UNIDADE 4
Aprendizado de Máquina
Agrupamento

Disciplina: Tópicos Especiais III (DEC7553)

Prof. Alexandre L. Gonçalves

E-mail: a.l.goncalves@ufsc.br

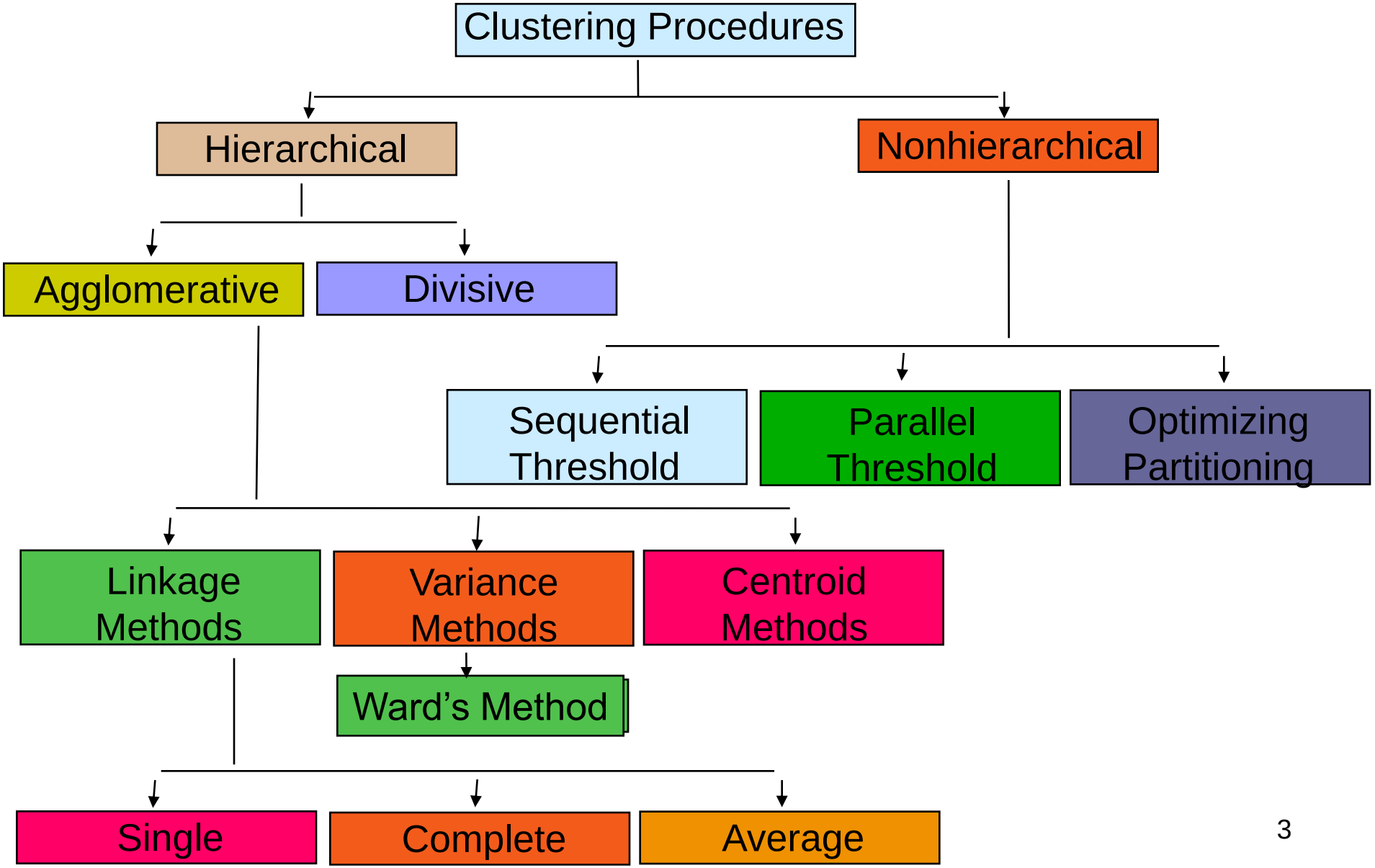
Agrupamentos

Introdução: Principais conceitos e dificuldades da tarefa de agrupamento;

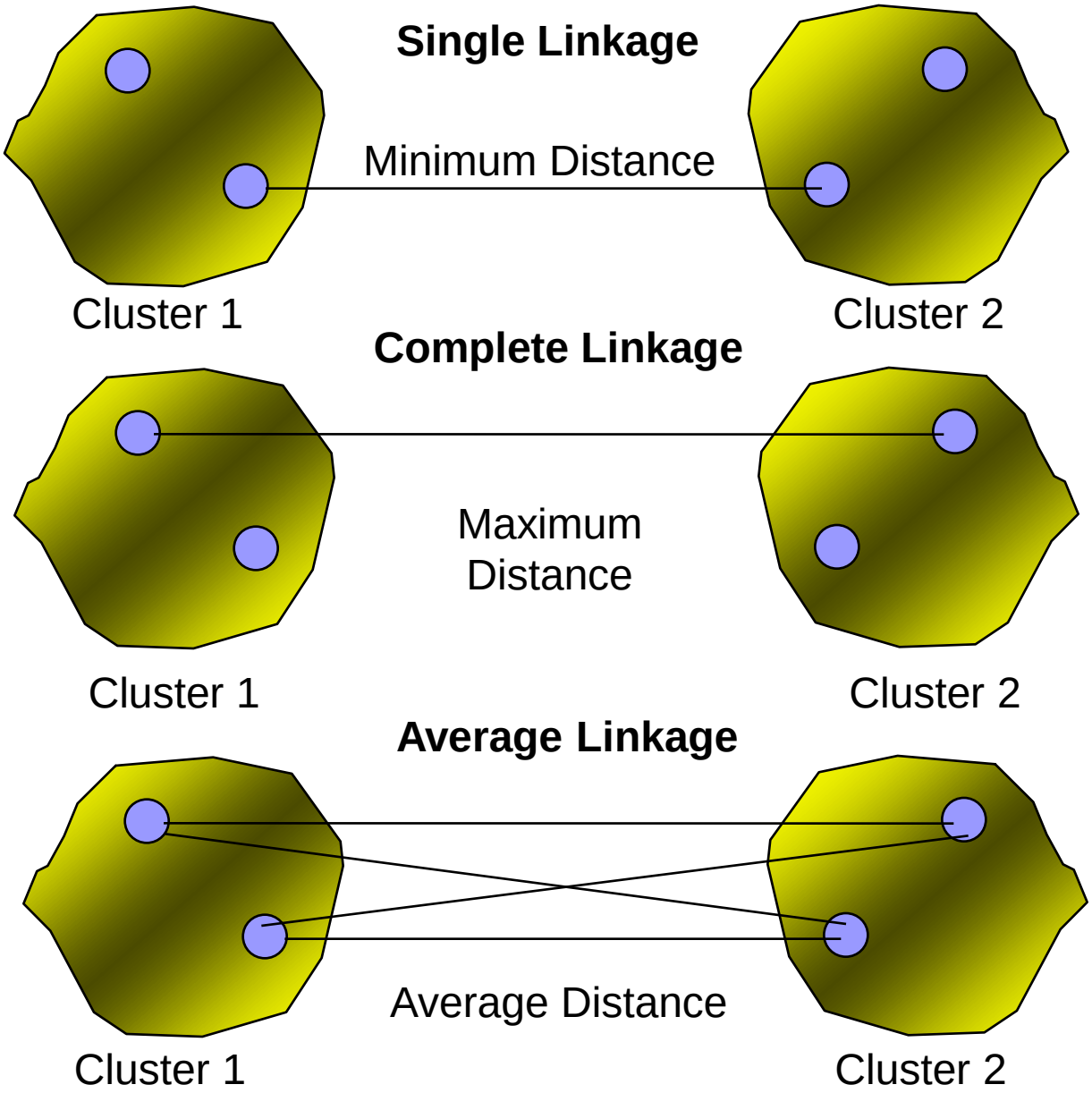
Algoritmos por particionamentos: Medidas de Similaridade, algoritmo *k*-means;

Algoritmos hierárquicos: método aglomerativo, dendrograma;

Classificação dos Procedimentos de Clustering



Método Linkage de Clustering



Algoritmos

Uma forma de **classificar** os **algoritmos** de **Clustering** é através de :

- Métodos por particionamento;
- Métodos hierárquicos;
- Métodos baseados em densidade;
- Métodos baseados em grades;
- Métodos baseados em modelos (Graph-based).

Conceitos

Conjunto de métodos usados para a construção de **grupos de objetos** com base nas **semelhanças** e **diferenças** entre os mesmos, de tal maneira que, os grupos obtidos sejam os mais homogêneos e bem separados possíveis.

A **Clusterização** é uma tarefa prévia à classificação. Sem classes, não se pode determinar a pertinência de um objeto em determinado contexto.

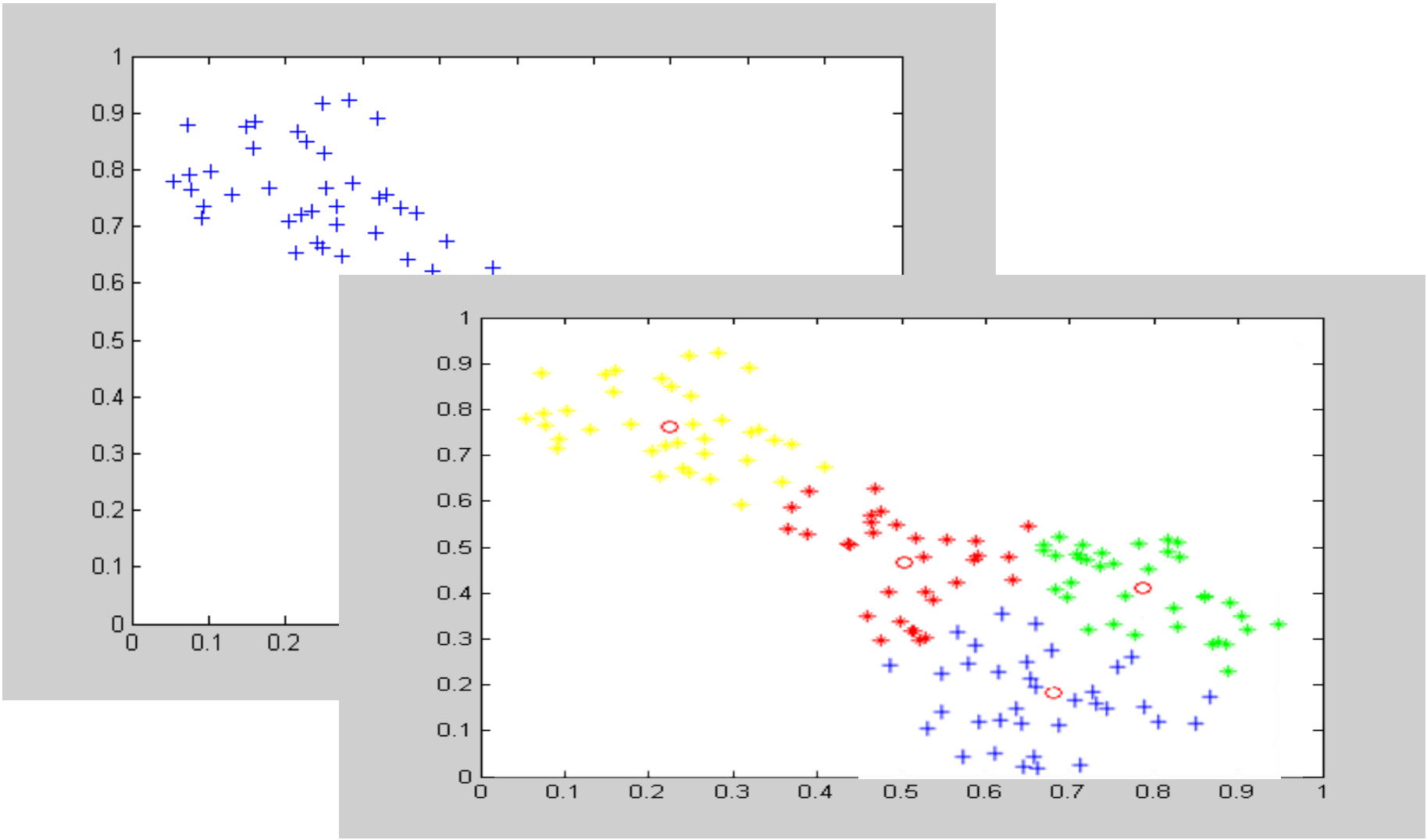
Com a existência das classes, ao se receber um **novo** objeto que pertença ao universo considerado, pode-se classificá-lo corretamente.

Conceitos

O problema de **Clustering** é descrito como:
Tendo um conjunto de dados, de objetos, tentar agrupar este conjunto de forma que os elementos que compõem cada grupo sejam mais parecidos entre si do que parecidos com os elementos dos outros grupos.

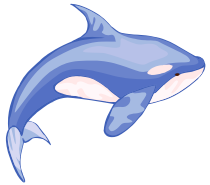
Em resumo, é colocar os iguais (ou quase iguais) juntos num mesmo grupo e os desiguais em grupos distintos.

Desafios



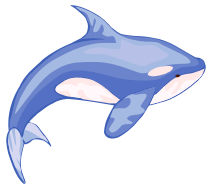
- Como agrupar os seguintes animais?



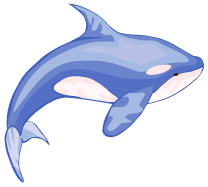





Com bico



Sem bico

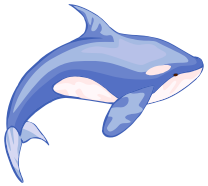


Água



Terra



Ovíparo



Mamífero

Desafios

Como **medir** a **similaridade** entre os itens? (como qualificar os itens, como trabalhar com dados **categóricos** e **numéricos**)

Como formar os **agrupamentos**?
(que variáveis fazem parte da geração dos agrupamentos)

Quantos grupos devem ser formados?
(como definir o número de agrupamentos, ou o raio de abrangência do agrupamento).

Dificuldades

Encontrar o melhor agrupamento para um conjunto de objetos não é uma tarefa simples, a não ser que ***n*** (número de objetos) e ***k*** (número de clusters) sejam extremamente pequenos, visto que o número de partições distintas em que podemos dividir ***n*** objetos em ***k*** *clusters* é aproximadamente:

$$N(n,k) = k^n/k! = 1/K! \sum_{i=0}^{K} ((-1)^i * C(k,i) * (k-i)^n)$$

Ex. k= 2 e n =5 então são 16 formas de dividir 5 elementos em 2 grupos. Para agrupar 25 objetos em 5 grupos: 2.436.684.974.110.751 maneiras possíveis.

E se o número de clusters é desconhecido, precisamos somar todas as partições possíveis para cada número de *clusters* entre 1 e 5.

$$\sum_{K=1}^{n} N(n,k)$$

Dificuldades

Por que a **efetividade** dos **algoritmos** de ***Clustering*** é um problema?

1. Quase todos os algoritmos de *Clustering* necessitam de **valores** para os **parâmetros** de **entrada** que são **difíceis** de **determinar**, especialmente para conjuntos de dados do mundo real contendo objetos com muitos atributos.

2. Os **algoritmos** são muito **sensíveis** a estes **valores** de **parâmetros**, frequentemente produzindo **partições muito diferentes** a medida que os parâmetros se modificam, mesmo para ajustes mínimos de parâmetros.

3. Conjuntos de dados de alta dimensionalidade têm uma **distribuição muito ampla** que pode ser difícil de ser **revelada** por um **algoritmo** de ***clustering***.

Medidas de Similaridade

As **medidas** de **similaridade** fornecem valores numéricos que expressam a "**distância**" (correlação ou associação) entre dois objetos.

Quanto **menor** o valor da "distância", **mais semelhantes** serão os objetos e estes deverão ficar no **mesmo cluster**.

Não há uma **medida** de **similaridade** que sirva para todos os **tipos** de **variáveis** que podem existir numa base de dados.
A **similaridade** pode ser medida de diversas formas:

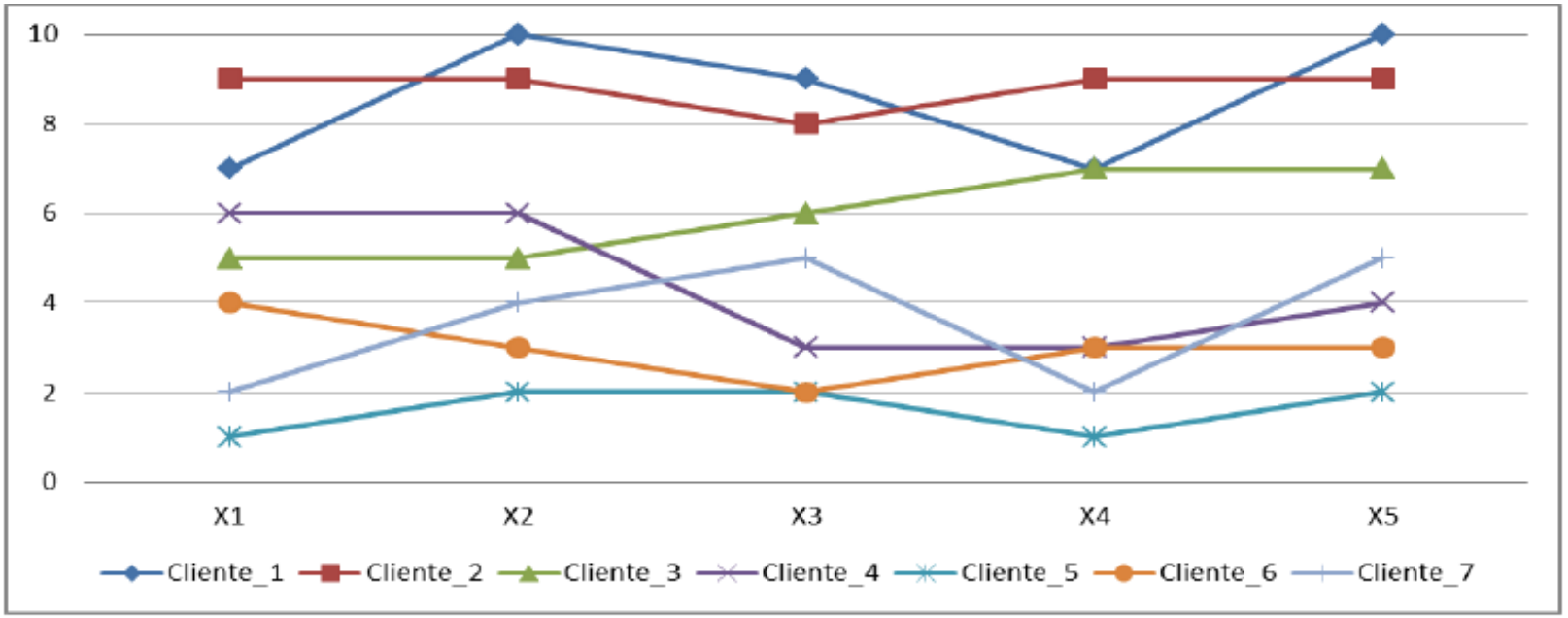
Medidas de Distância (e.g., distância euclidiana);

Medidas Correlacionais (e.g., correlação de Pearson);

Medidas de Associação (e.g., índice de Jaccard)

Medidas de Similaridade

	X1	X2	X3	X4	X5
Cliente_1	7,000	10,000	9,000	7,000	10,000
Cliente_2	9,000	9,000	8,000	9,000	9,000
Cliente_3	5,000	5,000	6,000	7,000	7,000
Cliente_4	6,000	6,000	3,000	3,000	4,000
Cliente_5	1,000	2,000	2,000	1,000	2,000
Cliente_6	4,000	3,000	2,000	3,000	3,000
Cliente_7	2,000	4,000	5,000	2,000	5,000



Medidas de Distância

- Representam a similaridade como a proximidade entre observações (instâncias) ao longo dos atributos. As medidas de distância são, na verdade, uma medida de dissimilaridade, em que os valores maiores denotam menor similaridade.

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

	Cliente_1	Cliente_2	Cliente_3	Cliente_4	Cliente_5	Cliente_6	Cliente_7
Cliente_1	0,00						
Cliente_2	3,32	0,00					
Cliente_3	6,86	6,63	0,00				
Cliente_4	10,24	10,20	6,00	0,00			
Cliente_5	15,78	16,19	10,10	7,07	0,00		
Cliente_6	13,11	13,00	7,28	3,87	3,87	0,00	
Cliente_7	11,27	12,16	6,32	5,10	4,90	4,36	0,00

Medidas Correlacionais

- Medidas correlacionais representam similaridades pela correspondência de padrões ao longo dos atributos; não analisa a magnitude de valores dos atributos, apenas o padrão global desses valores.

$$Pearson(x,y) = COV(X,Y) / \sqrt{VAR(X) * VAR(Y)}$$

$$Pearson(x,y) = \frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_i (X_i - \overline{X})^2} * \sqrt{\sum_i (Y_i - \overline{Y})^2}}$$

Medidas Correlacionais

	Cliente_1	Cliente_2	Cliente_3	Cliente_4	Cliente_5	Cliente_6	Cliente_7
Cliente_1	1,000						
Cliente_2	-0,147	1,000					
Cliente_3	0,000	0,000	1,000				
Cliente_4	0,087	0,516	-0,824	1,000			
Cliente_5	0,963	-0,408	0,000	-0,060	1,000		
Cliente_6	-0,466	0,791	-0,354	0,699	-0,645	1,000	
Cliente_7	0,891	-0,516	0,165	-0,239	0,963	-0,699	1,000

- As instâncias 1, 5 e 7 têm padrões semelhantes e correlações altas e positivas. Da mesma forma instâncias 2, 4 e 6 possuem correlações positivas. A instância 3 tem correlações baixas e correlações negativas, ou não possui correlação com as demais, de modo que talvez forme um grupo por si mesma. Portanto, as correlações representam padrões ao longo dos atributos, que vão além das magnitudes.

Algoritmos

Os métodos mais tradicionais de *Clustering* são os **métodos** por **particionamento** e os **métodos hierárquicos**.

Os *clusters* produzidos por um **método** por **particionamento** são em geral de **qualidade superior** aos produzidos por **métodos hierárquicos**.

Os mais conhecidos e geralmente usados **métodos** de **particionamento** são o ***k*-means** e o ***K*-mediana**, e suas variações.

Algoritmos

Métodos por particionamento

Os algoritmos por particionamento dividem a base de dados em ***k*** grupos, onde o número ***k*** é informado pelo usuário.

Inicialmente, o algoritmo escolhe ***k*** objetos como sendo os centros dos ***k*** *clusters*.

Os objetos são divididos entre os ***k*** *clusters* de acordo com a medida de similaridade adotada, de modo que cada objeto fique no *cluster* que forneça o menor valor de distância entre o objeto e o centro do mesmo.

Algoritmos

Então, o algoritmo utiliza uma estratégia iterativa de controle para determinar que **objetos** devem **mudar** de ***cluster*** de forma que se **otimize** uma **função objetivo**.

Após a divisão inicial, há duas possibilidades na escolha do "elemento" que vai **representar** o **centro** do ***cluster***, e que será a referência para o cálculo da medida de similaridade:

Algoritmos

Ou utilizamos a média dos objetos que pertencem ao *cluster* em questão, esta é a abordagem conhecida como **k-means**;

Ou escolhemos como representante o objeto que se encontra mais próximo ao centro de gravidade do cluster;

Esta abordagem é conhecida como **k-mediana**, e o elemento mais próximo ao centro é chamado de **mediana**.

Algoritmos

A função objetivo mais utilizada para espaços métricos é o critério do erro quadrático:

$$\mathbf{E = \Sigma_{i=1:K} \Sigma_{x \varepsilon Ci} || p_i - m_j ||^2}$$

onde, **E** é a soma do erro quadrático para todos os objetos na base de dados, $\mathbf{p_i}$ é o ponto no espaço representando um dado objeto, e $\mathbf{m_i}$ é o representante do cluster C_i.

Tanto $\mathbf{p_i}$ quanto $\mathbf{m_j}$ são multidimensionais.

A função representa a distância média de cada objeto ao seu respectivo representante.

Algoritmo K-means

O algoritmo **k-means** utiliza um parâmetro de entrada ***k***, e particiona um conjunto de ***n*** objetos em ***k*** clusters tal que, a similaridade intracluster resultante é alta, mas a similaridade intercluster é baixa.

O algoritmo **k-means**, trabalha bem quando os clusters são densos e compactos e bem separados uns dos outros.

Algoritmo K-means

O Algoritmo **k-means** para particionamento de objetos baseia-se no valor médio das distâncias dos objetos no cluster.

Entrada: O número de clusters **k,** e a base de dados com **n** objetos.
Saída: Um conjunto de **k** clusters que minimizam o critério do erro quadrático.
Método:

1. Escolha arbitrariamente **k** objetos da base de dados como os centros inicias dos clusters;
2. **Repita**
   2.1 (Re)atribua cada objeto ao cluster ao qual o objeto é mais similar, de acordo com o valor médio dos objetos no cluster;
   2.2. Calcule o valor médio dos objetos para cada cluster;
   Até que não haja mudança de objetos de um cluster para outro.

Algoritmo K-means

Uma estratégia que frequentemente produz bons resultados é primeiro aplicar um algoritmo hierárquico aglomerativo para determinar o número de *clusters* e então, para encontrar os agrupamentos utilizar o algoritmo **k-means** para melhorar os agrupamentos iniciais.

Algoritmo K-means

Dataset a ser agrupado

	Variáveis	
Item	x1	x2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

Algoritmo K-means

Passo 1

Particiona-se os itens em dois *clusters* (AB) e (CD) e calcula-se a coordenada (x1,x2) do centróide do *cluster*.

Cluster	Coordenadas dos centros x1	x2
(AB)	(5 + (-1)) / 2 = 2	(3 + 1) / 2 = 2
(CD)	(1 + (-3)) / 2 = -1	(-2 + (-2)) / 2 = -2

Algoritmo K-means

Passo 2

Calcula-se a similaridade de cada item em relação ao centróide e em relação a cada item no grupo mais próximo. Se um item é movido de um agrupamento para outro, o centróide do *cluster* dever ser atualizado.

$$
\begin{aligned}
d^2(A, (AB)) &= (5 - 2)^2 + (3 - 2)^2 = 10 \\
d^2(B, (AB)) &= (-1 - 2)^2 + (1 - 2)^2 = 10 \\
d^2(C, (AB)) &= (1 - 2)^2 + (-2 - 2)^2 = 17 \\
d^2(D, (AB)) &= (-3 - 2)^2 + (-2 - 2)^2 = 41 \\
d^2(A, (CD)) &= (5 + 1)^2 + (3 + 2)^2 = 61 \\
d^2(B, (CD)) &= (-1 + 1)^2 + (1 + 2)^2 = 9 \\
d^2(C, (CD)) &= (1 + 1)^2 + (-2 + 2)^2 = 4 \\
d^2(D, (CD)) &= (-3 + 1)^2 + (-2 + 2)^2 = 4
\end{aligned}
$$

Algoritmo K-means

Ocorre o deslocamento do item (B) para o segundo *cluster* e calcula-se novamente as coordenadas.

	Coordenadas dos centros	
Cluster	x1	x2
A	5	3
(BCD)	-1	-1

Algoritmo K-means

Calcula-se a distância dos itens em relação ao cluster, para verificar a parada do algoritmo.

	Distâncias dos centróides			
	Item			
Cluster	A	B	C	D
A	0	40	41	89
(BCD)	52	4	5	5

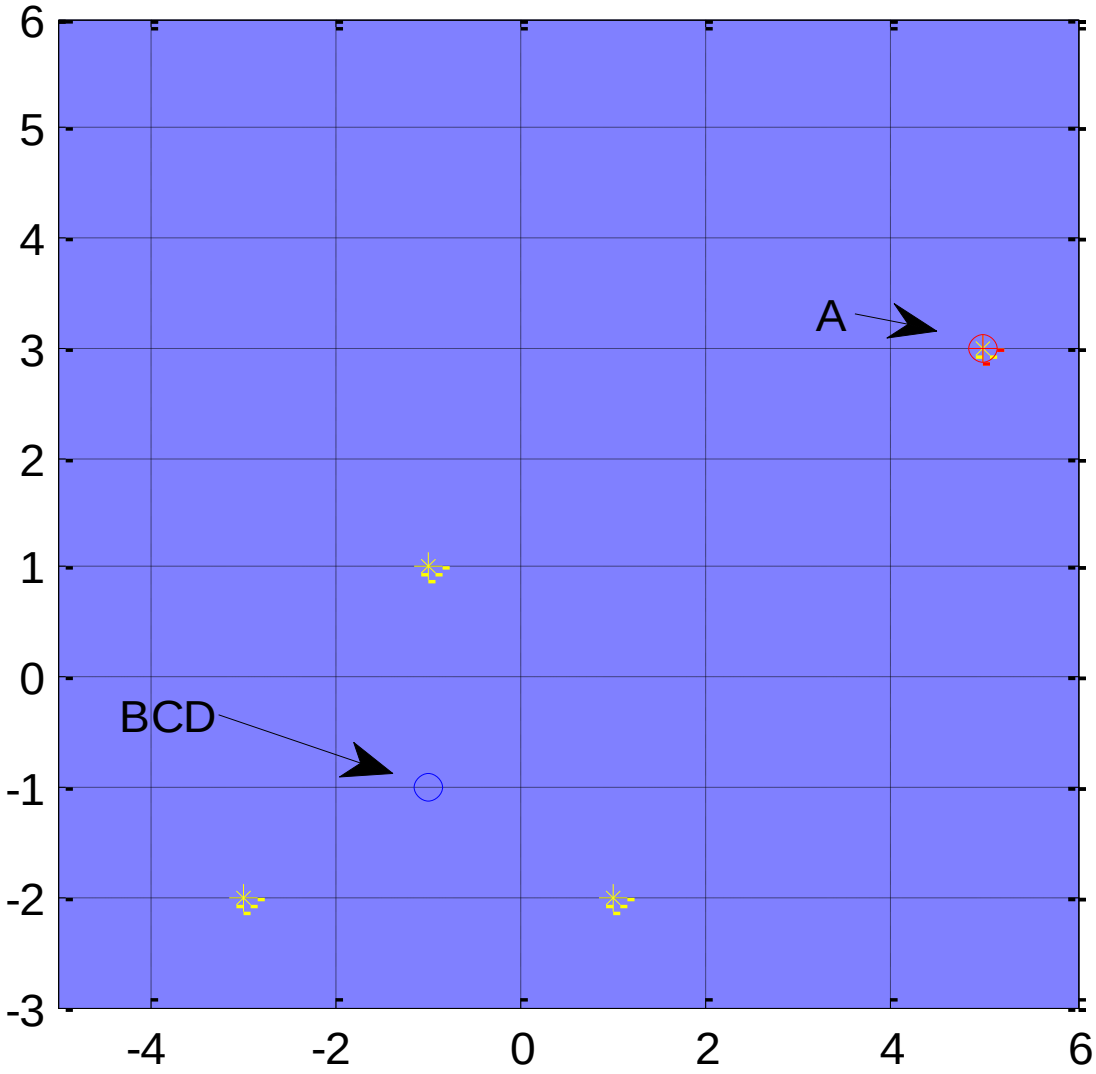
$$d^2(A, A) = (5-5)^2 + (3-3)^2 = 0$$

$$d^2(A, (BCD)) = (5+1)^2 + (3+1)^2 = 52$$

$$d^2(B, A) = (-1-5)^2 + (1-3)^2 = 40$$

$$d^2(B, (BCD)) = (-1+1)^2 + (1+1)^2 = 4$$

Algoritmo k-means



Bons Estudos!