o *K-means* (3 *clusters*)



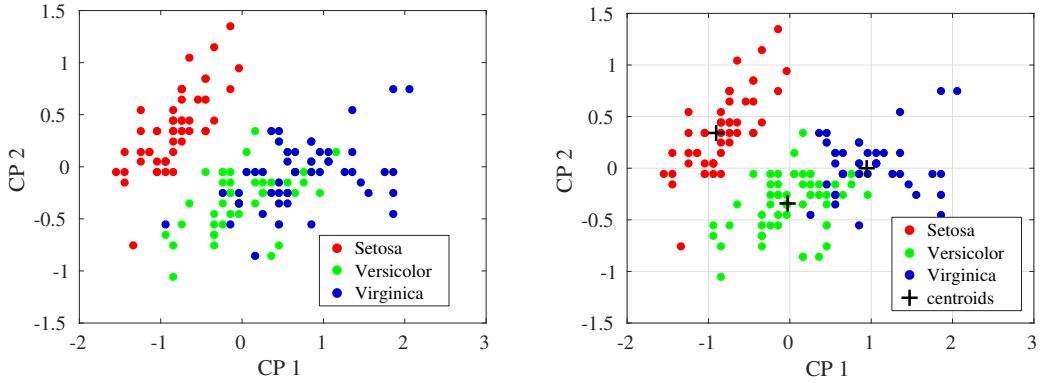
Fonte: Arquivo pessoal

Figura 4.2: Conjunto de Dados: Bananas.
À esquerda, pontos originais; à direita, o agrupamento usando o *K-means* (2 *clusters)*



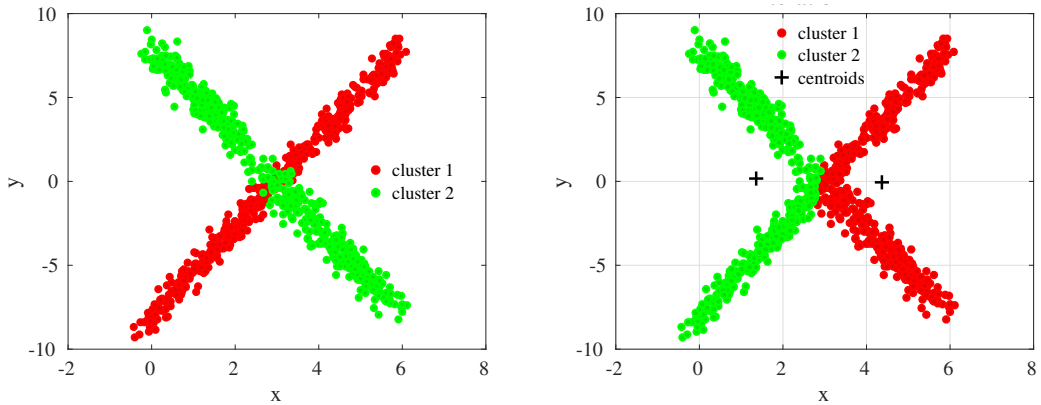
Fonte: Arquivo pessoal

Figura 4.3: Conjunto de Dados: Íris, usando a projeção sobre suas componentes principais.
À esquerda, pontos originais; à direita, o agrupamento usando o *K-means* (3 *clusters)*



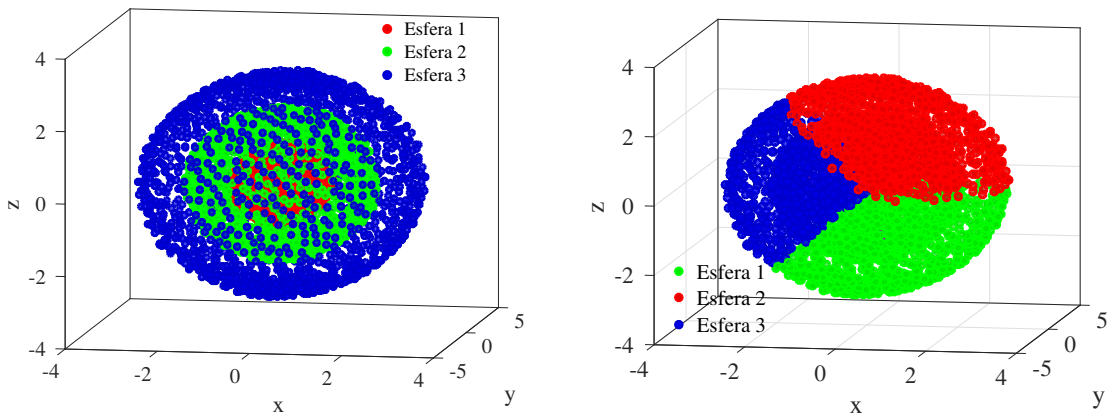
Fonte: Arquivo pessoal

Figura 4.4: Conjunto de Dados: Letra X.
À esquerda, pontos originais; à direita, o agrupamento usando o *K-means* (2 *clusters*)



Fonte: Arquivo pessoal

Figura 4.5: Conjunto de Dados: Esferas.
À esquerda, pontos originais usando projeção; à direita, o agrupamento usando o *K-means* (3 *clusters*)



Fonte: Arquivo pessoal

Capítulo 5

Aplicações do *K-means*

Neste capítulo, apresentamos a descrição e os resultados obtidos em dois conjuntos de dados que pertencem ao banco de dados da Universidade da Califórnia Irvine (UCI), publicamente disponíveis no repositório de Aprendizado de Máquinas da UCI [7].

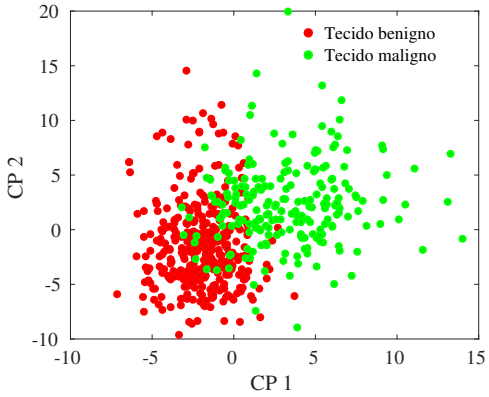
5.1 Conjunto de Dados: Câncer de Mama

Este conjunto de dados corresponde a um diagnóstico de câncer de mama de pacientes mulheres do estado de Winconsin (EUA). Assim, o objetivo do conjunto de dados é prever se uma paciente tem ou não câncer de mama com base em algumas medidas de diagnóstico (atributos) incluídas no conjunto de dados. Esses atributos foram calculados a partir de uma imagem digitalizada de um aspirado por agulha fina (PAAF) de uma massa mamária, e descrevem as características dos núcleos celulares presentes na imagem. Trata-se de um conjunto de dados constituído de 569 amostras (instâncias) organizadas em 2 classes de tecido mamário: a classe "*tecido benigno*" composta de 357 amostras; e a classe "*tecido maligno*" constituída de 212 amostras. Vale observar que cada amostra apresenta 30 atributos (variáveis independentes); e algumas das informações desses atributos, para cada núcleo celular, são:

- *Radius_mean* : a média das distâncias do centro aos pontos no perímetro;
- *Texture_mean* : desvio padrão dos valores da escala de cinza;
- *Perimeter_mean* : tamanho médio do tumor central;
- *Smoothness_mean* : média da variação local nos comprimentos do raio;
- *Concave points_mean* : média do número de porções côncavas do contorno;
- *Radius_se* : o erro padrão para a média das distâncias do centro até os pontos no perímetro;
- *Texture_se* : o erro padrão para o desvio padrão dos valores em escala de cinza.

Na Figura 5.1, apresentamos a projeção em $\mathbb{R}^2$ do conjunto de dados sobre suas componentes principais (CP 1 e CP 2). A classe tecido benigno é representado por as bolas da cor vermelho; e a classe tecido maligno, por as bolas da cor verde. Também, podemos observar que existe uma boa separação dos *clusters*, e também que existem alguns pontos conhecidos como *outliers* [1].

Figura 5.1: Dados originais usando a projeção sobre suas componentes principais do conjunto de dados Câncer de Mama

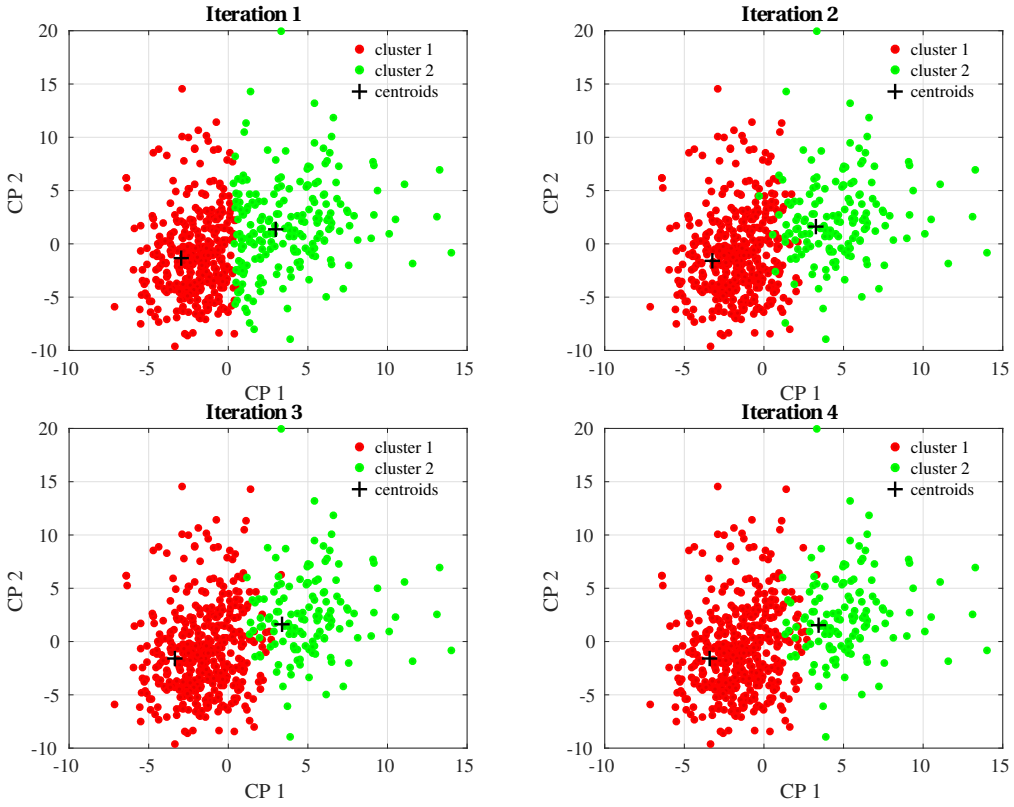


Fonte: Arquivo pessoal

Na Figura 5.2, apresentamos o funcionamento do *K-means* em 4 iterações, e a movimentação dos centroides de cada *cluster*. O melhor resultado é apresentado na figura da iteração 4. Vale observar que existe uma ótima separação dos *clusters*, e o resultado é muito semelhante a os dados originais apresentados na Figura 5.1.

[1] Outliers são pontos de dados muito distantes de outros pontos de dados

Figura 5.2: Agrupamento usando o *K-means* (2 *clusters*) e a projeção sobre suas componentes principais do conjunto de dados Câncer de Mama



Fonte: Arquivo pessoal

5.2 Conjunto de Dados: Diabetes

Este conjunto de dados é originalmente do Instituto Nacional de Diabetes e Doenças Digestivas e Renais. O objetivo do conjunto de dados é prever se um paciente (somente pacientes do sexo feminino) apresenta, de acordo com a Organização Mundial de Saúde, sinais de diabetes, com base em algumas medidas de diagnóstico (atributos) incluídas no conjunto de dados. Várias restrições foram colocadas na seleção dessas amostras em um banco de dados maior. Em particular, todos os pacientes aqui são mulheres com pelo menos 21 anos de idade da herança indígena Pima.

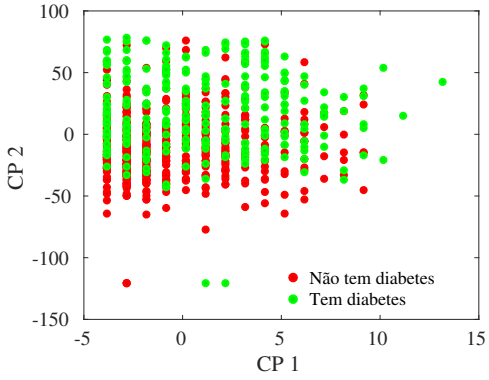
O conjunto de dados é constituída de 768 amostras (instâncias). Cada amostra pertence a uma das seguintes classes: " *não tem diabetes*" (interpretado como teste negativo para diabetes) formada por 500 amostras; e "*tem diabetes*" (interpretado como teste positivo para diabetes) formada por 268 amostras. A análise envolve 8 atributos numéricos (variáveis independentes) a saber:

- *Pregnancies*: número de vezes grávida;
- *Glucose* : concentração de glucose no plasma a 2 horas num teste de tolerância oral à glicose;

- *BloodPressure* : pressão arterial diastólica (mmHg);
- *Skin Thickness* : dobras cutâneas triciptal (mm);
- *Insulin* : nível de insulina no soro (MUU/ ml);;
- *BMI* : índice de massa corporal;
- *DiabetesPedigreeFunction* : função pedigree do diabetes;
- *Age* : idade.

Agora, na Figura 5.3 apresentamos a projeção em $\mathbb{R}^2$ do conjunto de dados sobre suas componentes principais (CP 1 e CP 2). A classe *não tem diabetes* é representado por bolas da cor vermelho; e a classe *tem diabetes*, por bolas da cor verde. É importante saber que os atributos desse conjunto de dados são complexos; porem, logo de usar o algoritmo *K-means* podemos observar que existe uma boa separação dos *clusters*, assim como também apresenta alguns *outliers*.

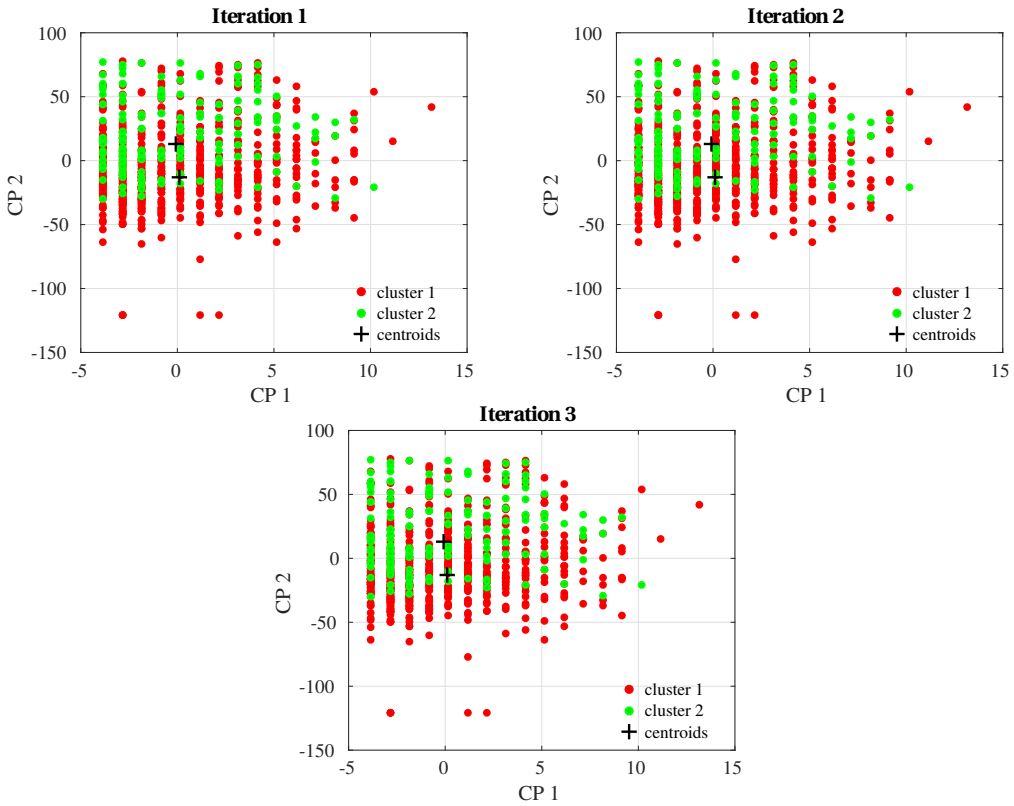
Figura 5.3: Dados originais usando a projeção sobre suas componentes principais do conjunto de dados Diabetes



Fonte: Arquivo pessoal

Na Figura 5.4, observamos o funcionamento do *K-means* em 3 iterações, e a movimentação dos centroides de cada *cluster*. O melhor resultado é apresentado na figura da iteração 3. É fácil observar que existe uma boa identificação dos *clusters*, e o resultado é semelhante a os dados originais apresentados na Figura 5.3.

Figura 5.4: Agrupamento usando o *K-means* (2 *clusters*) e a projeção sobre suas componentes principais do conjunto de dados Diabetes



Fonte: Arquivo pessoal

Capítulo 6

Considerações Finais

Neste trabalho estudou-se matematicamente a análise de convergência do algoritmo *K-means*, que se caracteriza por não necessitar de um supervisor que defina os padrões a serem gerados no processo iterativo, e principalmente por usar a distância euclidiana entre os elementos e os centroides para agrupar os dados em *clusters*, segundo um grau de similaridade que a máquina entende.

Vimos o funcionamento do algoritmo nas aplicações em dois conjuntos de dados voltados para Medicina, foram os do câncer de mama e diabetes. Escolhemos os dados do câncer de mama, por ser, segundo o Instituto Nacional do Câncer (INCA), a segunda doença mais comum entre as mulheres de todo o mundo, ficando atrás somente do câncer de pele. No ano de 2018 foram esperados mais de cinquenta e nove mil novos casos, mais de dezesseis mil mortes de mulheres e duzentos de homens acometidos pela doença. Escolhemos também os dados da diabetes, por segundo a Sociedade Brasileira de Diabetes (SBD), existir hoje no Brasil mais de treze milhões de pessoas diagnosticadas com essa doença, isso representa quase 7% da população brasileira. A clusterização do conjunto de dados de diagnóstico do câncer de mama e diabetes no algoritmo *K-means*, se mostrou bastante eficaz, ou seja, conseguiu resultados satisfatórios a nível de qualidade dos *clusters*, quando comparados com dados originais. Percebemos pouca ou quase nenhuma diferença entre elas, assim como a não superposição dos dados.

Dessa forma, as vantagens apresentadas podem ser facilmente percebidas quando comparamos dados originais com os dados clusterizados pelo algoritmo. Porém o método possui algumas desvantagens, como dificuldades em definir a quantidade k de *clusters*, limitado a atributos numéricos, cada item deve permanecer a único *cluster*, ou seja, não deve haver suposições de dados.

Isso nos permite ter uma visão introdutória do Aprendizado de Máquina, um tema de estudo que atualmente está sendo desenvolvido e possui avanços interessantes. Está se investindo muito em pesquisa direcionada a isso, uma vez que este é uma ramificação da inteligência

artificial, que busca compreender como as máquinas aprendem, como cálculos prévios interferem na produção de decisões da máquina e dos seus resultados. Suas aplicações estão presentes em inúmeras áreas, tais como: análise de redes sociais, síntese de proteínas, mecanismos de pesquisas, segmentação de clientes no setor comercial e bancário, etc.

Os estudos desenvolvidos nessa monografia servirão de base para estudos futuros.

Referências

[1] ALPAYDIN, Ethem. **Introduction to Machine Learning**. 2. ed. Cambridge: Massachusetts Institute of Technology, 2010.

[2] AMORIM, Ronan Gomes de. **Introdução à Análise Convexa**. 2013. 82 f. Dissertação (Mestrado Profissional em Matemática) - Instituto de Matemática e Estatística - Universidade Federal de Goiás. Goiânia, 2013.

[3] ANDRADE, João Santos. **Algoritmo do Ponto Proximal Generalizado para o Problema de Desigualdade Variacional em $\mathbf{R}^n$**. 2010. 67 f. Dissertação (Mestrado em Matemática) - Centro de Ciências da Natureza - Universidade Federal de Piauí. Teresina 2010.

[4] ANIL K, Jain. Data clustering: 50 years beyond K-means. **Pattern Recognition Letters**. 1 jun. 2010. Disponível em: $<$ https://www.journals.elsevier.com/pattern-recognition-letters $>$. Acesso em: 27 nov. 2019.

[5] BERKHIN, Pavel. **Survey of clustering data mining techniques**. Technical report. In: Accrue software. 2002.

[6] CAVAMURA ENDO, Daniela Hiromi. **Espaços Métricos**: uma introdução. 2015. 82 f. Monografia (Matemática) - Departamento de Matemática - Universidade Federal de São Carlos. São Carlos, 2015.

[7] DUA, Dheeru; GRAFF, Casey. **UCI Machine Learning Repository**. 2017. Disponível em : https://archive.ics.uci.edu/ml/datasets/. University of California, Irvine, School of Information and Computer Science. Accesso em: 7 nov. 2019.

[8] FREITAS NUNES, Diogo Henriques. **Um breve estudo sobre o algoritmo K-means**. 2016. 60 f. Dissertação (Mestrado em Matemática) - Departamento de Matemática - Faculdade de Ciências e Tecnologia de Coimbra. Coimbra, 2016.

[9] GERHARDT, Tatiana Engel.; SILVEIRA, Denise Tolfo. **Métodos de pesquisa**. 1. ed. Editora da UFRGS: Porto Alegre 2009.

[10] IZMAILOV, Alexey.; SOLODOV, Mikhail. **Otimização volume 2. Métodos Computacionais**. 2. ed. Impa: Rio de Janeiro, 2012.

[11] JAMES, Gareth.; WITTEN, Daniela.; HASTIE, Trevor.; TIBSHIRANI, Robert. **An Introduction to Statistical Learning**. 1. ed. Springer: Nova York 2013.

[12] LIMA, Elon. **Espaços Métricos**. 3. ed. Projeto Euclides: Rio de Janeiro 1993.

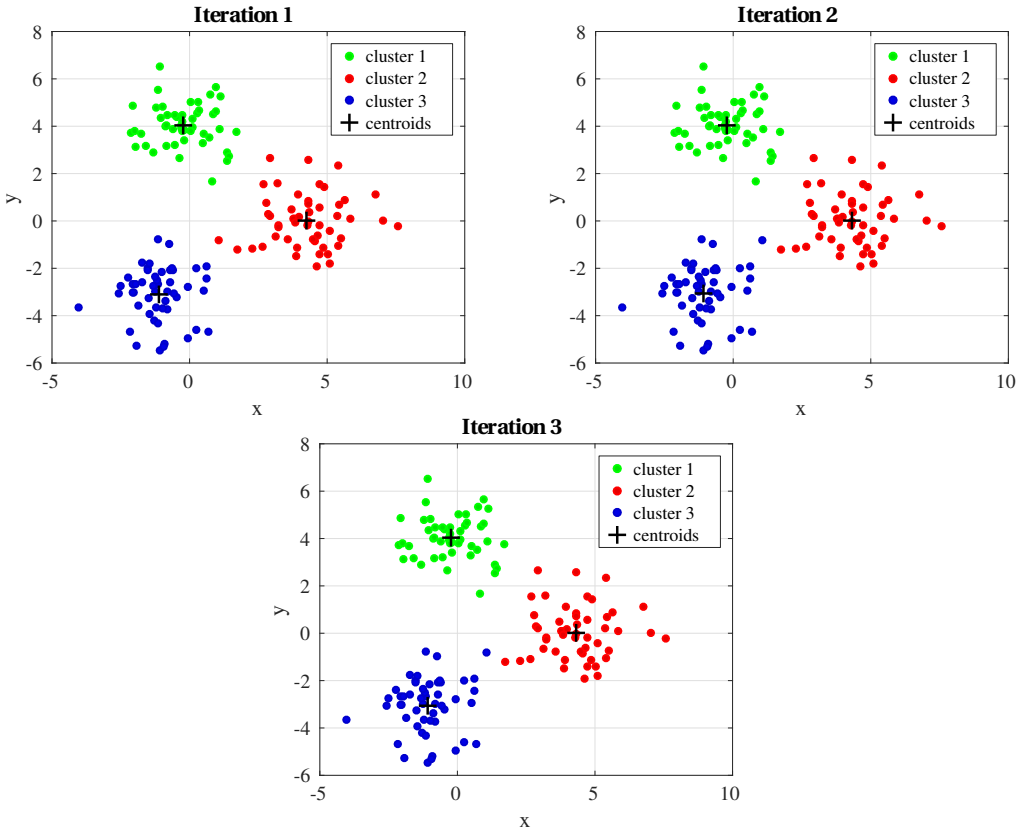
[13] MEDIUM. **Medium:** Machine Learning, 2014 - 2018. Página inicial. Disponível em: $<$ https://medium.com $>$. Acesso em: 30 nov. 2019.

Apêndice A

Dados Sintéticos

Trata-se de um conjunto composto por 150 amostras (instâncias) originados aleatoriamente. Cada amostra pertence a um dos três tipos de classes: *cluster* 1, *cluster* 2 e *cluster* 3. Cada uma das classes é composta por 50 amostras, e cada amostra possui dois atributos ("x" e "y").

Figura A.1: Conjunto de Dados: Dados Sintéticos



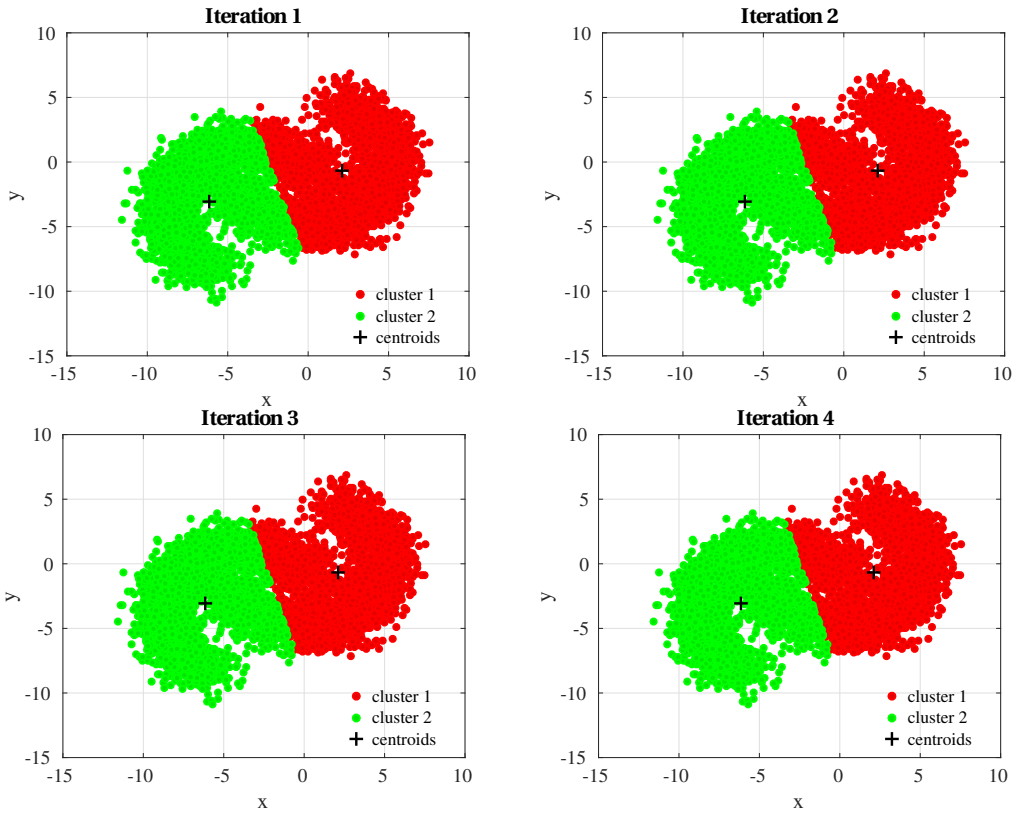
Fonte: Arquivo pessoal

Apêndice B

Bananas

Trata-se de um conjunto composto por 5000 amostras extraídas do Machine Learning Repositttory [7]. Cada amostra pertence a um dos 2 tipos de classes: *cluster* 1 e *cluster* 2. Cada uma das classes é composta por 2500 amostras, e cada amostra possui dois atributos ("x" e "y").

Figura B.1: Conjunto de Dados: Bananas



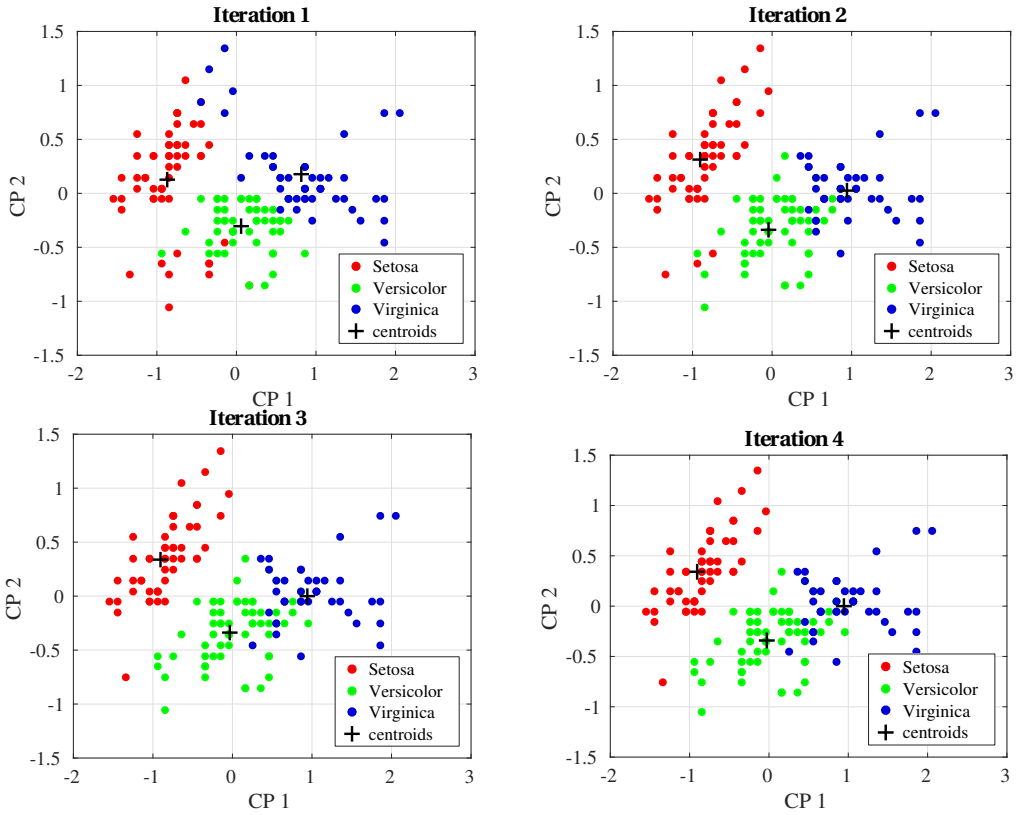
Fonte: Arquivo pessoal

Apêndice C

Íris

Trata-se de conjunto composto de 150 amostras de plantas Iris, extraída do Machine Learning Repositttory [7]. Cada amostra pertence a um dos três tipos de classes: Iris Setosa, Iris Versicolor e Iris Virgínica. Cada uma das classes é composta por 50 amostras, e cada indivíduo (planta) é descrito por quatro características quantitativas: comprimento da sépala, largura da sépala, comprimento da pétala e largura da pétala.

Figura C.1: Conjunto de Dados: Íris



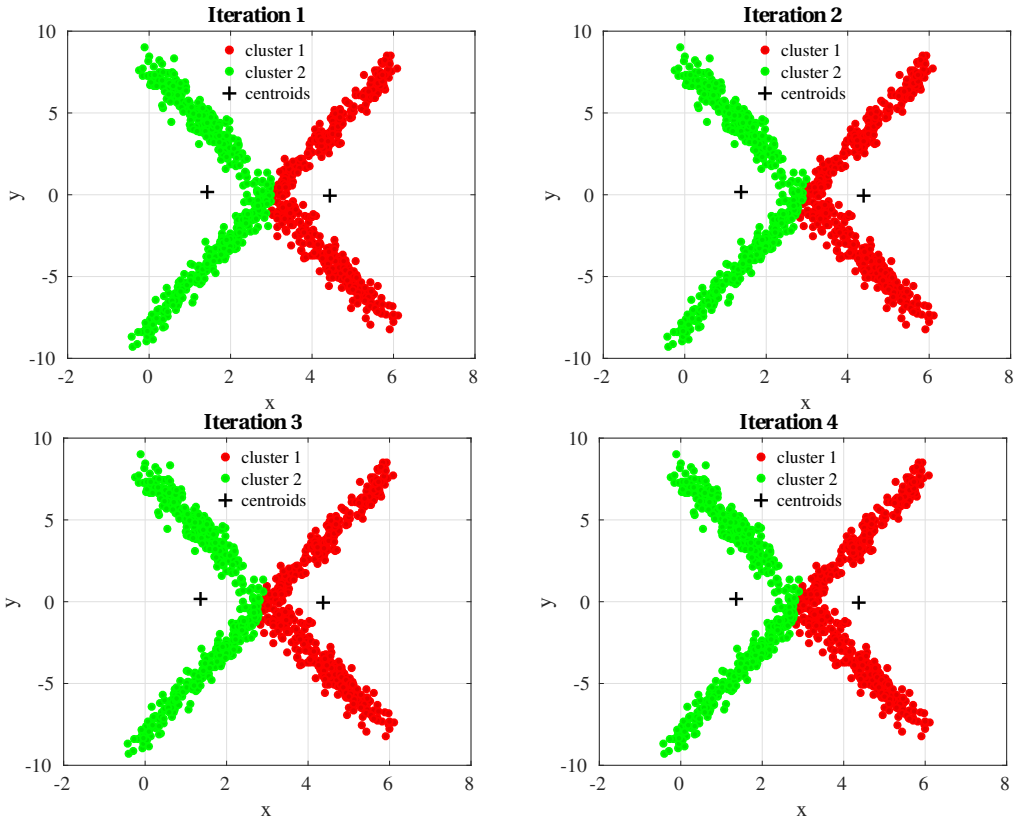
Fonte: Arquivo pessoal

Apêndice D

Letra X

Trata-se de um conjunto composto por 969 amostras extraídas do Machine Learning Repositttory [7]. Cada amostra pertence a um dos 2 tipos de classes: *cluster* 1 composta por 470 amostras e o *cluster* 2 composta por 499 amostras, e cada amostra possui dois atributos ("x" e "y").

Figura D.1: Conjunto de Dados: Letra X



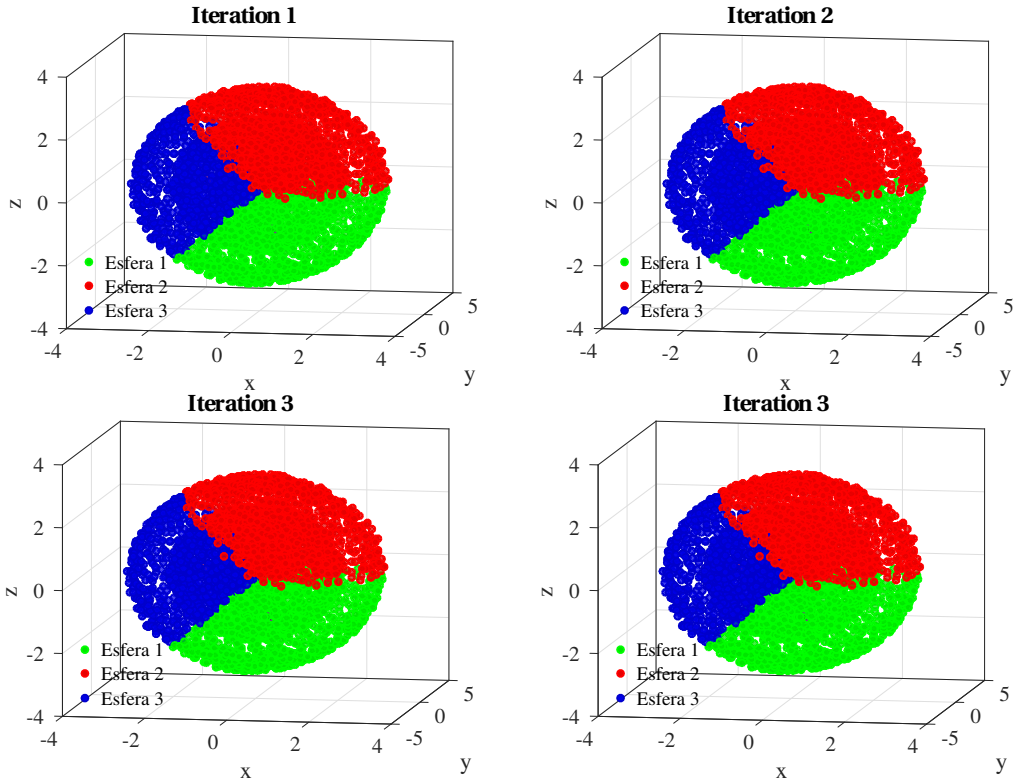
Fonte: Arquivo pessoal

Apêndice E

Esferas

Trata-se de um conjunto composto por 5580 amostras extraídas do Machine Learning Repositttory [7]. Cada amostra pertence a um dos três tipos de classes: Esfera 1, Esfera 2 e Esfera 3. Cada uma das classes é composta por 1860 amostras, e cada amostra possui três atributos ("x", "y", e "z").

Figura E.1: Conjunto de Dados: Esferas



Fonte: Arquivo pessoal