Clustering: k-means e Agglomerative

Jackson Nunes Marco Eugênio Araújo

Outubro de 2014

Sumário

- Contextualização
  - Classificação
  - Agrupamento (Clustering)
  - Cenários de Aplicação
- Clustering
  - Tipos de Clustering (Hierárquico e Particional)
  - Medidas de similaridade
  - Algoritmos Hierárquicos e Não Hierárquicos
- K-means
- Agglomerative
- Práticas envolvendo os dois tipos de Clustering

Classificação de Dados - Histórico

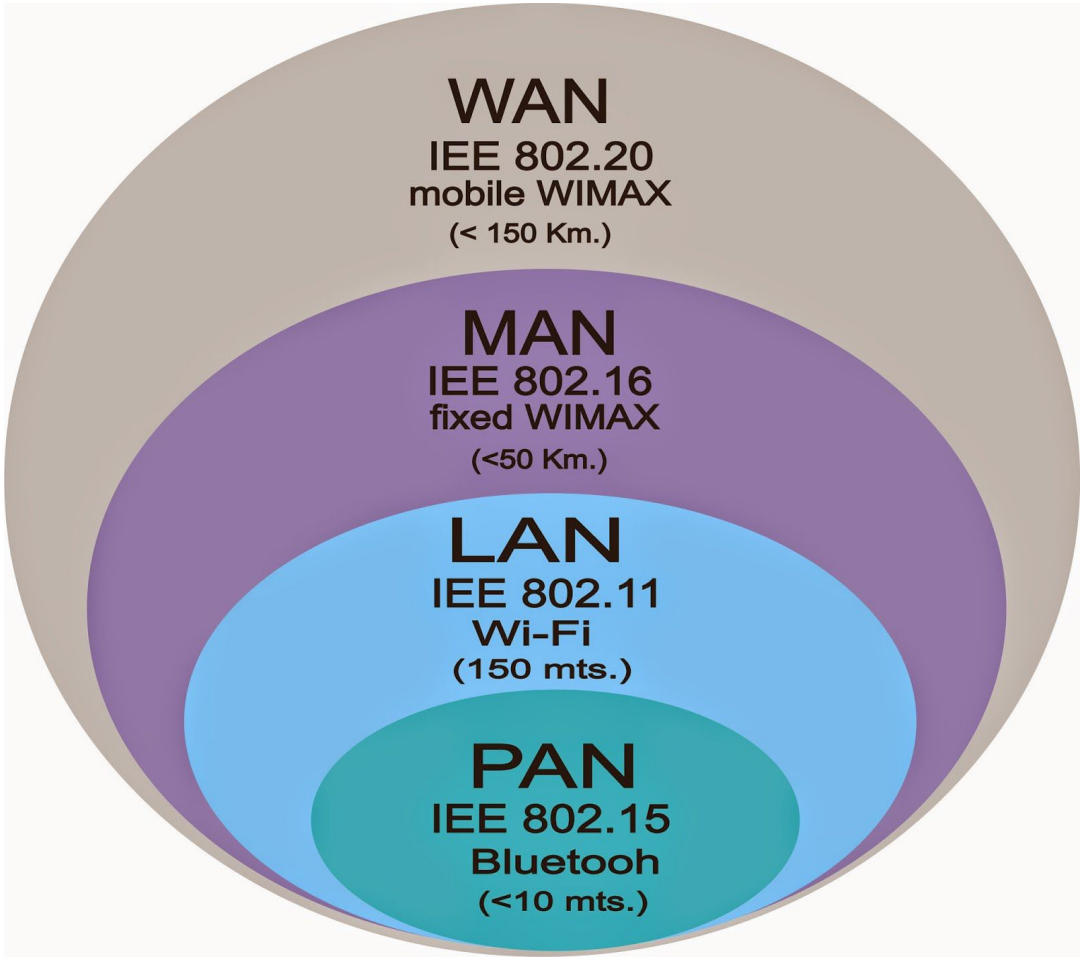
- Desde o homem primitivo, é natural a necessidade de separar coisas/objetos semelhantes em categorias, classificando-os.
- Aristóteles construiu um elaborado sistema para classificar as espécies do reino animal em vertebrados e os invertebrados.
- Theophrastos escreveu as primeiras referências da estrutura e classificação das plantas.

Classificação de Dados - Definições

- A tarefa de classificação consiste em construir um modelo de algum tipo que possa ser aplicado a dados não classificados visando categorizá-los em classes. Um objeto é examinado e classificado de acordo com classes pré-definidas (REZENDE, 2003).

- Classificação, em seu sentido mais amplo, consiste em palavras que nos ajudam a reconhecer e discutir os diferentes tipos de eventos, objetos e pessoas que encontramos.

Classificação de Dados



Por que Classificar?

- A classificação pode representar, simplesmente, um método conveniente para a organização de um grande conjunto de dados de modo que possa ser compreendido mais facilmente e de forma mais eficiente.

- Pode ser útil em pesquisas de mercado, identificando produtos preferenciais ("nichos de mercado")

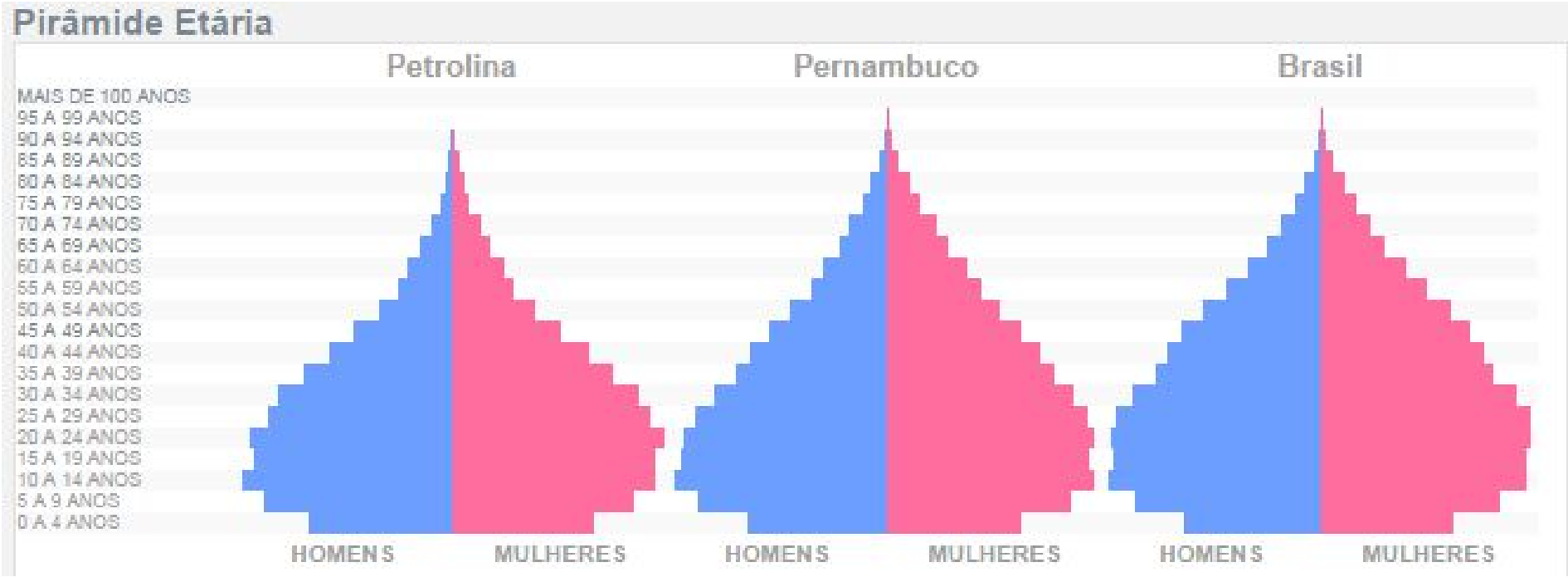
- A necessidade de classificar está presente em várias áreas do conhecimento, como nas ciências biológicas, sociais, na medicina, informática, entre outras.

Classificação de Dados

- Classificação associa elementos em classes que contenham características semelhantes ou comuns.
- Os classificadores podem ser:
  - Supervisionados
  - Não-Supervisionados

Classificação Supervisionada

- Classifica objetos em diferentes categorias
- É baseada em características e número de classes previamente definidas.



Classificação Supervisionada

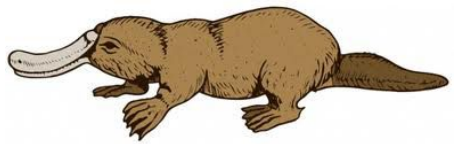
Idade	Petrolina		Pernambuco		Brasil	
	Homens	Mulheres	Homens	Mulheres	Homens	Mulheres
0 a 4 anos	10.549	10.319	277.508	268.115	5.638.154	5.444.151
5 a 9 anos	13.803	13.248	378.324	366.005	7.623.749	7.344.867
10 a 14 anos	15.282	14.737	423.568	411.963	8.724.960	8.440.940
15 a 19 anos	14.391	14.772	407.498	406.100	8.558.497	8.431.641
20 a 24 anos	14.762	15.499	402.836	414.746	8.629.807	8.614.581
25 a 29 anos	13.510	14.407	379.000	400.641	8.460.631	8.643.096
30 a 34 anos	12.752	13.605	344.709	372.344	7.717.365	8.026.554
35 a 39 anos	10.860	11.733	301.541	333.661	6.766.450	7.121.722
40 a 44 anos	9.017	9.909	271.173	305.896	6.320.374	6.688.585
45 a 49 anos	7.205	7.844	233.862	268.313	5.691.791	6.141.128
50 a 54 anos	5.347	5.998	191.000	225.663	4.834.828	5.305.231
55 a 59 anos	3.906	4.516	152.743	190.010	3.902.183	4.373.673
60 a 64 anos	3.240	3.839	128.560	160.049	3.040.897	3.467.956
65 a 69 anos	2.397	2.741	95.597	124.093	2.223.953	2.616.639
70 a 74 anos	1.629	2.086	73.653	100.594	1.667.289	2.074.165
75 a 79 anos	865	1.240	46.054	66.426	1.090.455	1.472.860
80 a 84 anos	557	832	31.232	46.240	668.589	998.311
85 a 89 anos	287	506	16.348	24.574	310.739	508.702
90 a 94 anos	145	231	6.460	11.060	114.961	211.589
95 a 99 anos	41	72	1.870	3.534	31.528	66.804
Mais de 100 anos	10	15	387	1.212	7.245	16.987

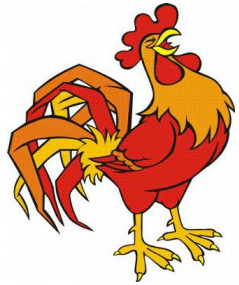
Classificação Não Supervisionada

- Classifica objetos em diferentes categorias através de um algoritmo de análise de agrupamento (Clustering)
- Não há conhecimento prévio das características
- Podem extrair características escondidas dos dados e desenvolver hipóteses a respeito de sua natureza

Classificação Não Supervisionada

- Como agrupar os animais abaixo?

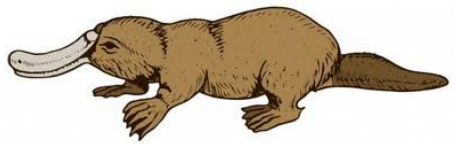


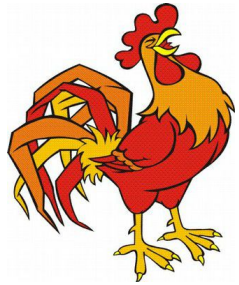


Classificação Não Supervisionada

- Como agrupar os animais abaixo?

Com bico

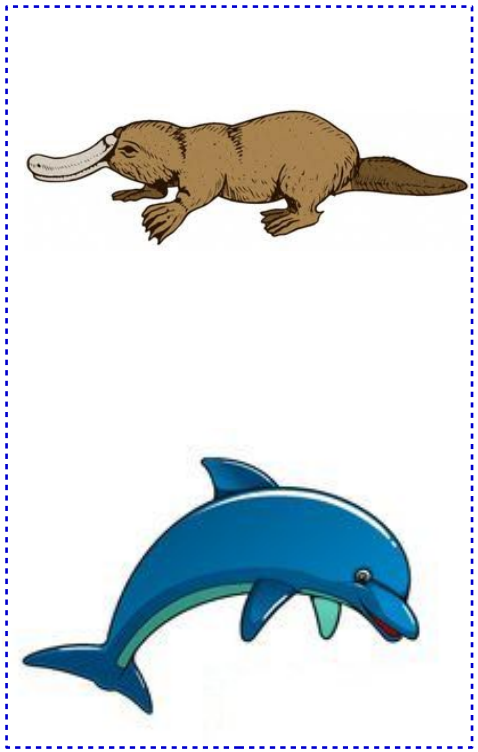




Sem bico

Classificação Não Supervisionada

- Como agrupar os animais abaixo?

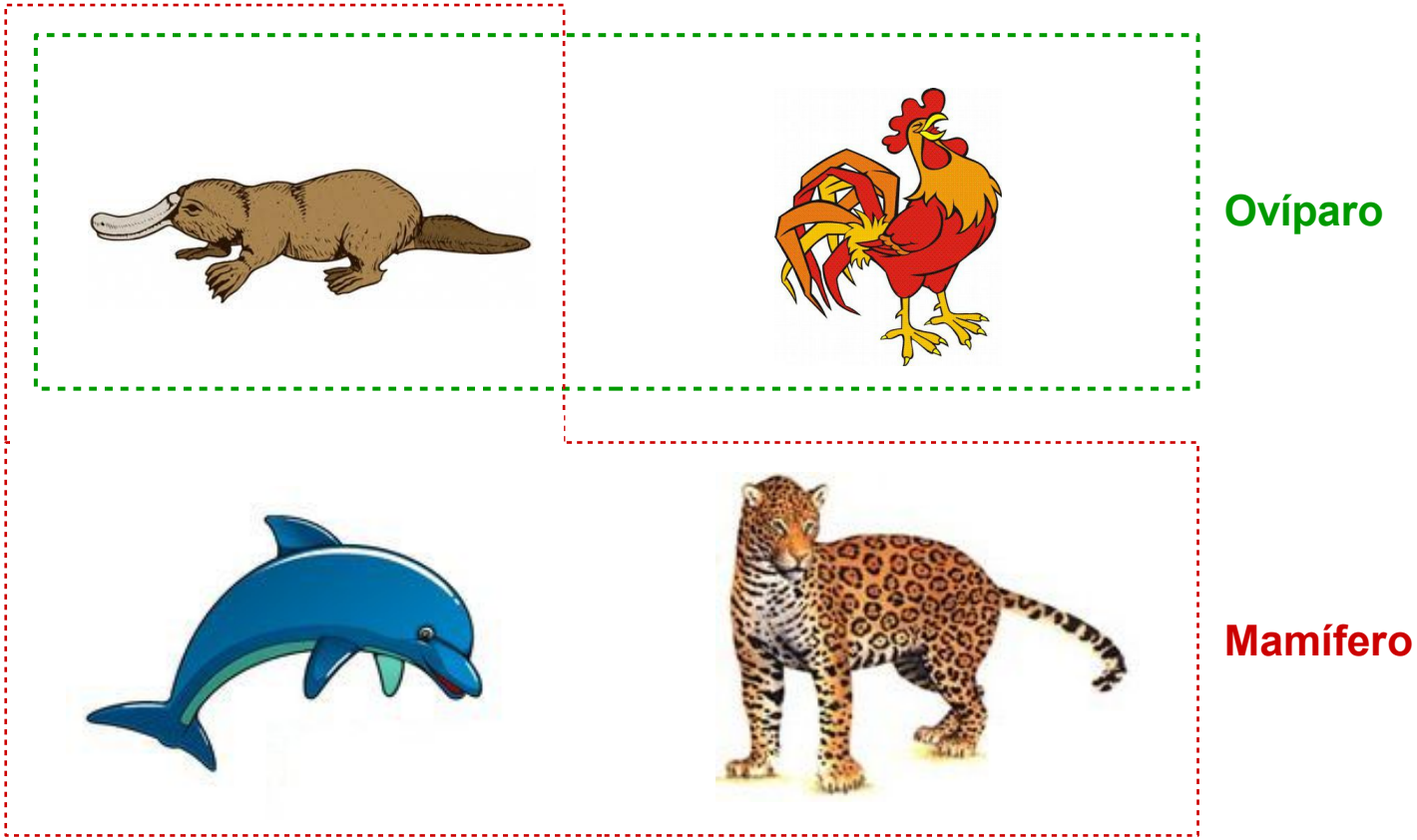


Aquático

Terrestre

Classificação Não Supervisionada

- Como agrupar os animais abaixo?

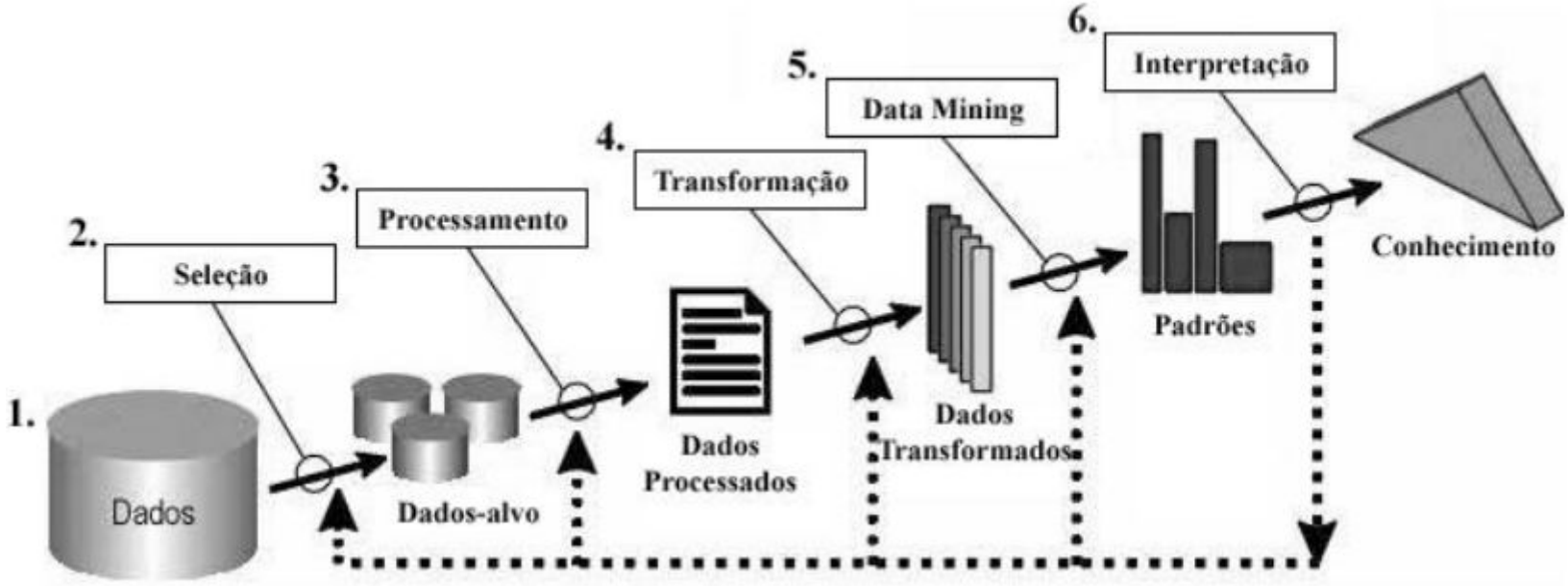


Clustering - Por que analisar agrupamento de dados?

- Crescimento contínuo do tamanho e da complexidade de diversos conjuntos de dados armazenados em banco
- Dificuldade de analisar dados, produzidos e armazenados em larga escala
- Inviabilidade de análise através de métodos tradicionais (planilhas, relatórios informativos operacionais, etc)
- A noção de descoberta de relações úteis a partir de dados gerados e processados

Clustering - Como surgiu?

- Extração de Conhecimento em Bases de Dados (ECBD)
- Data Mining
- Clustering



Clustering (Agrupamento) - Definições

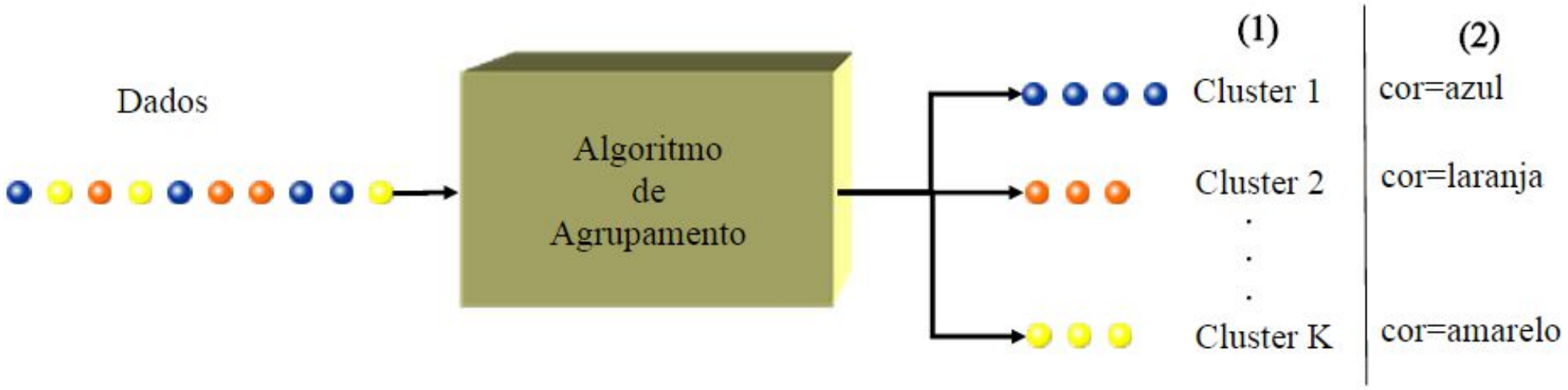
- Análise de agrupamento, ou clustering, é o nome dado para o grupo de técnicas computacionais cujo propósito é separar objetos em grupos, baseando-se nas características que estes possuem.

- A ideia básica consiste em colocar em um mesmo grupo objetos que sejam similares, de acordo com algum critério pré-determinado.

- O objetivo é que os objetos pertencentes a um mesmo grupo sejam similares (relacionados ou próximos) entre si e diferentes (não relacionados ou distantes) dos objetos que compõe os demais grupos.

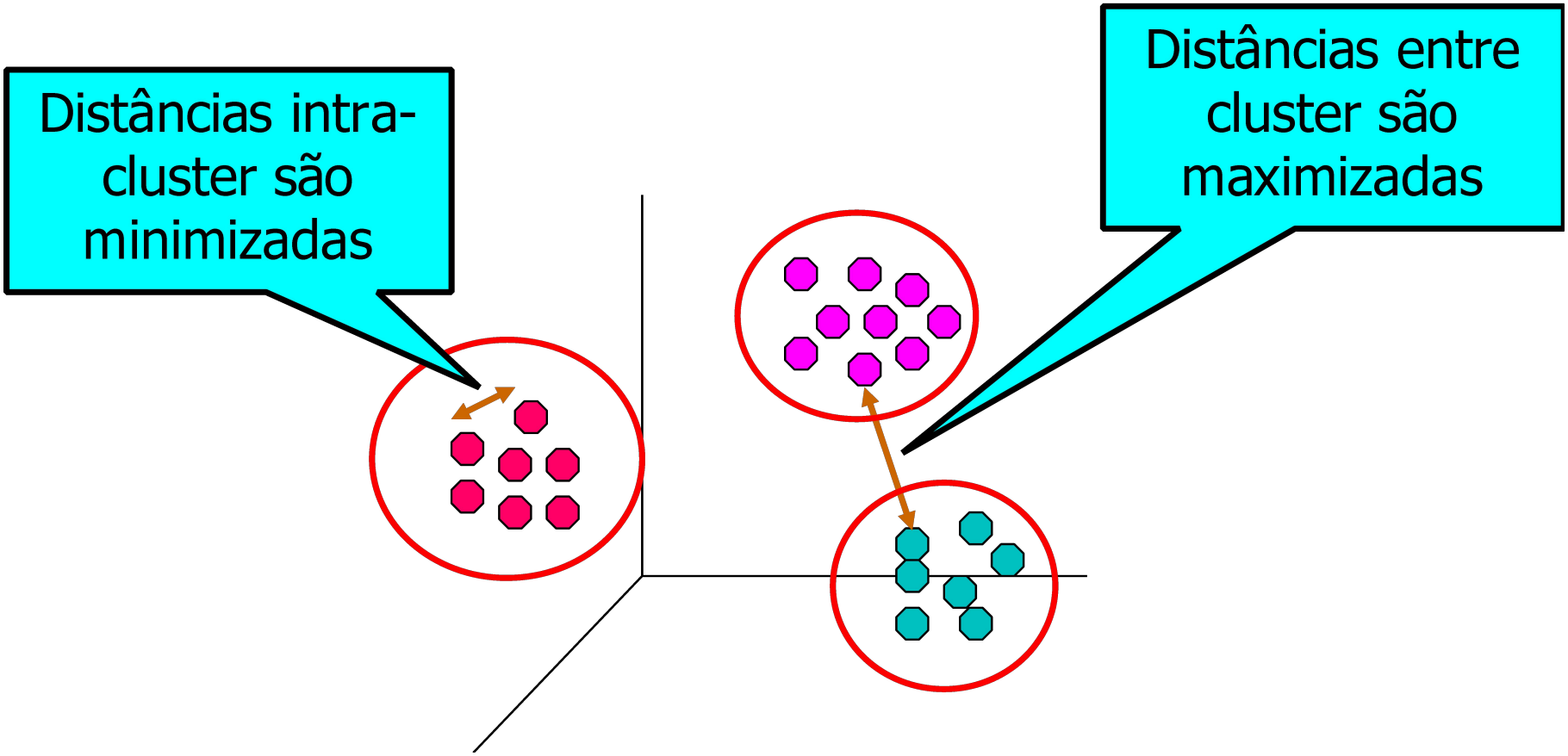
Clustering (Agrupamento)

- Análise de Clustering tem como objetivo:
  - Separar objetivamente grupos homogêneos
  - Maximizar a similaridade de objetos dentro de um mesmo grupo
  - Minimizar a similaridade de objetos entre grupos distintos
  - Atribuir uma descrição para os grupos formados



Clustering (Agrupamento)

- Dado um conjunto de objetos, colocá-los em grupos baseados na similaridade entre eles.



Clustering (Agrupamento)

- A noção de cluster pode ser ambígua

Quantos clusters?

Seis Clusters

Dois Clusters

Quatro Clusters

Clustering (Agrupamento) - Aplicações

- No comércio o clustering pode ajudar a descobrir grupos distintos de clientes e caracterizar estes grupos com base no seu padrão de compras.
- Em biologia pode ser usado para agrupar plantas, ajudar a identificar toxinas, classificar genes pela similaridade das suas funções.
- Classificar e agrupar problemas de saúde pública ou doenças hereditárias.
- Pode também ser uma ajuda na identificação das áreas de terrenos com usos similares, por observação de imagem de satélite.
- Pode ainda ser usado para ajudar a classificar informação e documentos descobertos na Web

Clustering - Características

- A análise de Clustering exige métodos que apresentem as seguintes características:
  - Ser capaz de lidar com dados com alta dimensionalidade;
  - Ser “escalável” com o número de dimensões e com a quantidade de elementos a serem agrupados;
  - Habilidade para lidar com diferentes tipos de dados;
  - Capacidade de definir agrupamentos de diferentes tamanhos e formas;
  - Exigir o mínimo de conhecimento para determinação dos parâmetros de entrada;
  - Ser robusto à presença de ruído;
  - Apresentar resultado consistente independente da ordem em que os dados são apresentados;

Clustering - Características

- É preciso medir a similaridade entre os elementos a serem agrupados.
- A similaridade é expressa como uma função distância ou métrica.

Seja M um conjunto, uma métrica em M é uma função d: $M \times M \rightarrow \Re$, tal que para quaisquer x, y, z $\in M$, tenhamos:

1. $d_{xy} > 0$ - para todo $x \neq y$

2. $d_{xy} = 0 \Leftrightarrow x = y$

3. $d_{xy} = d_{yx}$

Clustering – Medidas de Similaridade

- Medidas de Similaridade para cálculo de distância entre do elementos:
  - Distância Euclidiana
  - Distância Euclidiana Quadrática
  - Distância *Manhattan*
  - Distância de *Chebychev*
- A distância é normalmente representada na forma de matriz
- A Distância Euclidiana é a mais utilizada.

Medidas de Similaridade – Distância Euclidiana

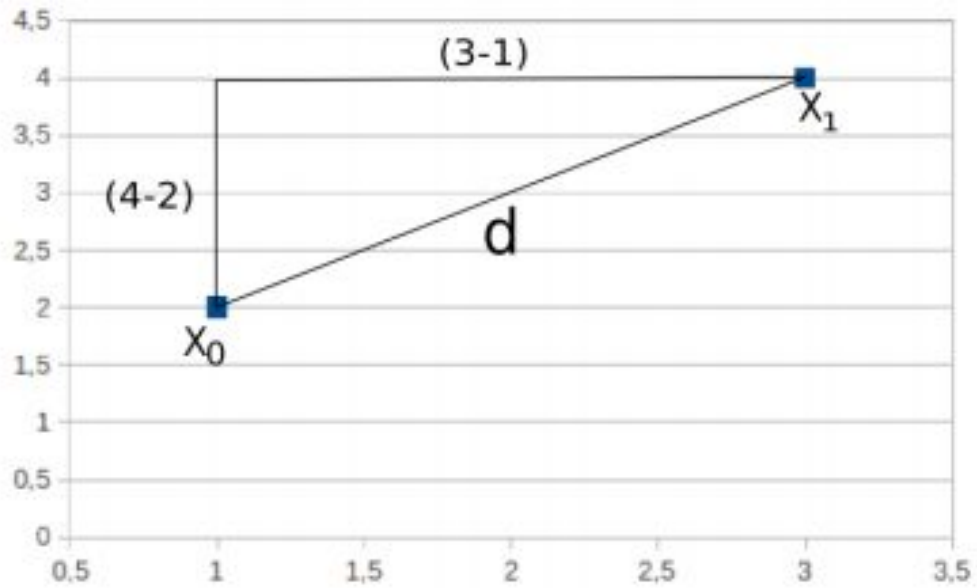
- Distância geométrica no espaço multidimensional.
- A distância euclidiana entre dois elementos

  $X = [X_1, X_2, ..., X_p]$ e $Y = [Y_1, Y_2, ..., Y_p]$, é definida por:

$$d_{xy} = \sqrt{(X1 - Y1)^2 + (X2 - Y2)^2 + ... + (Xp - Yp)^2} = \sqrt{\sum_{i=1}^{p} (Xi - Yi)^2}$$

Medidas de Similaridade – Distância Euclidiana

- Calcular distância entre os elementos X0 = (1,2) e X1 = (3,4)



$$d_{x_0x_1} = \sqrt{(3-1)^2 + (4-2)^2} = \sqrt{8} = 2,83$$

Medidas de Similaridade – Distância Euclidiana Quadrática

- A distância euclidiana quadrática é definida pela expressão:

$$d_{xy} = (X_1 - Y_1)^2 + (X_2 - Y_2)^2 + \ldots (X_p - Y_p)^2 = \sum_{i=1}^{p} (X_i - Y_i)^2$$

- Considerando-se os mesmos pontos X0 e X1 do exemplo anterior, observa-se a intensificação da distância:

$$d_{x_0 x_1} = (3-1)^2 + (4-2)^2 = 8$$

Medidas de Similaridade – Distância *Manhattan*

- A distância de Manhattan é definida pela expressão:

$$d_{xy} = |X_1 - Y_1| + |X_2 - Y_2| + \ldots |X_p - Y_p| = \sum_{i=1}^{p} |X_i - Y_i|$$

- Em muitos casos, a distância *Manhattan* apresenta resultados similares ao da Euclidiana
- Porém, utilizando os dados do exemplo anterior:

$$d_{x_0 x_1} = |3-1| + |4-2| = |2| + |2| = 4$$

Medidas de Similaridade – Distância *Chebychev*

- A distância de Chebychev é apropriada no caso em que se deseja definir dois elementos como diferentes, se apenas umas das dimensões é diferente.
- Ela é definida por:

$$d_{xy} = maximo(|X_1 - Y_1| + |X_2 - Y_2| + ...|X_p - Y_p|)$$

- Utilizando os pontos X_2= (9,2) e X_3 = (2,5), tem-se:

$$d_{x_2x_3} = máximo(|9-2|,|2-5|) = (|7|,|-3|) = 7$$

Metriz de Similaridade – Exemplo

- Considerando os elementos da tabela, obtemos a matriz de similaridade.

		Benchmars (/1000)	
CPU MARK		**High End CPU's**	**Overclocked CPU's**
Intel Xeon E5-2690 v2 @ 3.00GHz	**A**	17,30	18,05
Intel Xeon E5-2660 v3 @ 2.60GHz	**B**	16,67	17,39
Intel Xeon E5-2687W @ 3.10GHz	**C**	14,54	15,27
Intel Xeon E5-1650 v2 @ 3.50GHz	**D**	12,49	13,83
Intel Xeon E5-2630 v3 @ 2.40GHz	**E**	13,62	13,63
Intel Xeon E3-1280 v3 @ 3.60GHz	**F**	10,42	11,44

Matriz de Similaridade – Exemplo

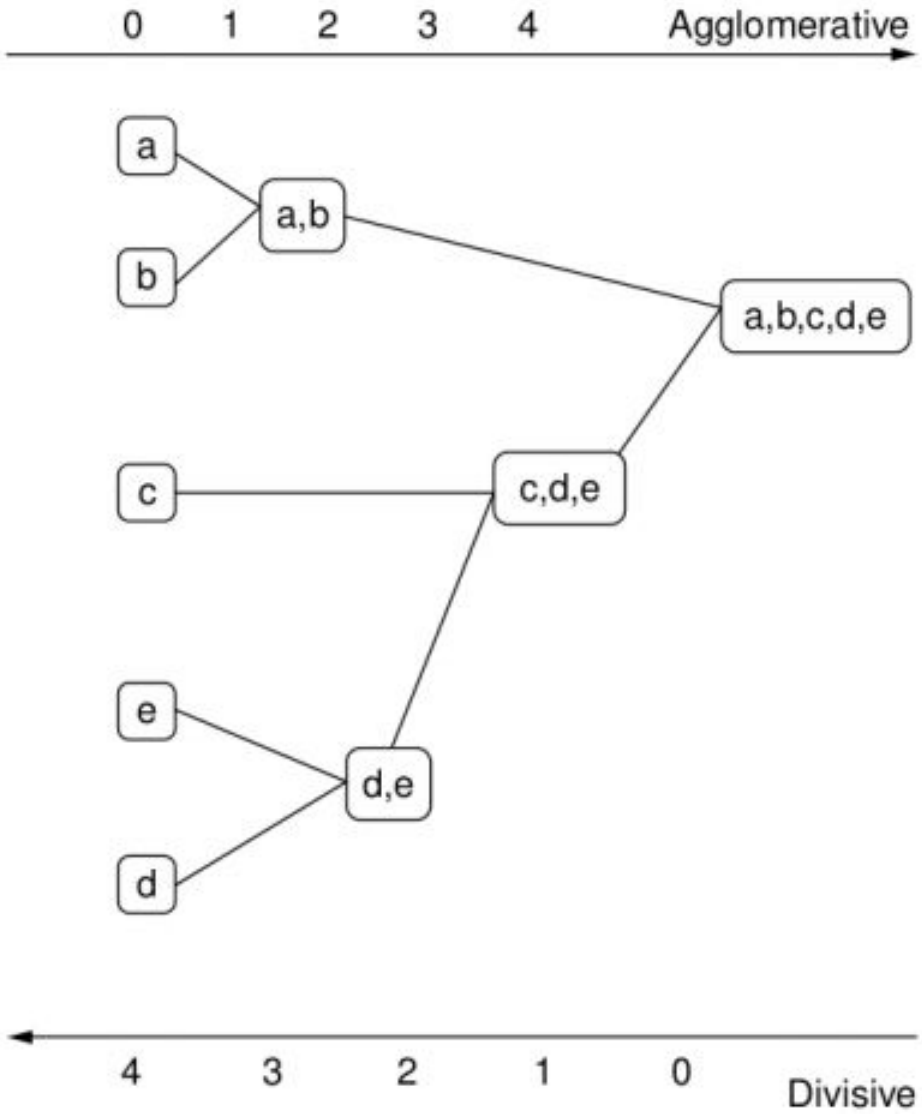
- Matriz de similaridade obtida através da aplicação da Distância Euclidiana.

TABELA DE SIMILARIDADE: Distância Euclidiana						
	A	**B**	**C**	**D**	**E**	**F**
A	0,00	0,92	3,92	6,40	5,76	9,55
B	0,92	0,00	3,00	5,48	4,84	8,63
C	3,92	3,00	0,00	2,50	1,88	5,63
D	6,40	5,48	2,50	0,00	1,14	3,17
E	5,76	4,84	1,88	1,14	0,00	3,88
F	9,55	8,63	5,63	3,17	3,88	0,00

Clustering - Métodos Hierárquicos

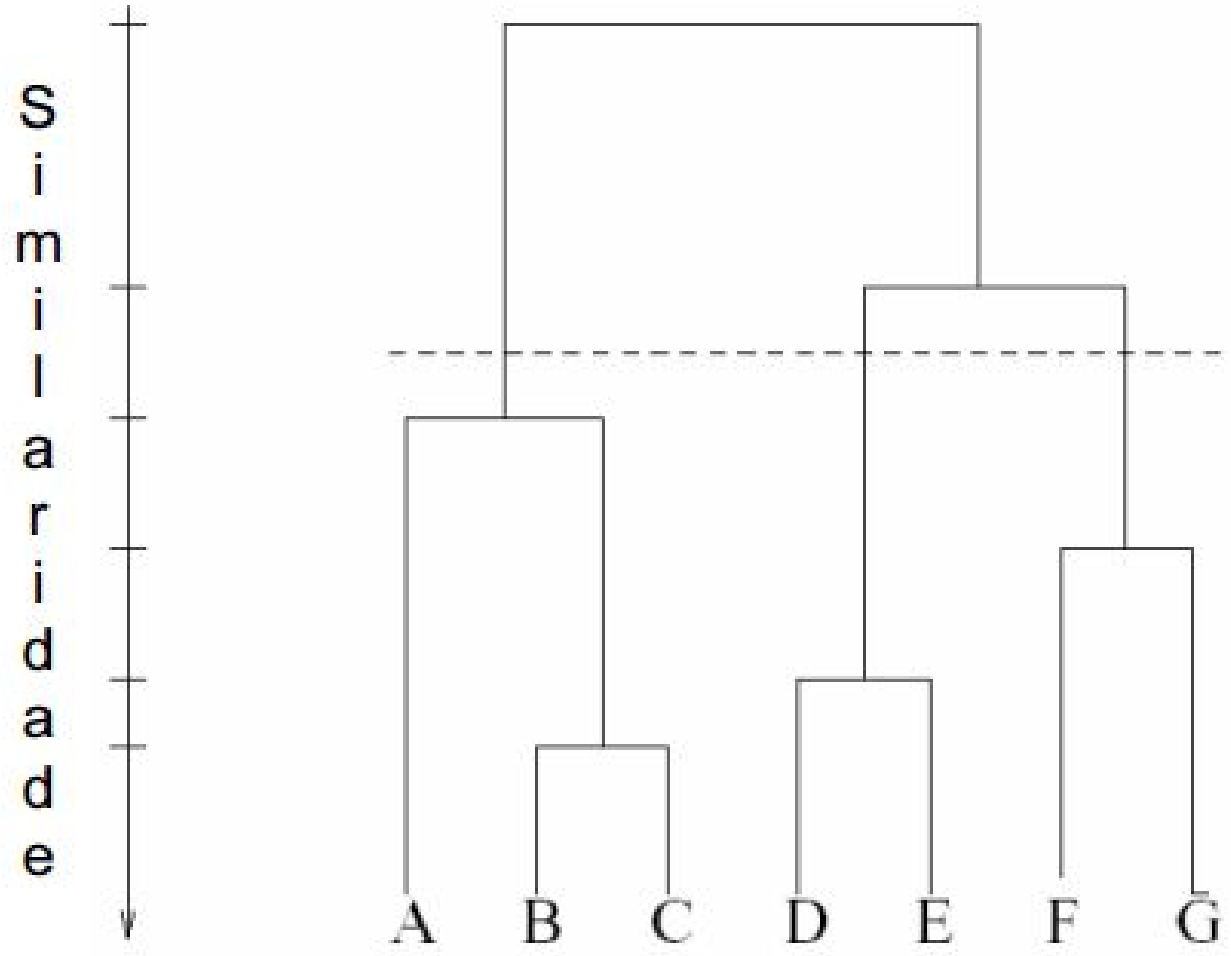
- Agrupamentos sucessivos ou divisões de elementos, que são agregados ou desagregados.
- Criam uma decomposição hierárquica de um dado conjunto de objetos
- São subdivididos em métodos aglomerativos e divisivos.
- Após executado o método hierárquico (divisão ou junção), a operação não pode ser desfeita.

Clustering - Métodos Hierárquicos



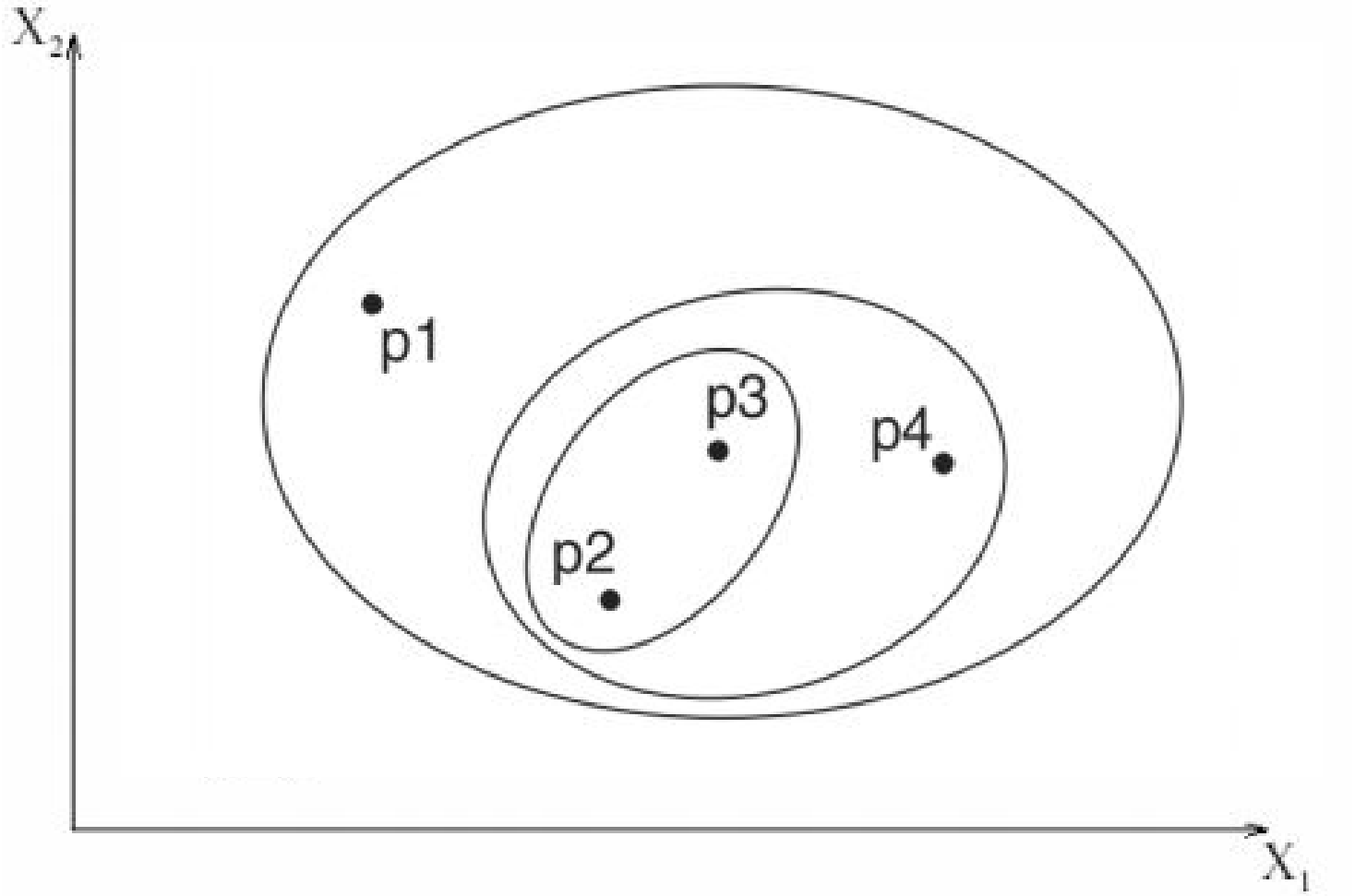
Métodos Hierárquicos - Dendograma

Representação bi-dimensional, também chamado de diagrama de árvore.



Métodos Hierárquicos - Diagrama

Diagrama que representa os clusters aninhados com agrupamento de padrões, níveis de semelhança (proximidade).



Clustering: Agglomerative

- Cada elemento inicia-se representando um grupo, e a cada passo, um grupo ou elemento é ligado a outro de acordo com sua similaridade.
- No último passo, é formado um grupo único com todos os elementos.
- Existem três grandes conceitos de agrupamento aglomerativo:
  - Métodos de ligação (single linkage, complete linkage, average linkage, median linkage);
  - Métodos de centróide;
  - Métodos de soma de erros quadráticos ou variância (método de Ward).

Clustering: Agglomerative

- De modo geral, os métodos aglomerativos utilizam os passos do um algoritmo padrão a seguir:

Entrada: Uma base de dados com ***N*** elementos.
Saída: Um conjunto de grupos.
1. Iniciar com ***N*** grupos, contendo um elemento em cada grupo e uma matriz de similaridade $D_{N \times N}$;
2. Repetir;
3. Localizar a menor distância d_{UV} (maior similaridade);
4. Atualizar a matriz ***D***, retirando os elementos ***U*** e ***V***;
5. Atualizar a matriz ***D***, adicionando as novas distâncias do grupo ***(U, V)***;
6. Até ***N*-1**, quando todos elementos estarão em um único grupo.

Clustering: Agglomerative – Single linkage

- O método de ligação por vizinho mais próximo emprega a distância de valor mínimo.

$$d_{(UV)W} = \min(d_{UW}, d_{VW})$$

- Apresenta bons resultados tanto para distâncias Euclidianas quanto para outras distâncias.
- Tendência a formar longas cadeias (encadeamento).
- Com uma cadeia longa, torna-se difícil definir um nível de corte para classificar os elementos em grupos

Clustering: Agglomerative – Complete linkage

- Nesse método, é empregada a distância máxima, dada por:

$$d_{(UV)W} = \max(d_{UW}, d_{VW})$$

- Apresenta bons resultados tanto para distâncias Euclidianas quanto para outras distâncias.
- Tendência a formar grupos compactos.

Clustering: Agglomerative – Demias métodos

- Average Linkage (ligação por média):

$$d_{(UV)W} = \frac{(N_u.d_{UW} + N_v.d_{VW})}{N_u + N_v}$$

- Centroid Linkage (ligação por centróide):

$$d_{(UV)W} = \frac{N_U.d_{UW} + N_V.d_{VW}}{N_U + N_V} - \frac{N_U.N_V.d_{UV}}{(N_U + N_V)^2}$$

Clustering: Agglomerative – Exercício 1

		Benchmarks (/1000)			
CPU MARK		**High End CPU's**	**Overclocked CPU's**	**Single Thread Performance**	**Power Performance (/100)**
Intel Xeon E5-2690 v2 @ 3.00GHz	**A**	17,304	18,051	1,819	1,33
Intel Xeon E5-2660 v3 @ 2.60GHz	**B**	16,666	17,389	1,883	1,59
Intel Xeon E5-2687W @ 3.10GHz	**C**	14,538	15,271	1,863	1,49
Intel Xeon E5-1650 v2 @ 3.50GHz	**D**	12,492	13,831	1,939	1,45
Intel Xeon E5-2630 v3 @ 2.40GHz	**E**	13,618	13,629	1,896	1,6
Intel Xeon E3-1280 v3 @ 3.60GHz	**F**	10,42	11,437	2,337	1,27

Métodos não Hierárquicos ou Particionados

- A ideia central é escolher uma partição inicial dos elementos e, em seguida, alterar os membros dos grupos para obter-se o melhor particionamento (ANDERBERG, 1973)
- Quando comparado com o método hierárquico, este método é mais rápido, pois não é necessário calcular e armazenar, durante o processamento, a matriz de similaridade

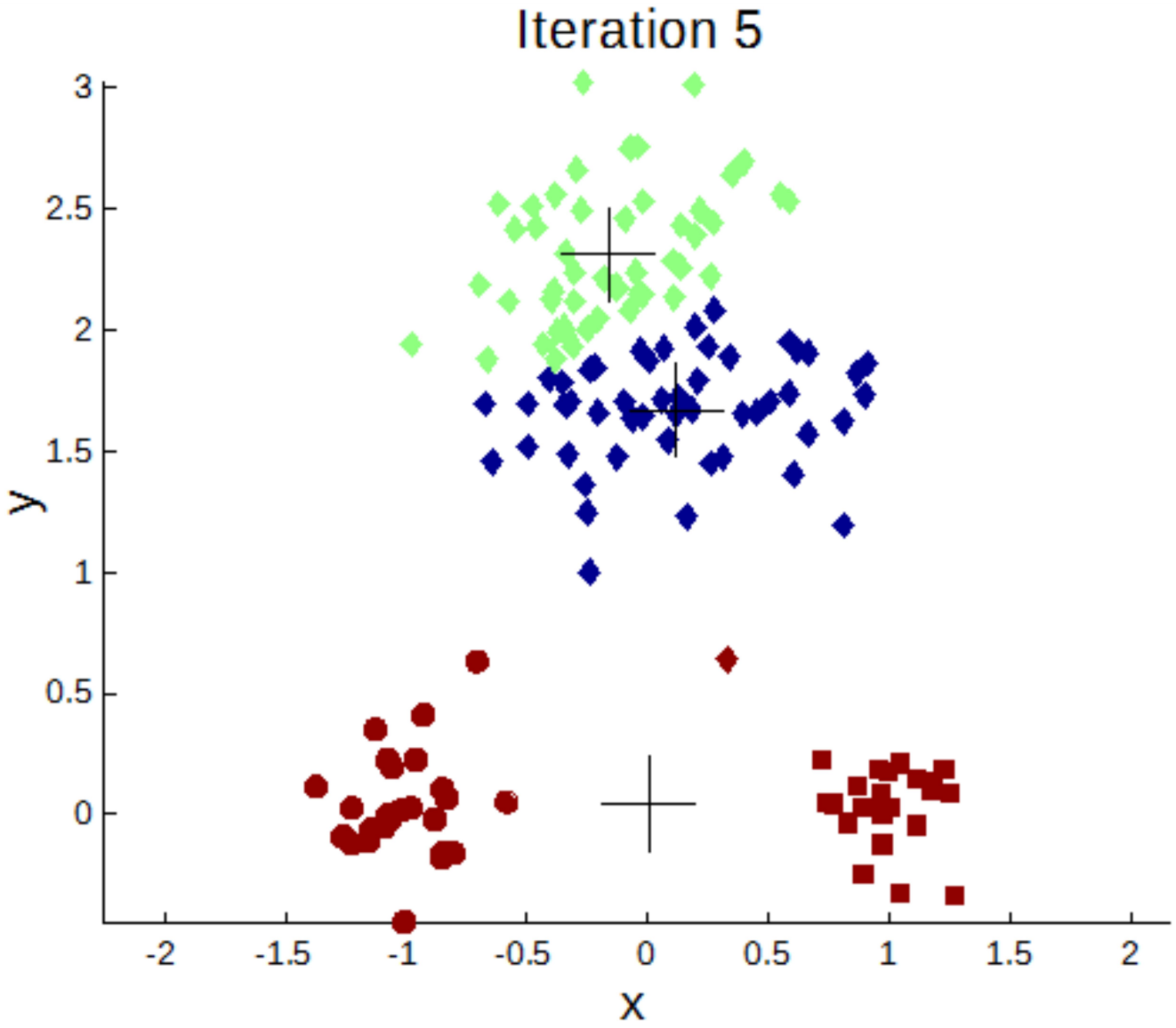
Algoritmo não Hierárquico – *K-means*

- A ideia é fornecer uma classificação de informações de acordo com os próprios dados (analisar e comparar)
- O método *k-means* toma um parâmetro de entrada, K, e particiona um conjunto de N elementos em K grupos
- Começa com uma partição inicial aleatória e continua atribuindo aos *clusters* novos padrões com base na similaridade entre o padrão e o *cluster* até que um critério de convergência conhecido seja atingido

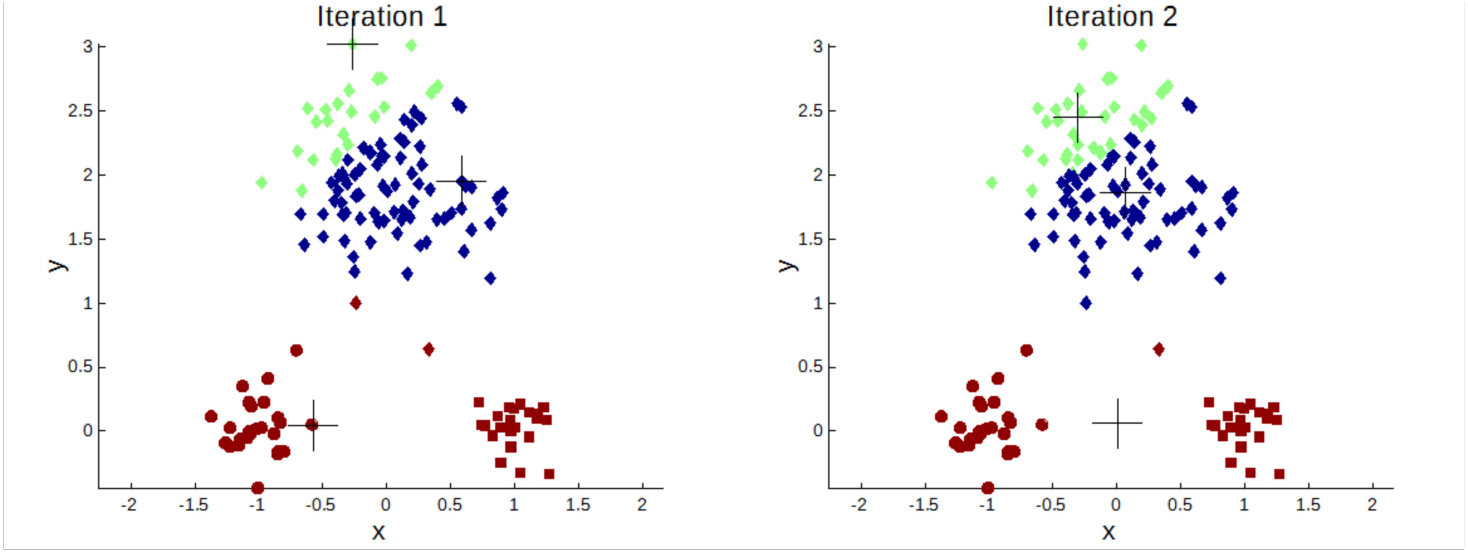
Algoritmo não Hierárquico – *K-means*

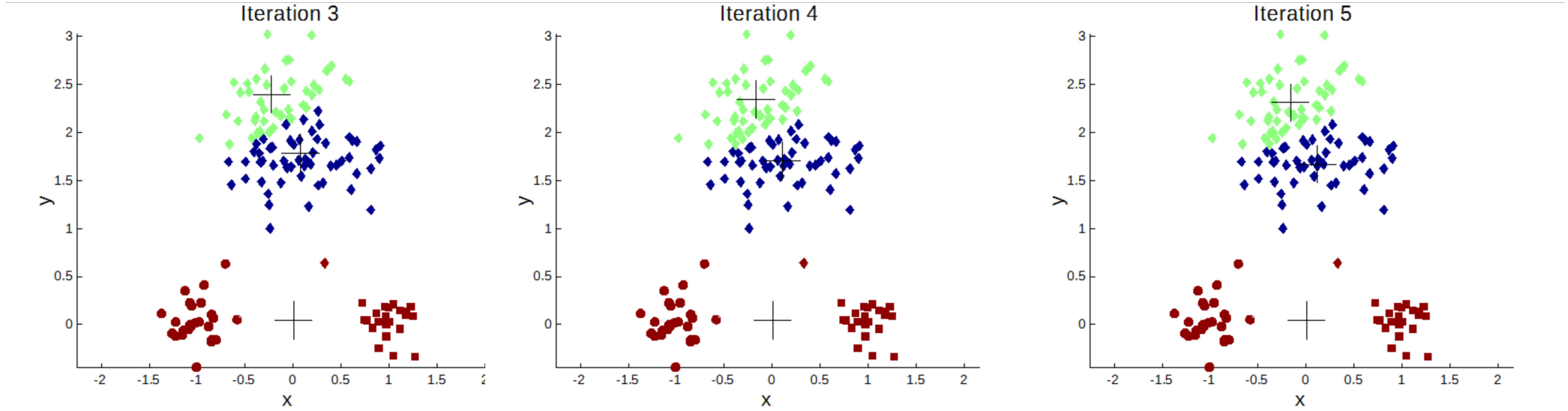
Entrada: O número de grupos, ***K***, e a base de dados com ***N*** elementos.
Saída: Um conjunto de ***K*** grupos.
1. Escolher arbitrariamente ***K*** elementos da base de dados como os centros iniciais dos grupos;
2. Repetir;
3. (re)Atribua cada elemento ao grupo ao qual o elemento é mais similar, de acordo com o valor médio dos elementos no grupo;
4. Atualizar as médias dos grupos, calculando o valor médio dos elementos para cada grupo;
5. Até que não haja mudanças de elementos de um grupo para outro.

K-means – Escolha dos Centróides Iniciais



K-means – Escolha dos Centróides Iniciais

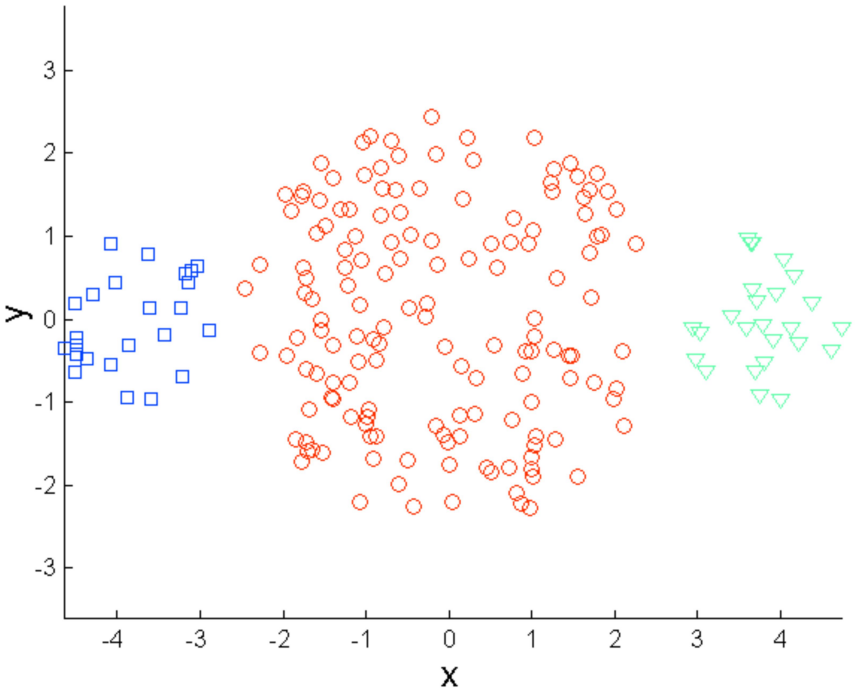




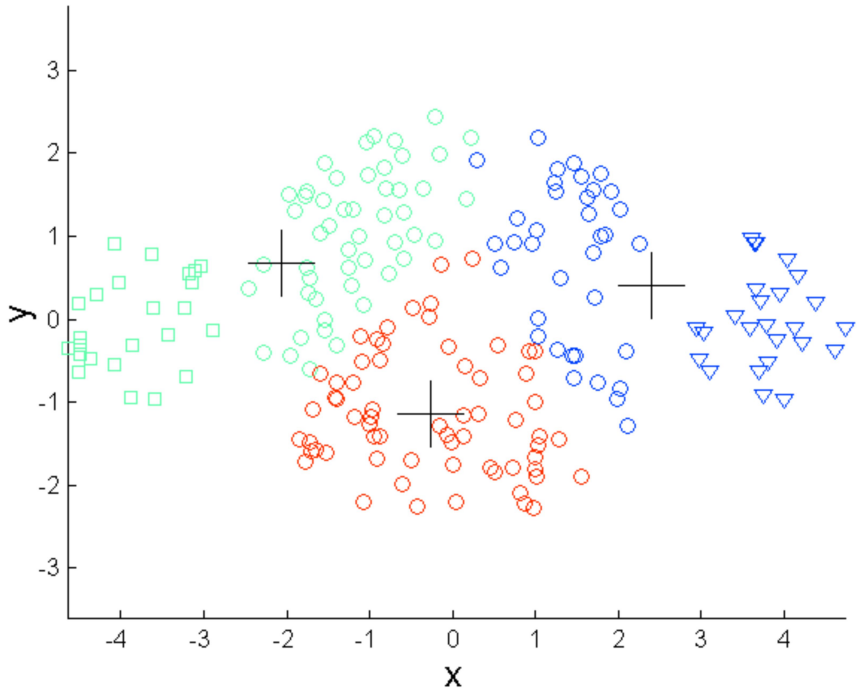
K-means – Limitações

- K-means possui limitações quando os clusters tem as seguintes características:
  - Tamanhos diferentes
  - Densidades diferentes
  - Formato não esférico

K-means – Limitações: tamanhos diferentes

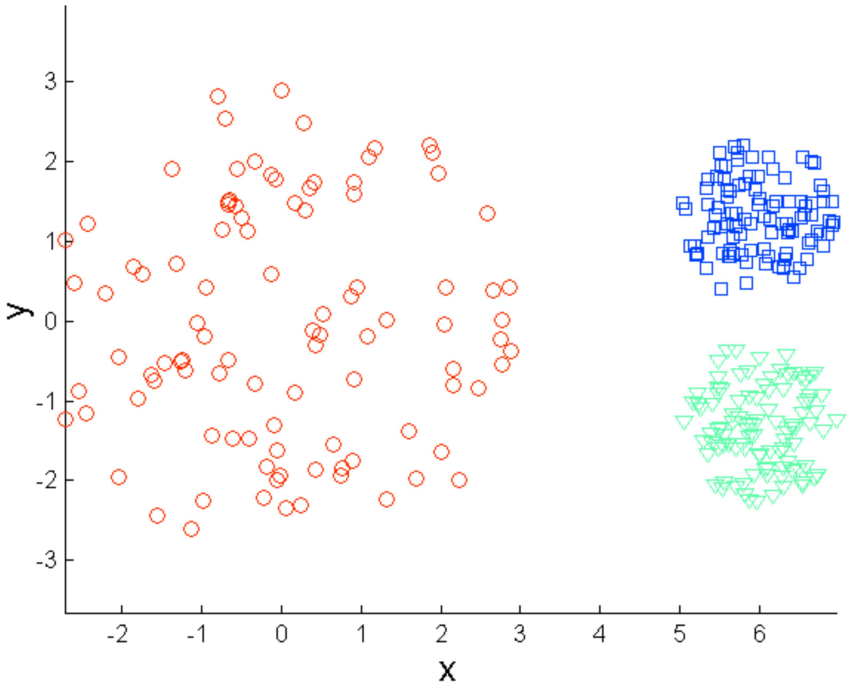


Pontos originais

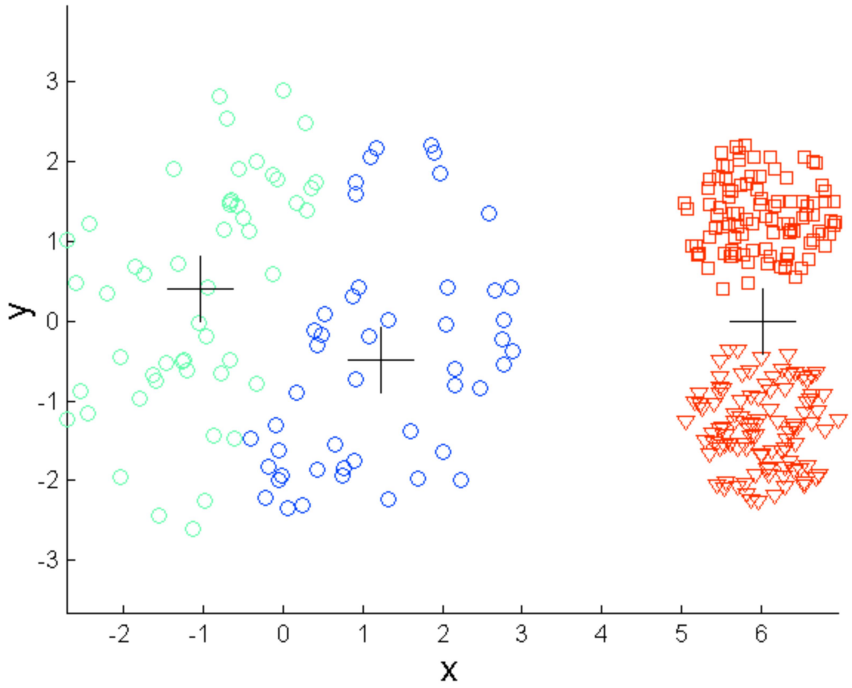


K-médias (3 Clusters)

K-means – Limitações: densidades diferentes

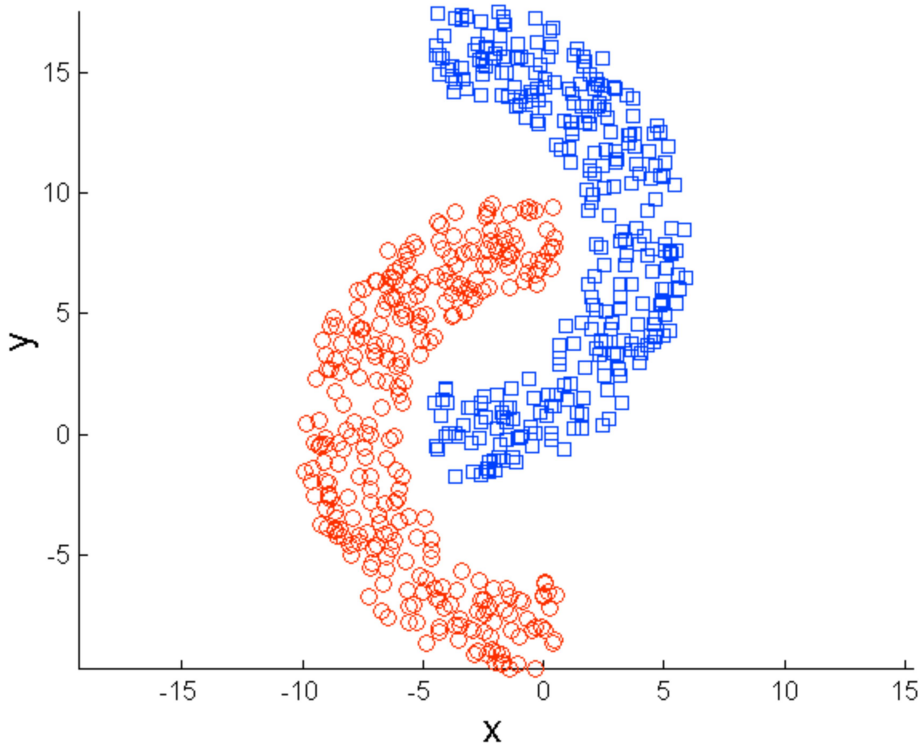


pontos originais

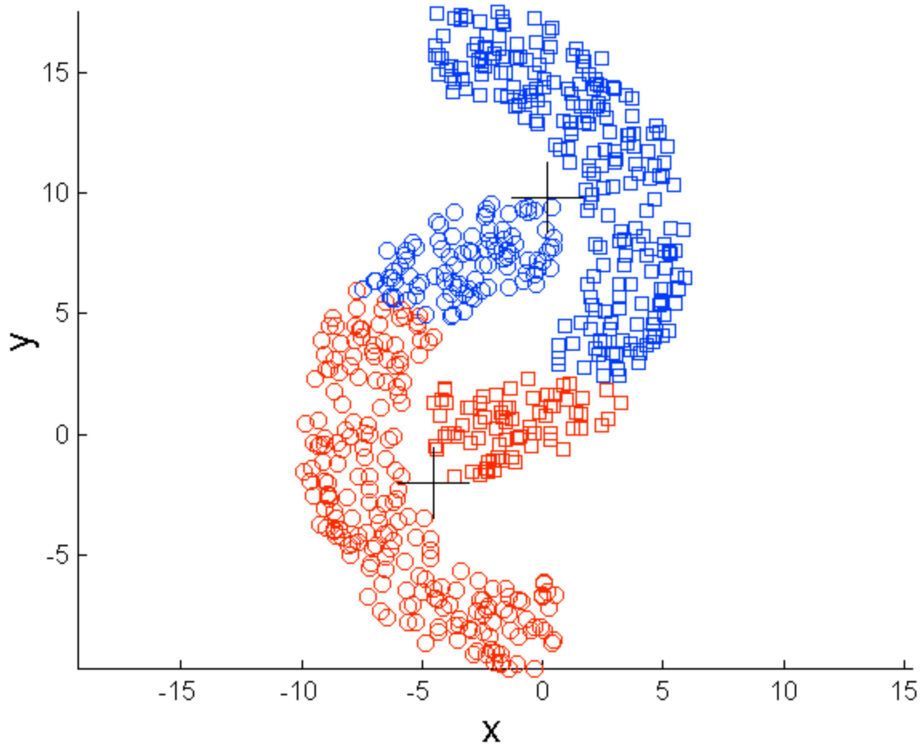


K-médias (3 Clusters)

K-means – Limitações: formatos não esféricos



Pontos originais



K-médias (2 Clusters)

K-means – Exemplo com Data Mining

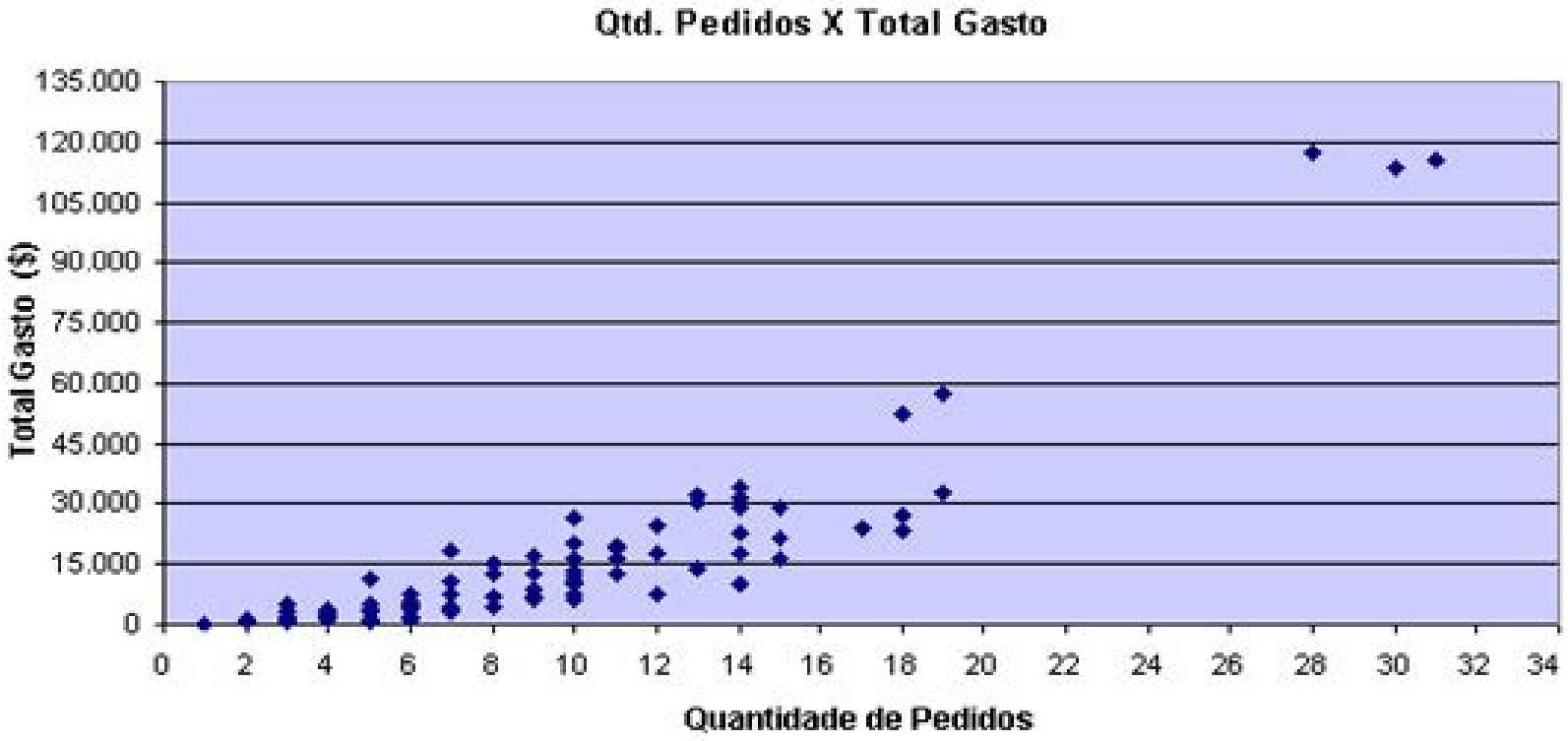
- Vamos considerar que uma determinada empresa vende produtos para clientes por meio de pedidos compostos por itens.
- Com base neste modelo, o departamento de marketing da empresa deseja segmentar os clientes para poder oferecer descontos diferenciados e outros benefícios
- A segmentação dos clientes deve dividir todos os clientes da base de dados em três categorias: **Clientes Ouro, Clientes Prata e Clientes Bronze**

K-means – Exemplo com Data Mining

- Variáveis de classificação:

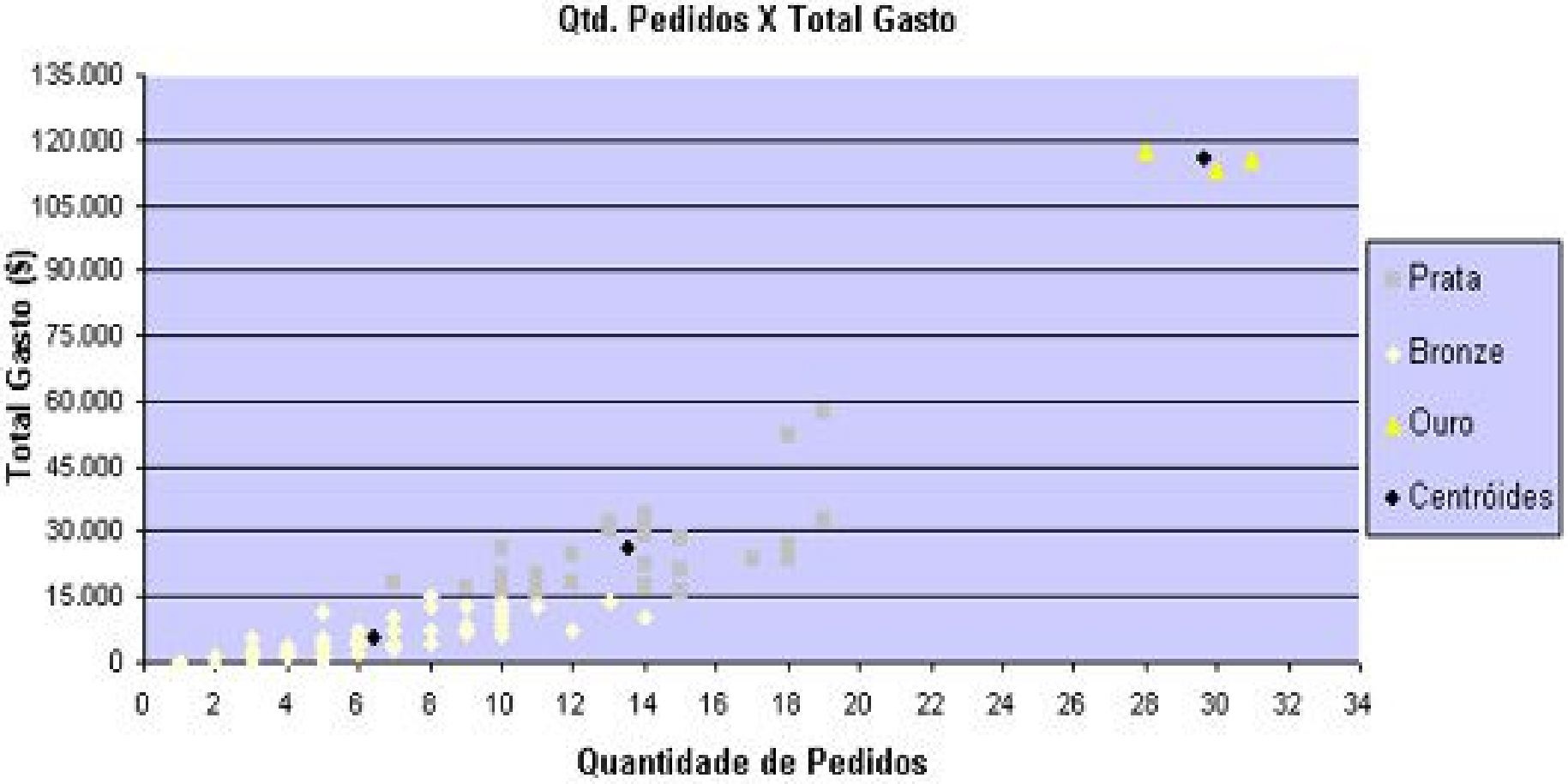
1. Quantidade de pedidos
2. Total gasto ($)

- *K* = *3* (Cliente Ouro, Prata e Bronze)



K-means – Exemplo com Data Mining

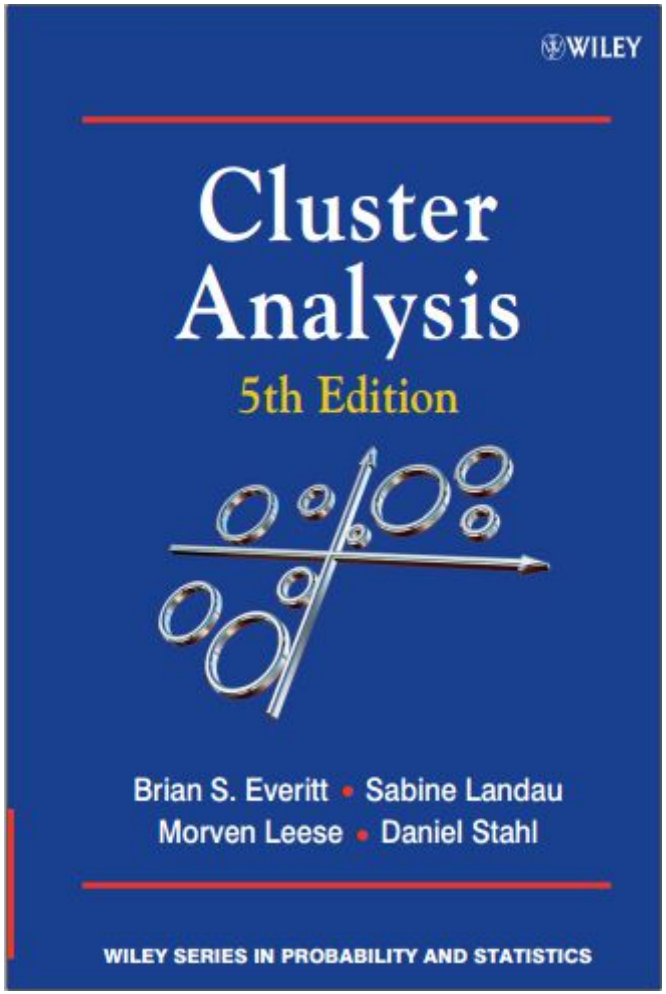
- Com a utilização do algoritmo pode-se classificar os clientes existentes da maneira que o departamento de marketing desejou



K-means – Exercício 2

- Utilizar a tabela de *Benchmark* do Exercício 1

Referências



Perguntas?

Obrigado!

Jackson Nunes
jns@cin.ufpe.br

Marco Eugênio Araújo
maea@cin.ufpe.br