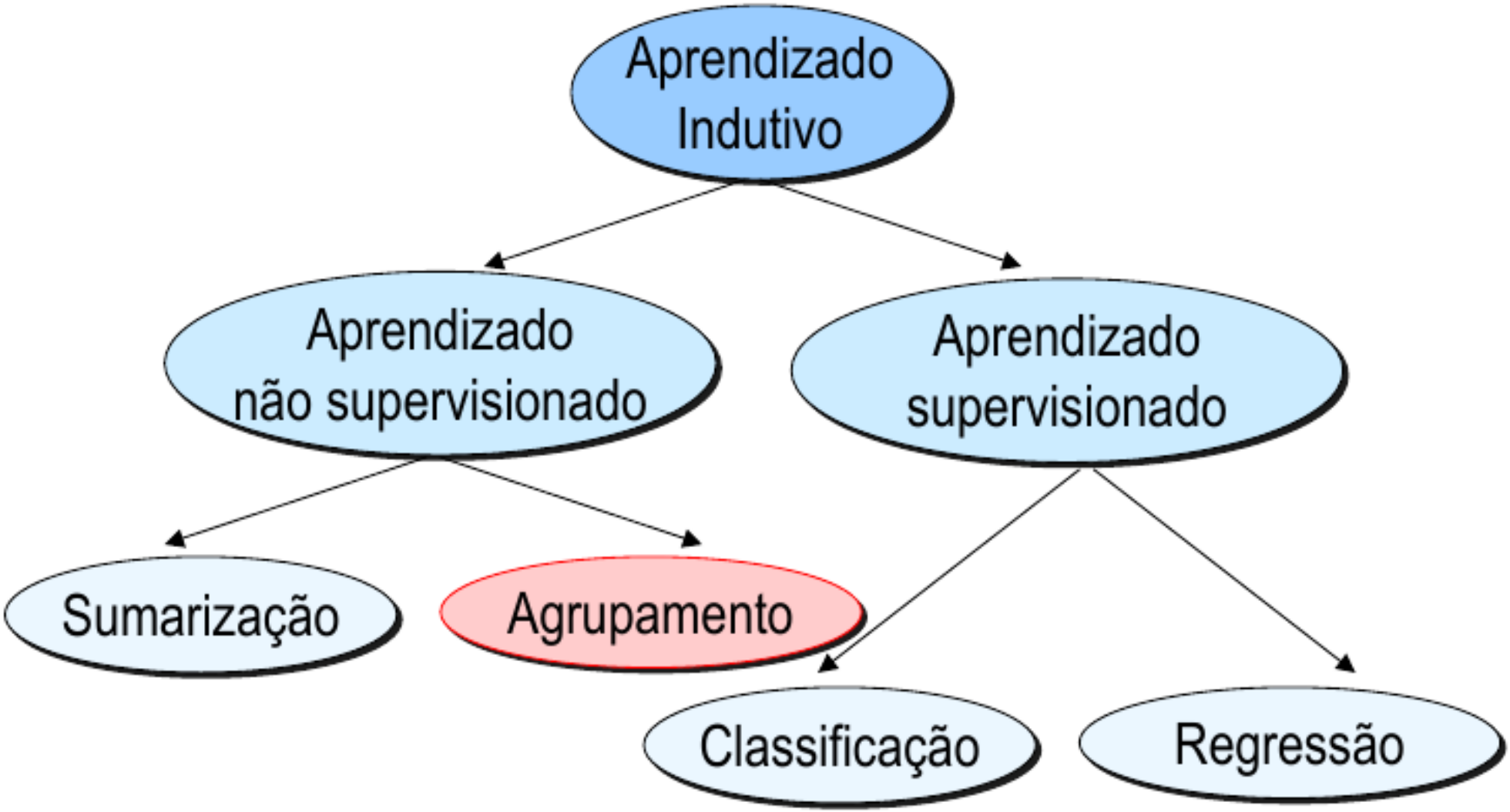
Aprendizado de Máquina: Agrupamento de Dados

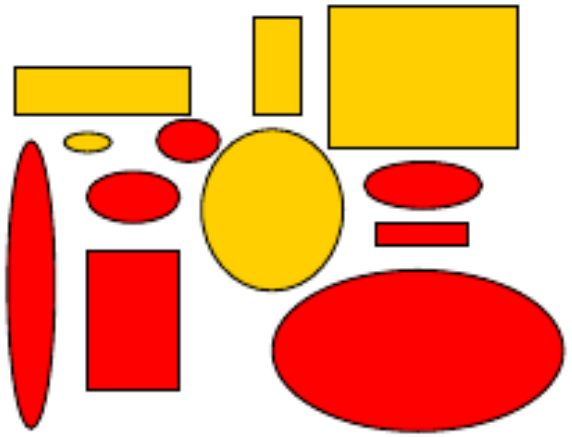
Prof. Arnaldo Candido Junior
UTFPR – Medianeira

Agrupamento



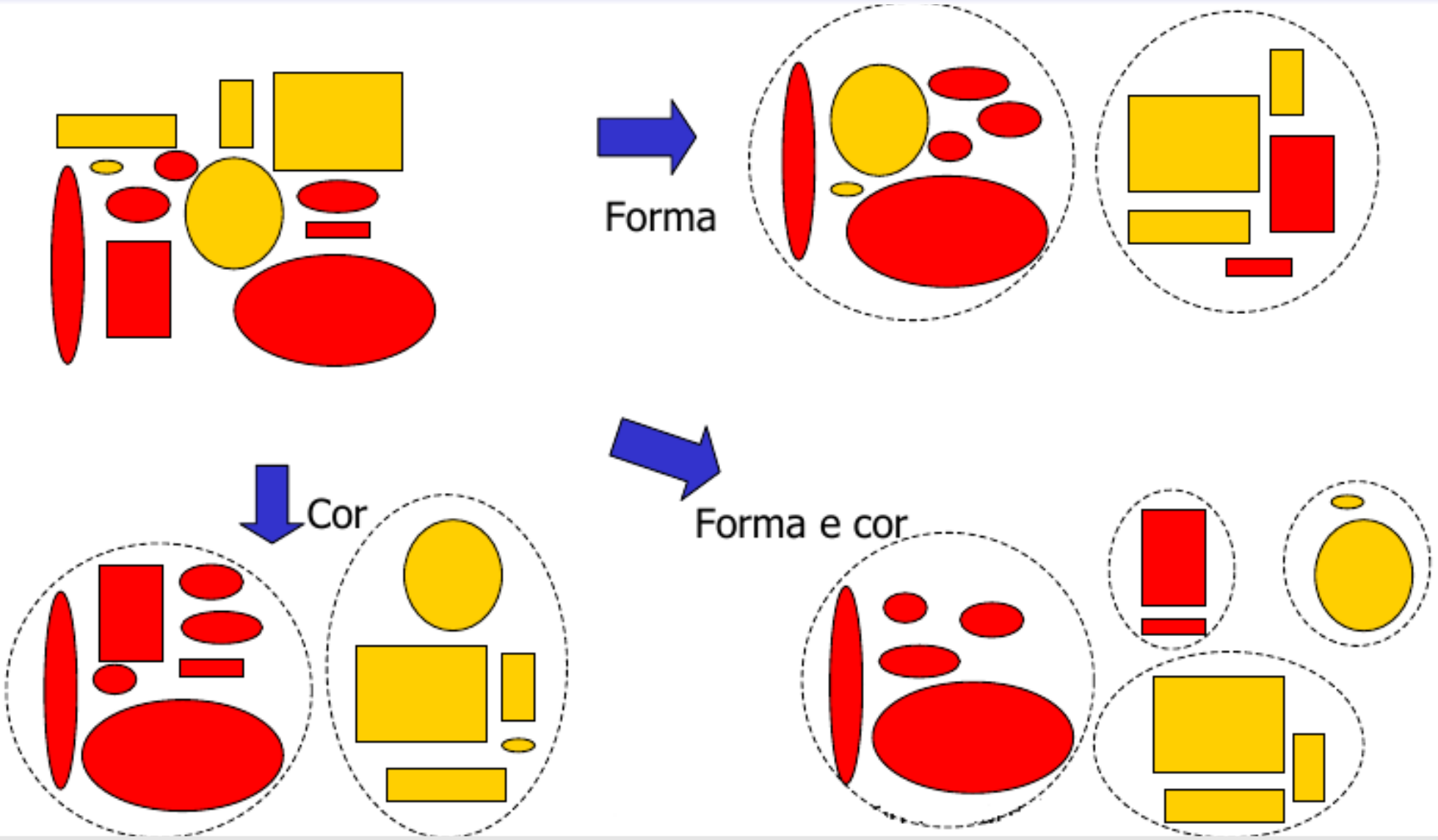
Agrupamento (2)

- Organização de um conjunto de objetos em grupos (clusters)
  - De acordo com alguma forma de semelhança ou relação entre eles



Como organizar?

Agrupamento (3)



Agrupamento (4)

- Agrupar dados em grupos (clusters) que:
  - Possuam um significado. Ex.: capturar a estrutura natural dos dados
  - Sejam o passo inicial para outros propósitos: sumarização de dados, compressão de dados
  - Sejam úteis para algum propósito
  - Sejam classes em potencial

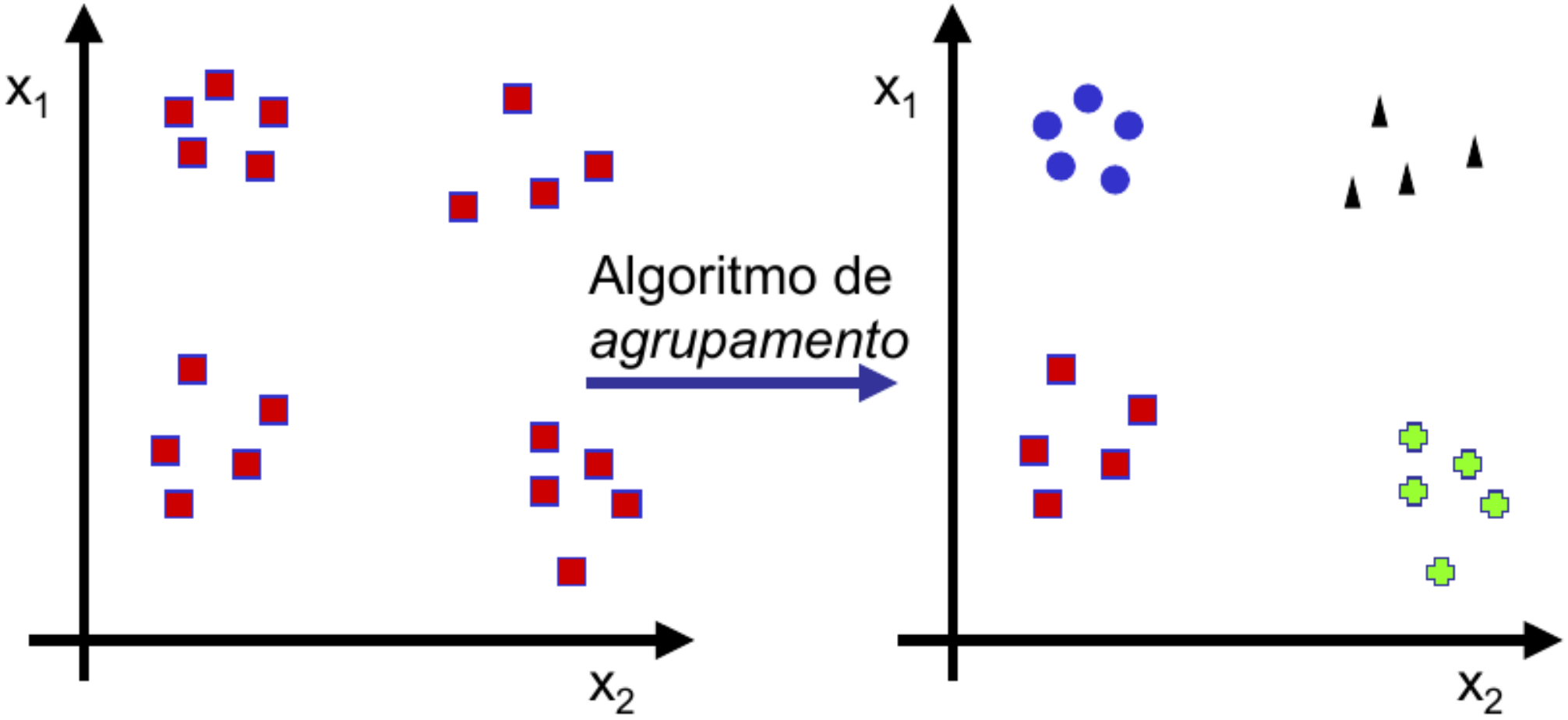
Agrupamento (5)

- Análise de cluster: estudo de técnicas para encontrar classes automaticamente
  - Fornece uma abstração das instâncias individuais presentes nos dados para seus respectivos grupos
- Algumas técnicas representam cada cluster por um protótipo: representante das instâncias do cluster
  - Pode ser utilizado para várias análises

Análise de Cluster

- Agrupa instâncias utilizando apenas informações sobre instâncias e seus relacionamentos
- Objetivo: instâncias dentro de um grupo sejam semelhantes entre si e distintas de instâncias de outros grupos
- Quanto maior a homogeneidade dentro dos grupos e a diferença entre os grupos, melhor
- Em várias aplicações, noções do que é um cluster não esta bem definida

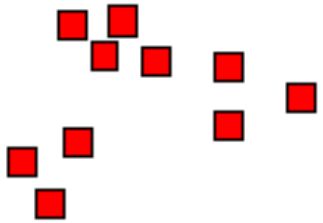
Análise de Cluster (2)



Diferentes Alternativas

Quantos clusters?

Seis clusters



Dois clusters

Quatro clusters

Análise de Cluster

- Definição do que é um cluster
  - Impreciso
  - Depende de:
    - Natureza dos dados
    - Resultados desejados
  - Existem várias definições de cluster

Análise de Cluster (2)

- Algoritmos de agrupamento
  - São não supervisionados
  - Ferramentas de análise de dados
  - Agrupam exemplos semelhantes de acordo com alguma medida de (dis)similaridade
  - Estruturas diferentes são detectadas para diferentes valores dos parâmetros

Principais Etapas

- Passos: cada passo é subjetivo e influenciado pela experiência e conhecimento do especialista
  - 1. Pré-processamento: seleção de características, normalização
  - 2. Definição de medida de (dis)similaridade

Principais Etapas (2)

- 3. Definição de critério de agrupamento
  - Define como os grupos são formados

Principais Etapas (3)

- 4. Verificar tendência de agrupamento
- 5. Definir algoritmo de agrupamento
- 6. Validação dos clusters
  - Verificar se escolha dos parâmetros do algoritmo e formato do cluster casam com o agrupamento natural dos dados
- 7. Interpretação
  - O especialista interpreta os resultados obtidos junto com informações sobre o problema

Tipos de Agrupamento

- Seja X = {$\hat{x}_1$, $\hat{x}_2$,..., $\hat{x}_n$} o conjunto de todos as instância
  - Tarefa: colocar cada $\hat{x}_i$ em um dos *m* clusters C_1, C_2, ..., C_m
- Clusters podem ser de dois tipos:
  - Tipo 1: duro (crisp)
  - Tipo 2: fuzzy

Tipos de Agrupamento (2)

- Cluster crisp
- Cada exemplo X_i pertence ou não a cada cluster C_j

$$C_i \neq \varnothing,\ i = 1, \ldots, m \qquad \bigcup_{i=1}^{m} C_i = X$$

$$C_i \cap C_j = \varnothing,\ i \neq j,\ i, j \in \{i, 2, \ldots, m\}$$

- Exemplo em C_i mais semelhante a outros em C_i que àqueles em C_j, $i \neq j$

Tipos de Agrupamento (3)

- Cluster Fuzzy
  - Usa uma função de pertinência para definir o quanto um elemento pertence a um grupo

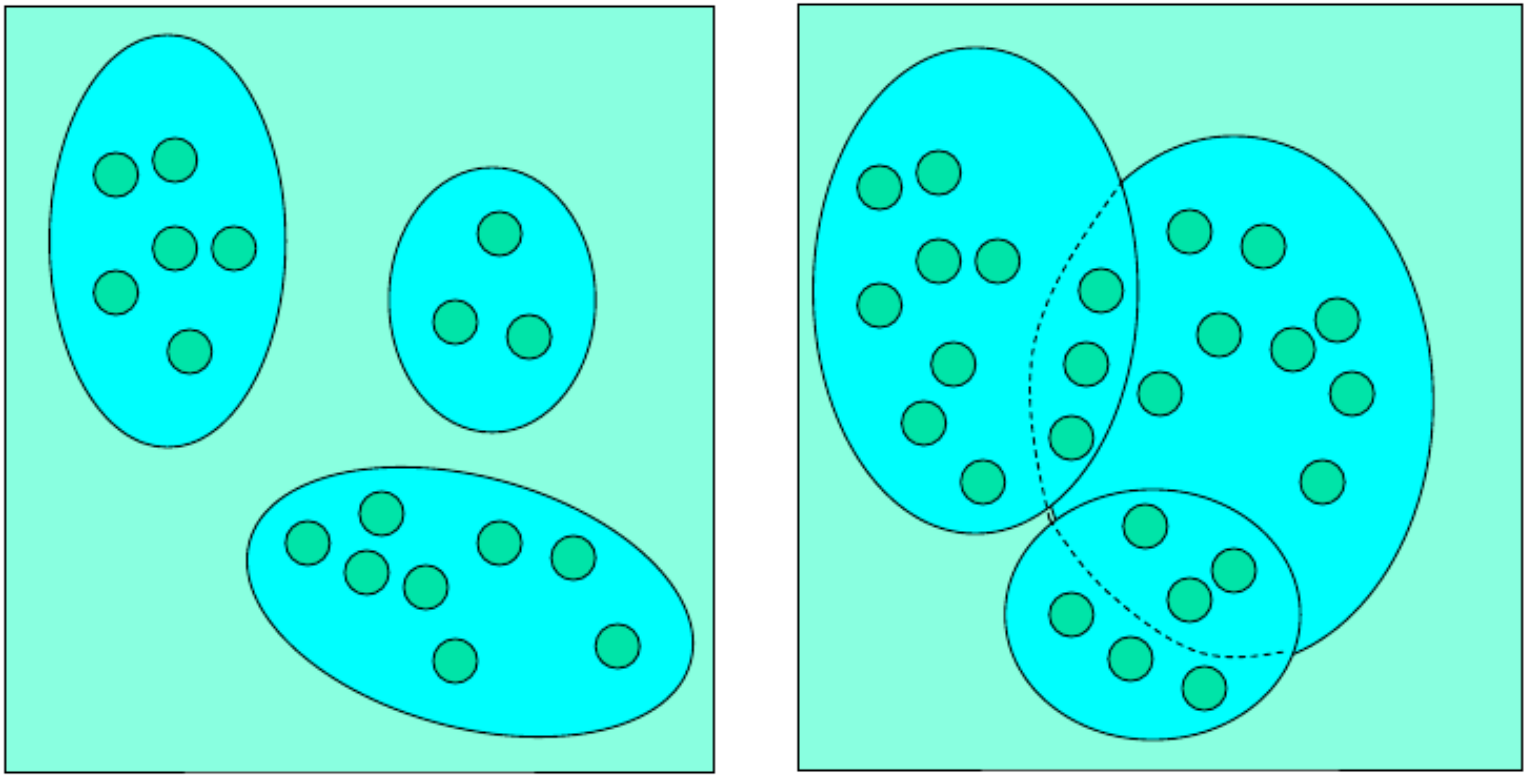
$$\mu_j : X \rightarrow [0, 1]$$

$$\sum_{j=1}^{m} \mu_j(x_i) = 1, i \in \{1, \ldots, n\}$$

m = número de grupos
n = número de objetos

$$0 < \sum_{i=1}^{n} \mu_j(x_i) < n, \; j \in \{1, \ldots, m\}$$

Tipos de Agrupamento (4)



Algoritmos de clusterização

- Busca exaustiva
  - Tentar todos os possíveis clusters de tamanho *m* para vários valores de *m*
  - Números de Stirling do segundo tipo
    - Número de formas de particionar *n* dados em *m* subconjuntos não vazios

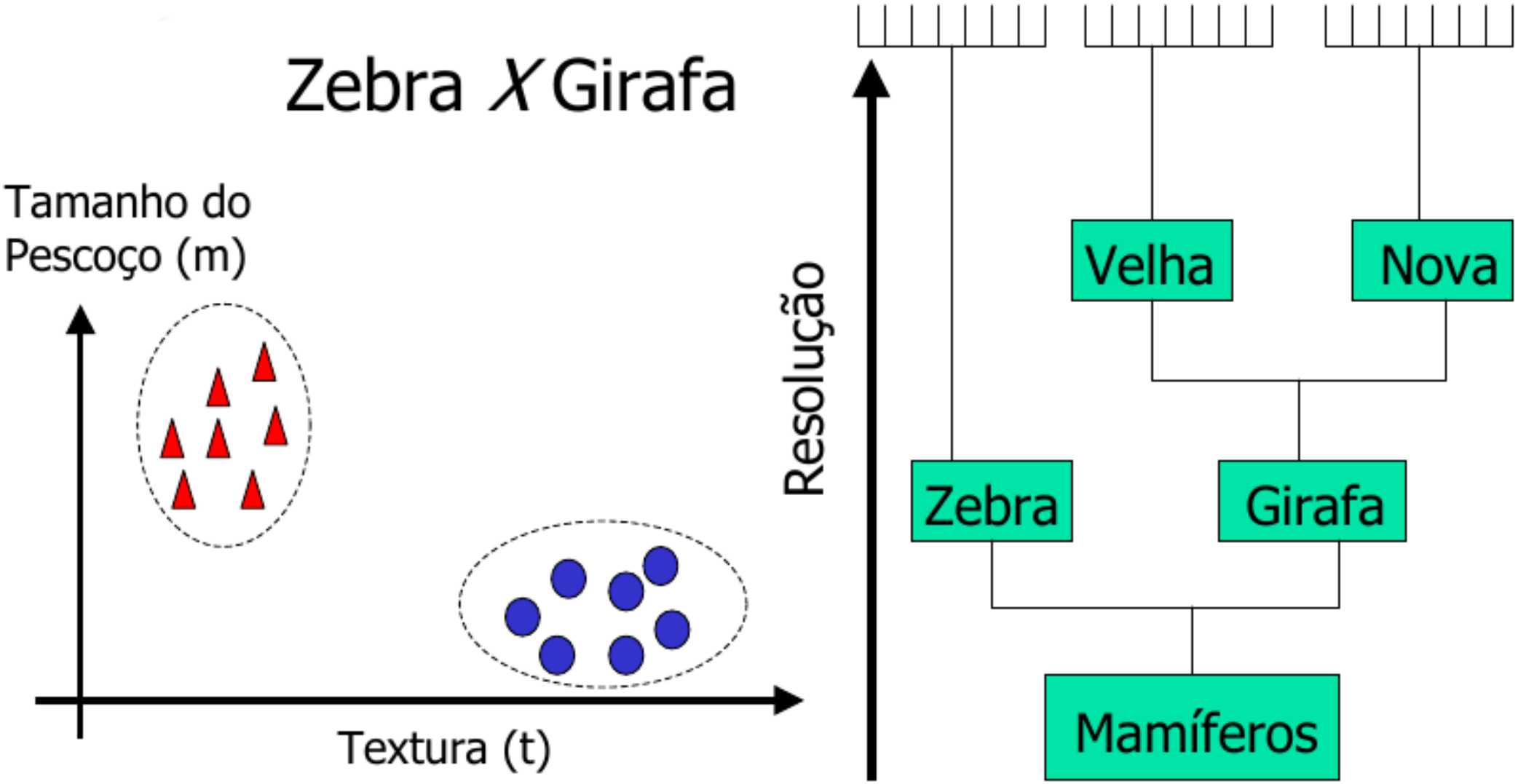
$$>> \begin{pmatrix} n \\ m \end{pmatrix} \geq \left(\frac{n}{m} \right)^m$$

- Método de força bruta é impraticável

Algoritmos de clusterização (2)

- Algoritmos particionais
- Algoritmos hierárquicos
- Algoritmos baseados em otimização de função de custo
- Algoritmos baseados em grafos
- Outros algoritmos

Particional X Hierárquico



Algoritmos Particionais

- Principais características
  - Produzem um único nível de agrupamento (plano)
  - A maioria utiliza abordagem “gulosa” (*greedy*)
    - Sempre procura escolher a melhor alternativa atual, sem considerar consequências futuras
    - Uma vez tomada uma decisão, ela não é mais alterada
    - Geralmente, resultado depende da ordem de apresentação dos exemplos

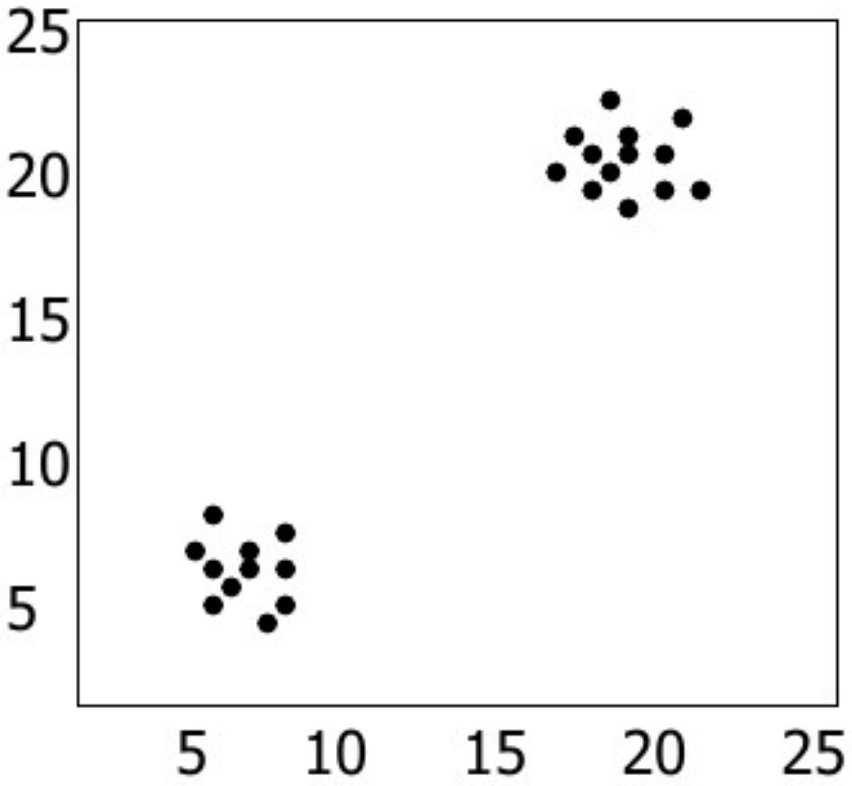
Algoritmo Particional Básico

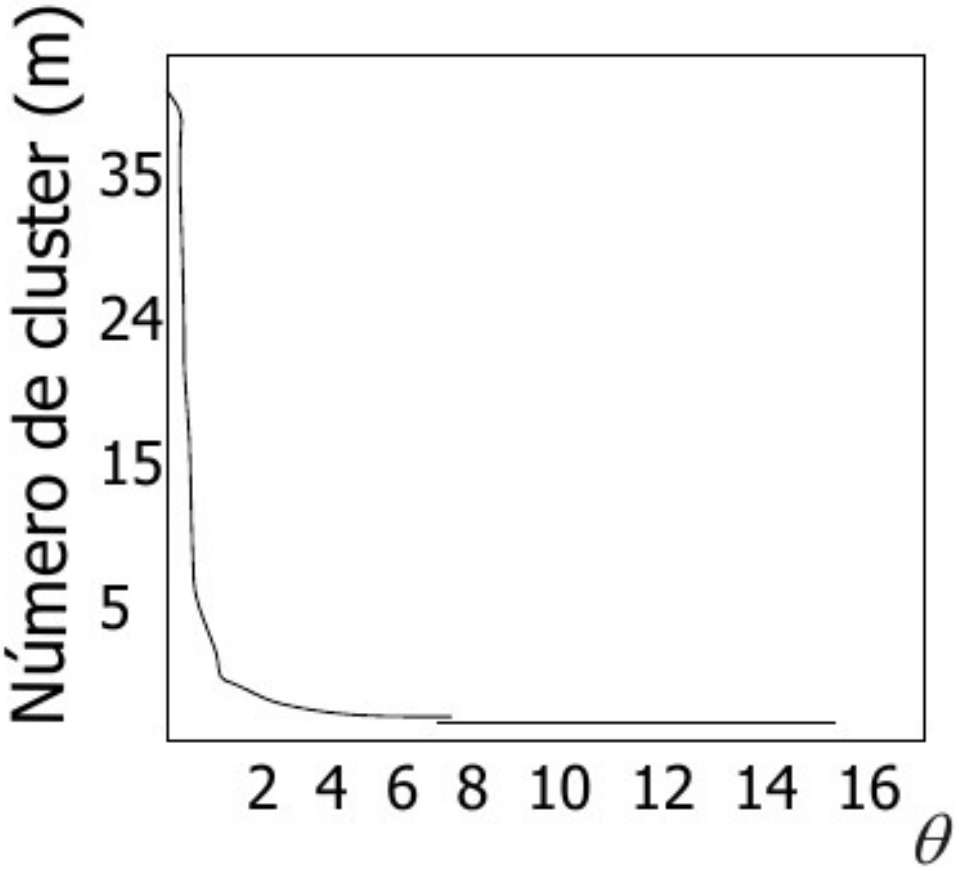
Entrada: θ, q
/* q, número máximo de clusters, é opcional) */
1 Inicializar m = 1, $C_1 = \{\hat{x}_1\}$
2 Para i = 2 até n faça
    C_k é o cluster mais próximo de $\hat{x}_i$
    Se $d(C_k, x_i) > \theta$ e m < q /* usar centros
    Então m = m + 1
        $C_m = \{x_i\}$
    Senão $C_k = C_k \cup \{x_i\}$ /* atualizar centros

Algoritmo Particional Básico (2)

- Sensitividade (granularidade)
  - Se θ for grande, poucos (grandes) clusters são formados
  - E vice-versa
- Como estimar valor de θ ?
  - Executar para vários valores de θ e m

Algoritmo Particional Básico (3)





Exemplos de Algoritmos Particionais

- K-médias
- K-médias ótimo
- K-médias sequencial
- SOM
- DENCLUE
- CLICK
- CAST
- SNN

Algoritmo k-médias

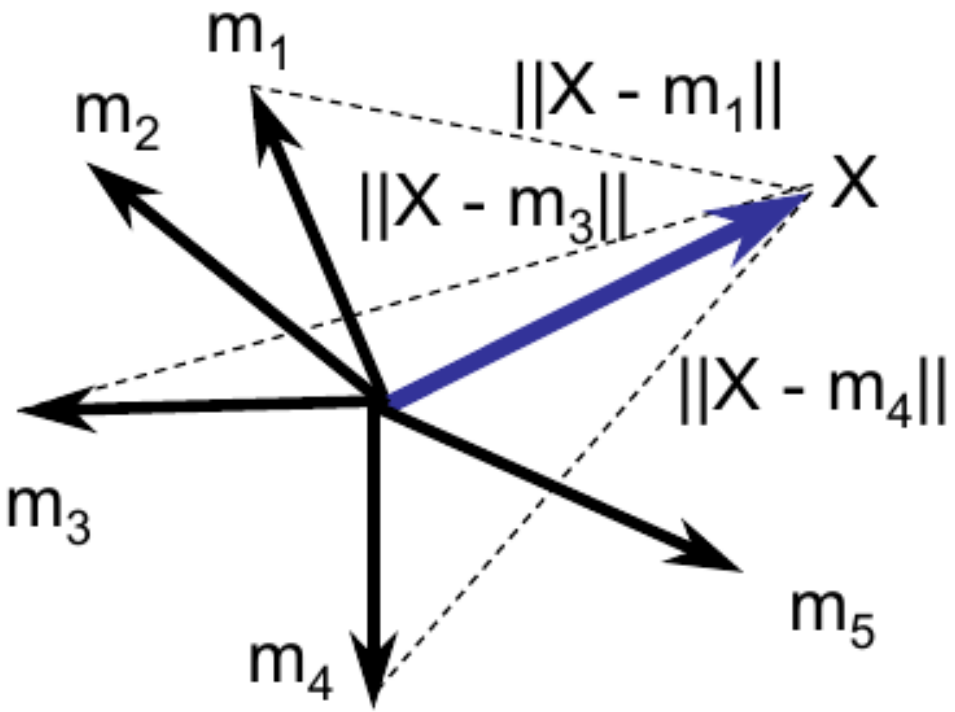
- Dataset contém n instâncias $\hat{x}_1$, $\hat{x}_2$, ..., $\hat{x}_N$
- Definir o númeo k de clusters, onde k < n
- Seja $\hat{m}_i$ a média das instâncias do cluster C_i

Algoritmo k-médias (2)

- Se os clusters estão bem separados
  - Critério de menor distância pode ser utilizado para definir a que cluster um instância pertence
  - $\hat{\mathbf{x}}_j \in$ cluster C_i se $||\hat{\mathbf{x}}_j - \mathbf{m}_i||$ é a menor de todas as k distâncias entre $\hat{\mathbf{x}}_j$ e $\hat{\mathbf{x}}_j - \mathbf{m}_i$, $j = 1, 2, ..., k$ e $i \neq j$
  - Onde $||u||$ é a norma de um vetor u

Medidas de Distância

- Calculam $|| \hat{x} - \hat{m}_i ||$ para i = 1 até k e escolhem o grupo com a menor distância



Algoritmo k-médias

1 Sugerir médias $\mathbf{m}_1$, $\mathbf{m}_2$, ..., $\mathbf{m}_k$ iniciais

2 Repetir

Usar as médias sugeridas para agrupar as instâncias nos k clusters

Para i variando de 1 a k

Substituir $\mathbf{m}_i$ pela média de todos os exemplos do cluster $\mathbf{C}_i$

Até nenhuma das médias mudar

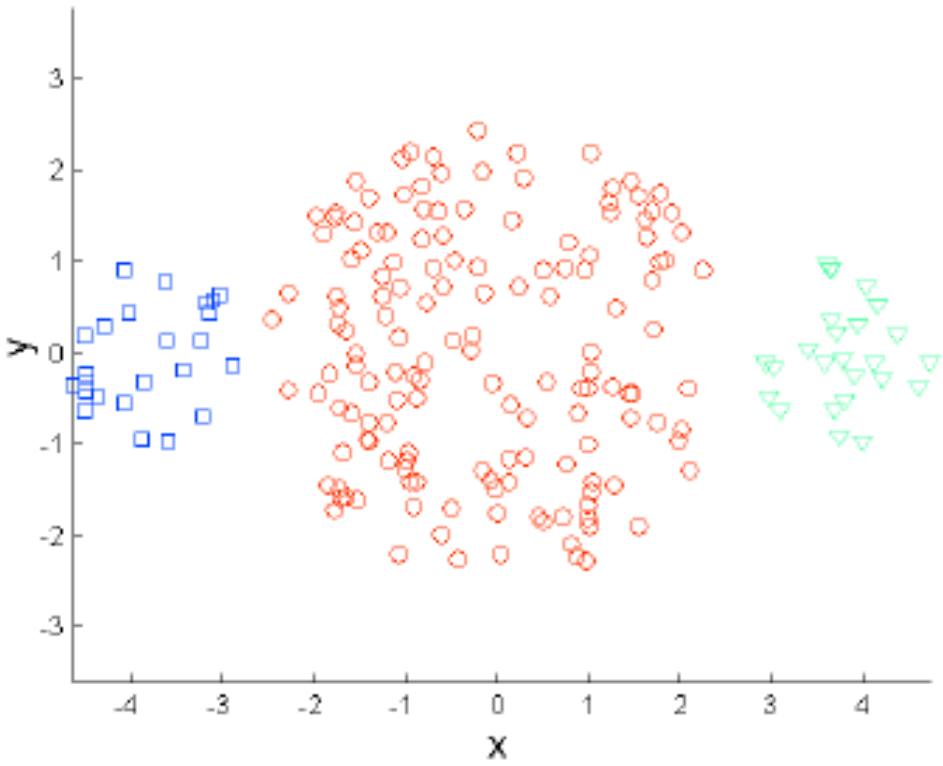
Algoritmo k-médias (2)

- Médias iniciais
  - Vetores aleatórios
  - Elementos aleatoriamente escolhidos do conjunto de treinamento

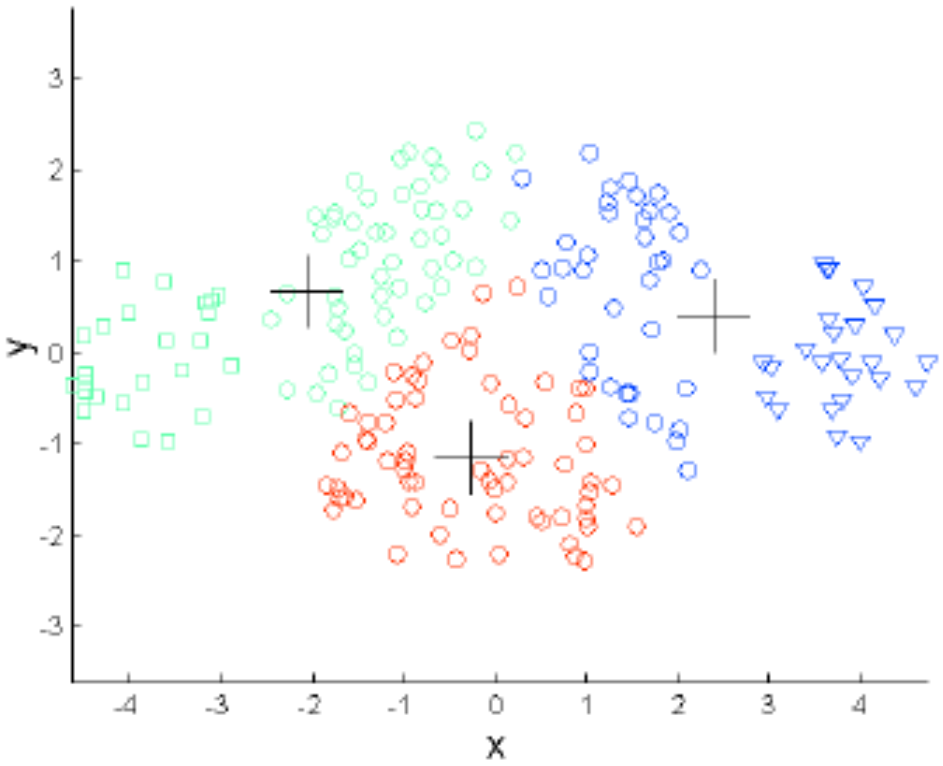
Limitações do k-médias

- Escolha do valor de k
- Problemas para k-médias:
  - Grupos de diferentes densidades/tamanhos
  - Formatos não hiper-esféricos
  - *Outliers*

Tamanhos diferentes

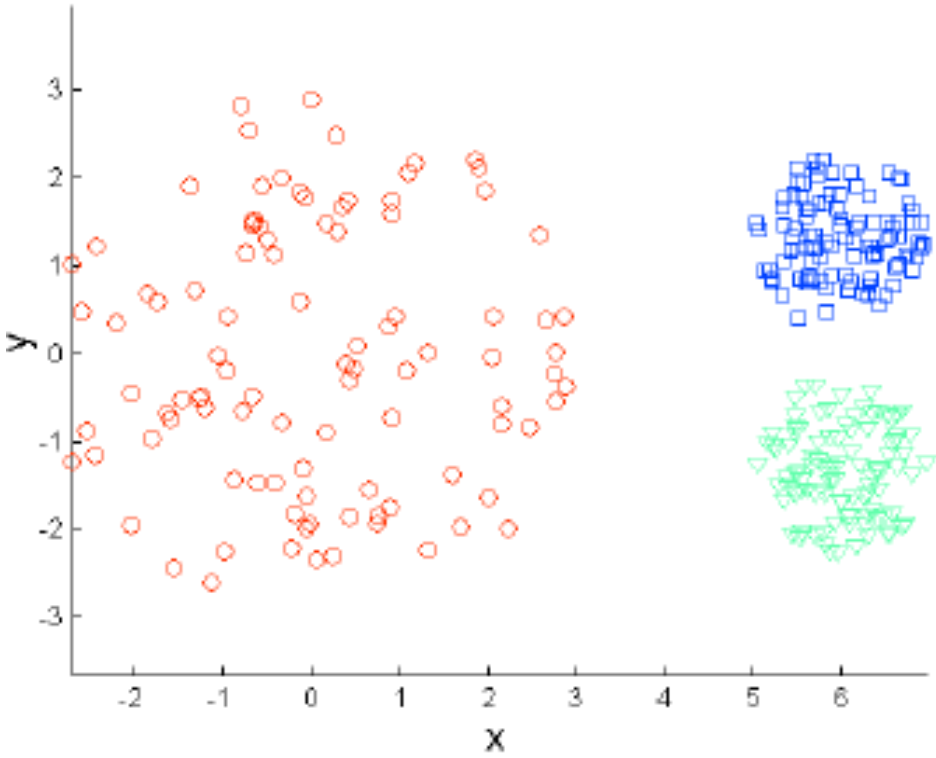


Dados originais

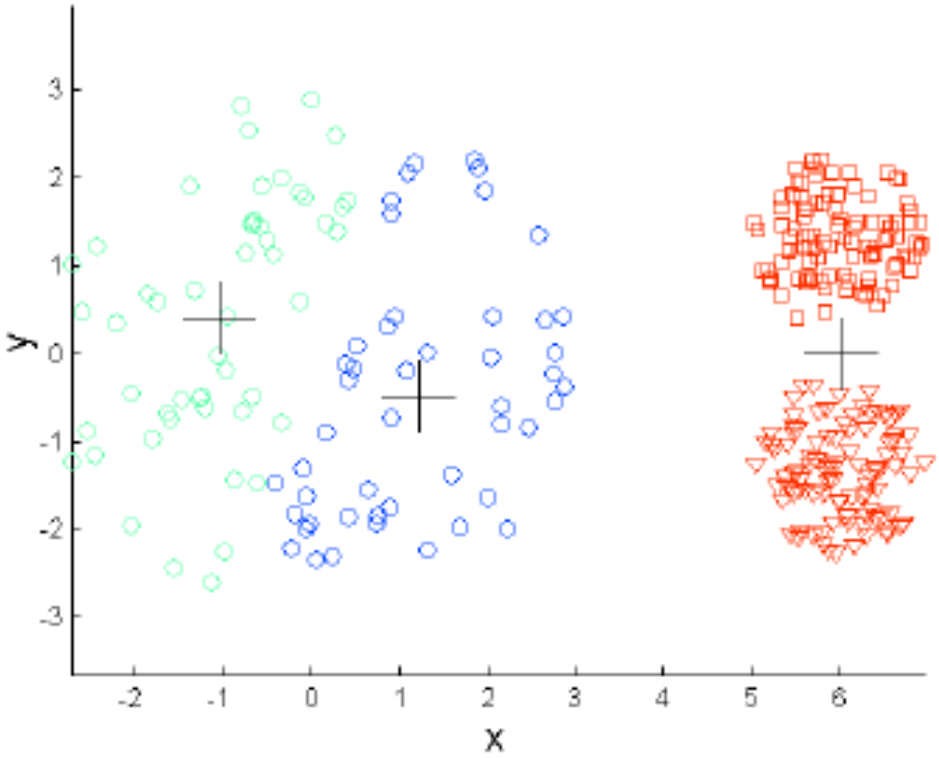


K-médias (3 Clusters)

Densidades diferentes

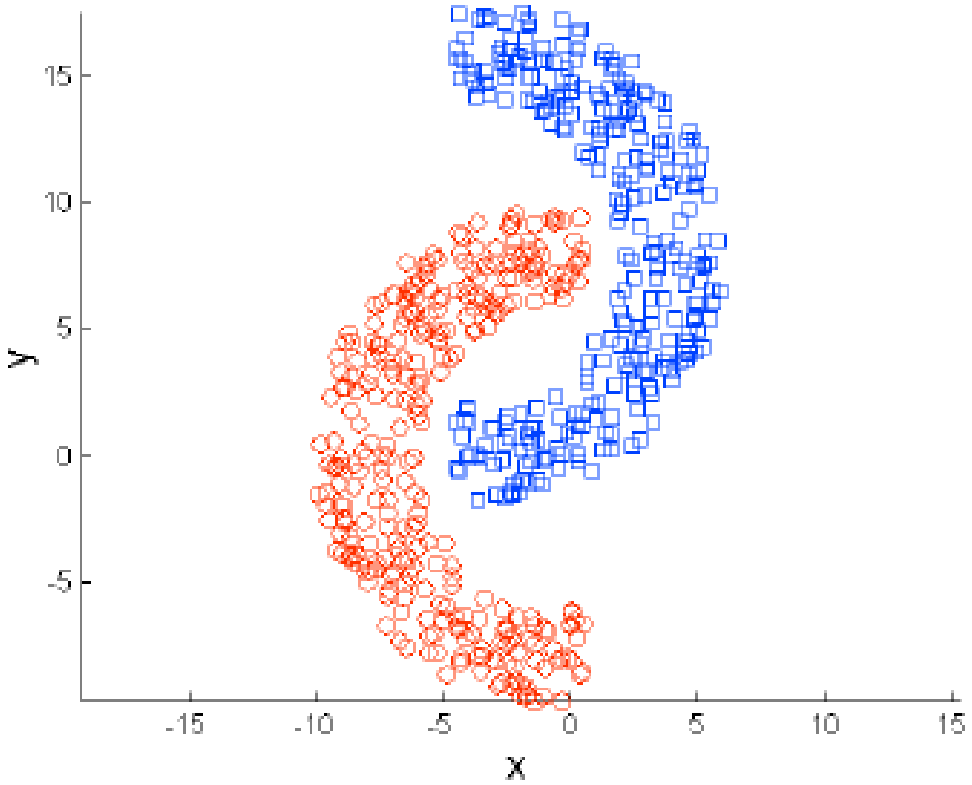


Dados originais

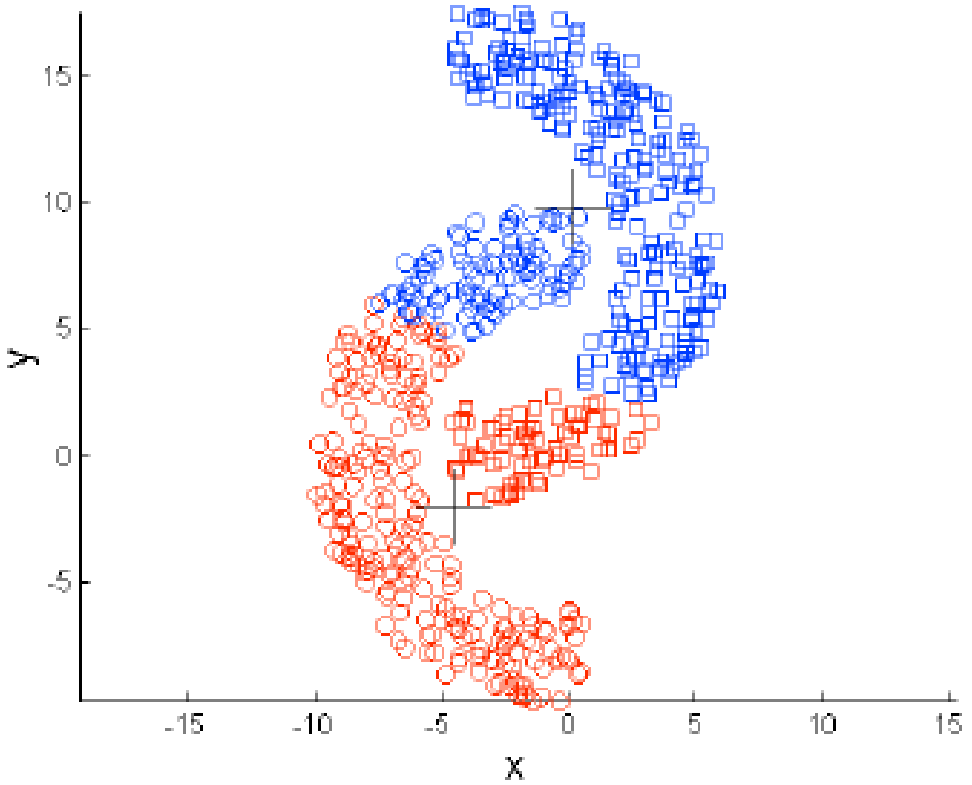


K-médias (3 Clusters)

Formatos não hiper-esféricos



Dados originais

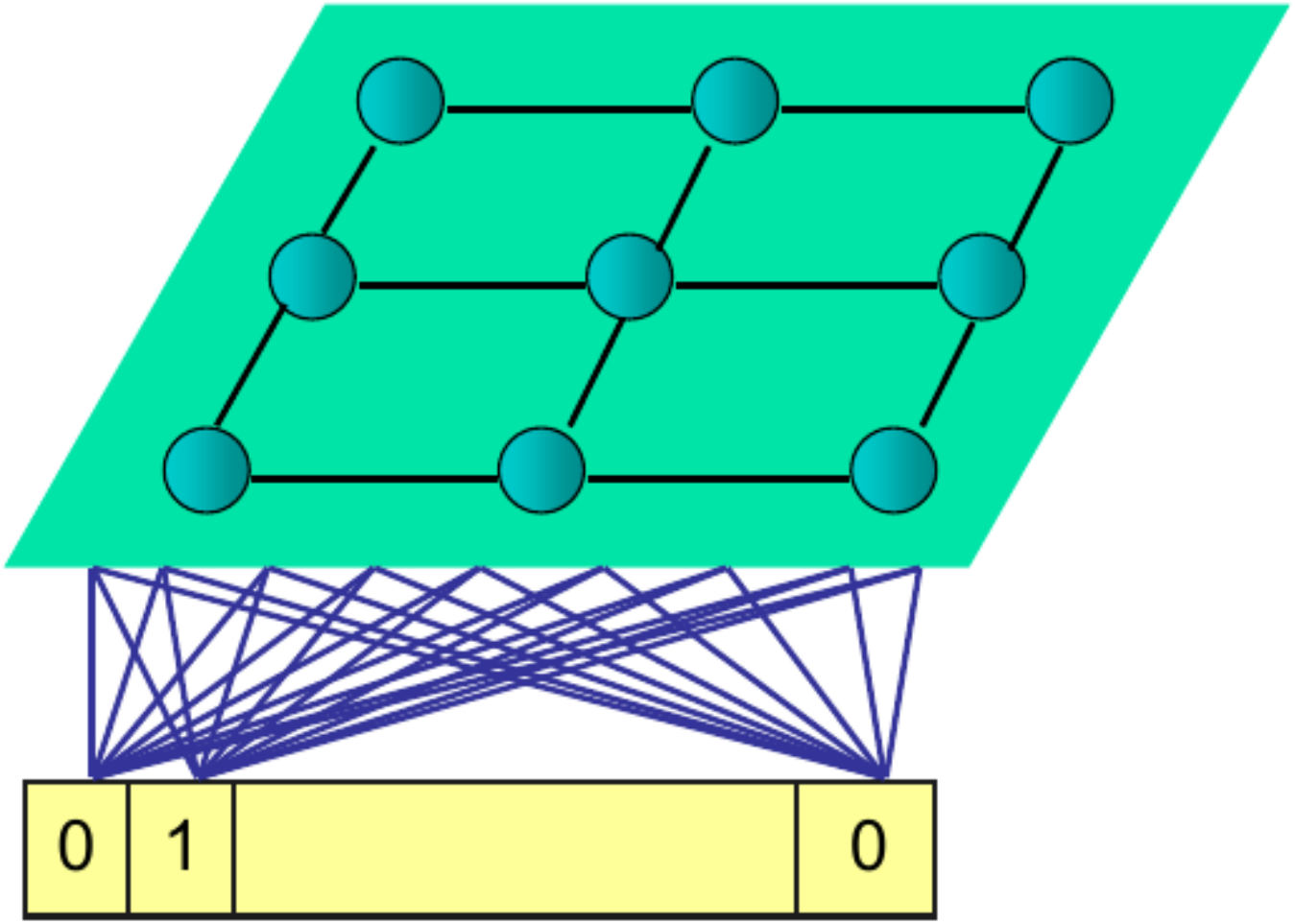


K-médias (2 Clusters)

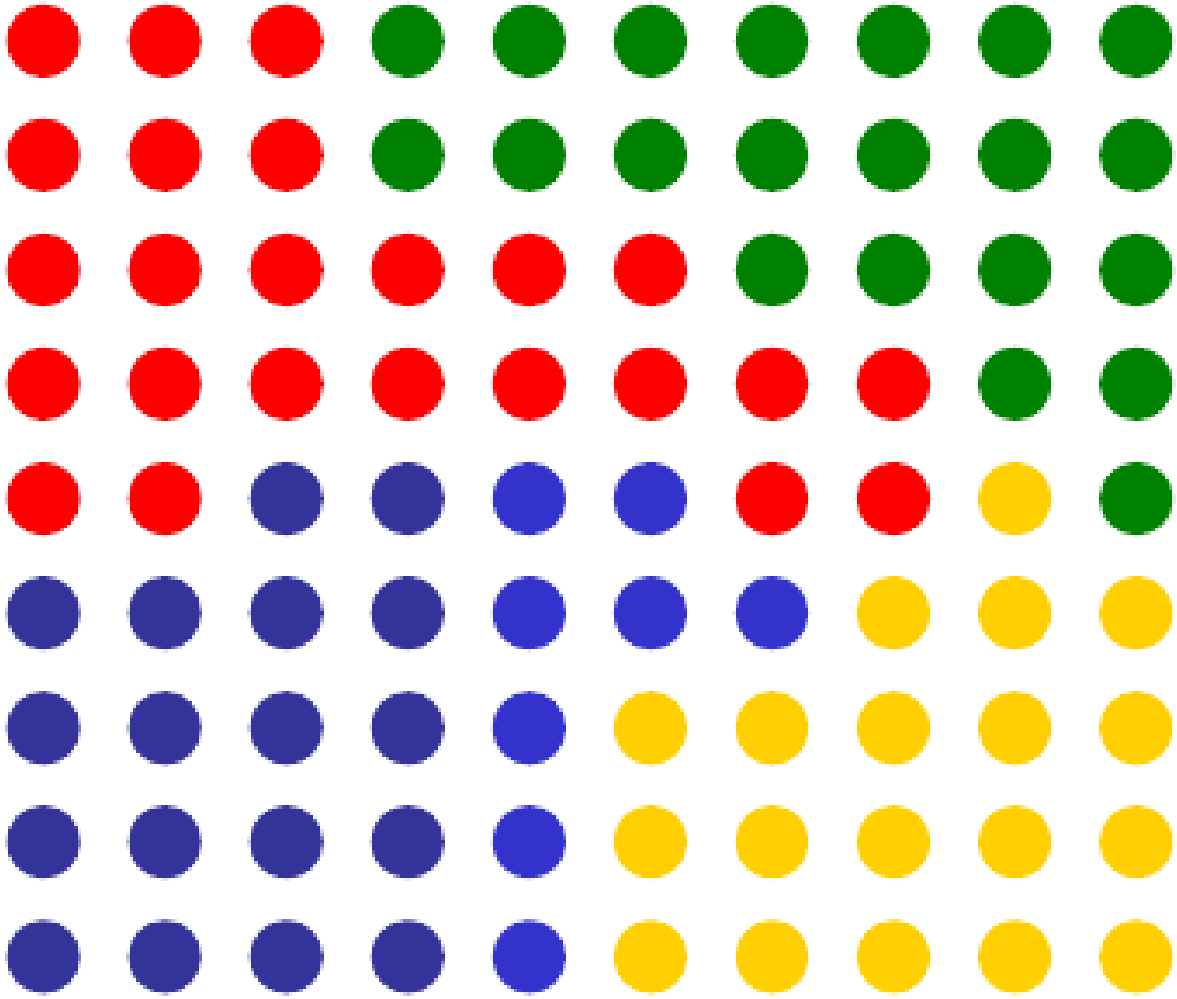
Self-Organizing Maps (SOM)

- Rede neural não supervisionada proposta na década de 80
- Cria grupos em um grid de nós (mapa topográfico)
  - Aprendizado Competitivo
- Treinamento guiado por:
  - Taxa de aprendizado e
  - Taxa de redução de raio de vizinhança

Redes SOM



Redes SOM (2)



Exercício

- Seja o seguinte cadastro de pacientes:

Nome	Febre	Enjôo	Manchas	Dores	Diagnóstico
João	sim	sim	pequenas	sim	doente
Pedro	não	não	grandes	não	saudável
Maria	sim	sim	pequenas	não	saudável
José	sim	não	grandes	sim	doente
Ana	sim	não	pequenas	sim	saudável
Leila	não	não	grandes	sim	doente

Exercício (2)

- Agrupar os dados em dois grupos usando o algoritmo K-médias
  - Usar k = 2
  - Rodar duas iterações
  - Usar distância Manhattan
  - Informação sobre a classe não será usada
  - Usar João como primeira média do grupo 1
  - Usar Leila como primeira média do grupo 2

Exercício (3)

- Em que grupos seriam colocados os novos casos?
  - (Luis, não, não, pequenas, sim)
  - (Laura, sim, sim, grandes, sim)

Algoritmos Hierárquicos

- Utilizam diagrama de árvore (dendograma)
  - Produzem uma sequência (hierarquia) de agrupamentos
- Historicamente utilizados em áreas que utilizam estrutura de agrupamento hierárquica
  - Ex.: biologia

Algoritmos Hierárquicos (2)

- Conceito de representação hierárquica de dados foi desenvolvido inicialmente na biologia
  - Algoritmos de agrupamento hierárquicos lembram a estrutura hierárquica da taxonomia de Linnaean
  - Biólogos geralmente preferem agrupamentos hierárquicos

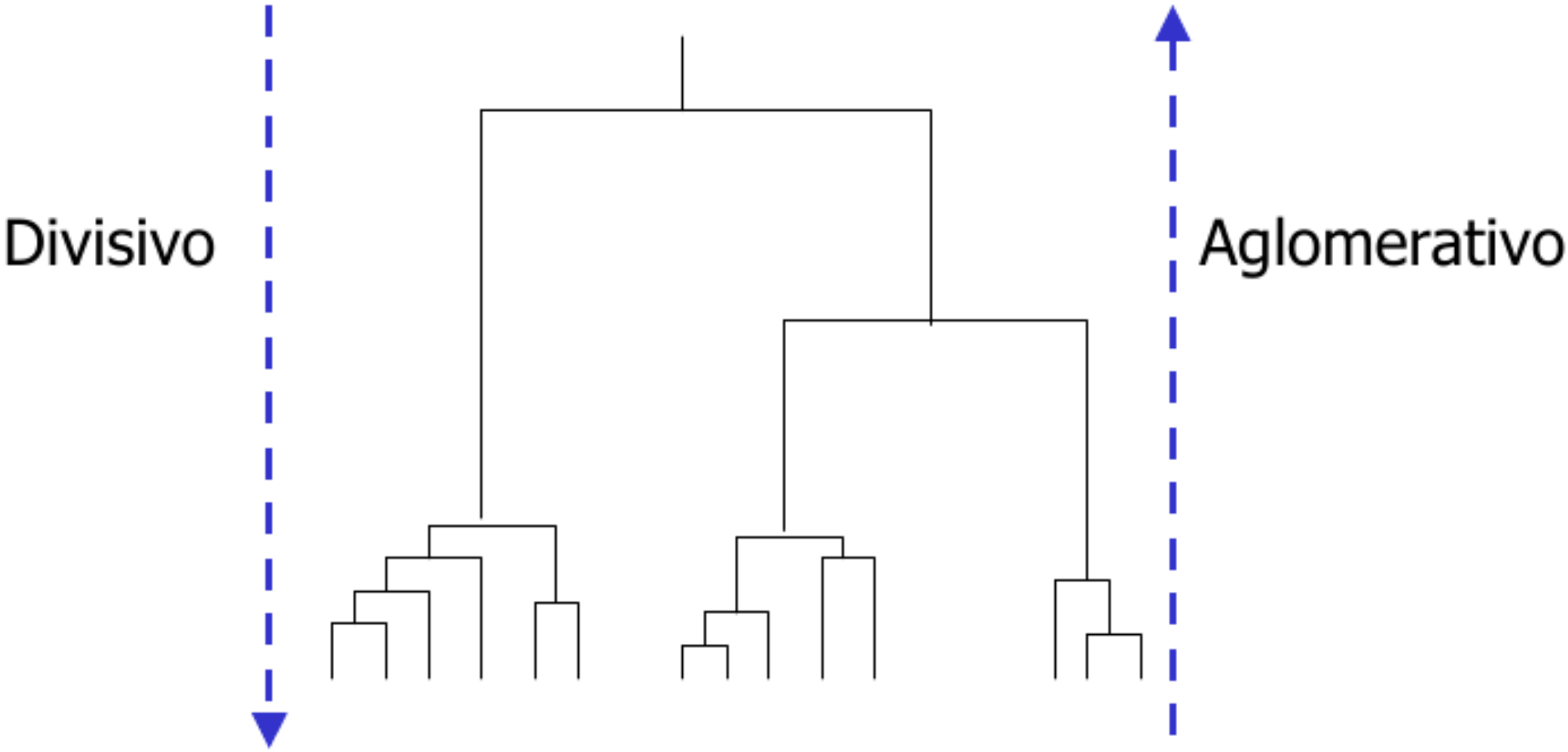
Algoritmos Hierárquicos (3)

- Aplicações na biologia geralmente não se preocupam com o número ótimo de clusters
  - Biólogo geralmente está interessado na estrutura da árvore completa

Algoritmos Hierárquicos (4)

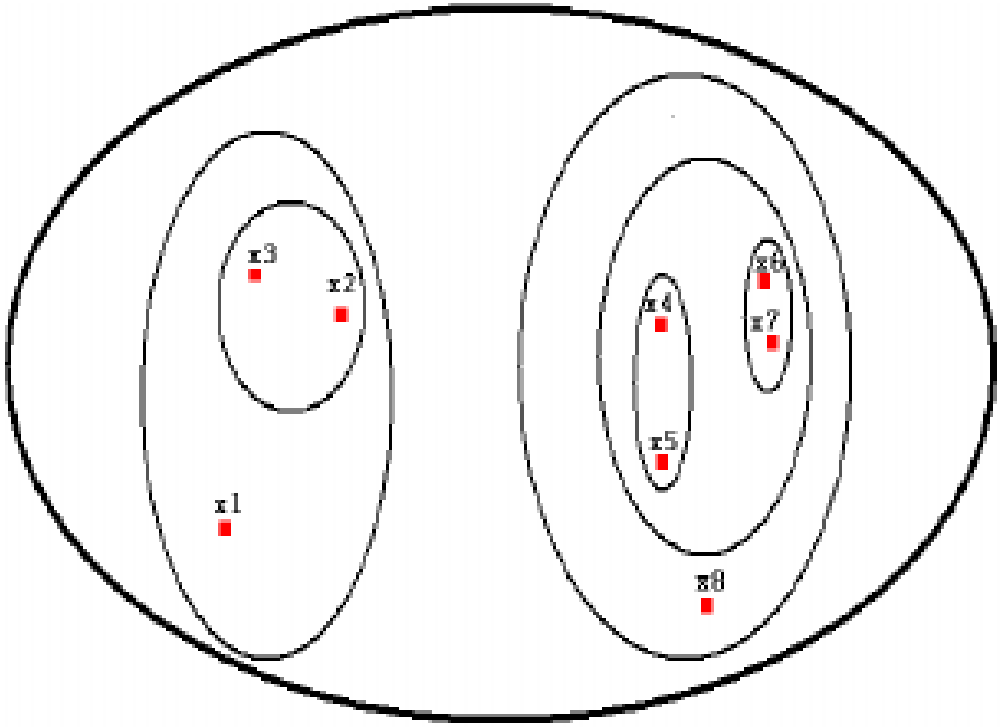
- Podem ser de dois tipos:
  - **Aglomerativos**: combinam, repetidamente, dois clusters em um
    - A cada passo, combina os dois clusters mais próximos
  - **Divisivos**: dividem, repetidamente, um cluster em dois
    - A cada passo, divide o cluster menos homogêneo em dois novos clusters

Exemplo



Exemplo (2)

- Não precisa ser apenas dendograma
  - Diagrama de Venn



Algoritmos Hierárquicos

- Definições:
  - Seja $P_t = \{C_1, C_2, ..., C_m\}$ uma partição no nível t de $X = \{\hat{x}_1, \hat{x}_2, ..., \hat{x}_n\}$
    - C_t é um agrupamento crisp
  - Diz-se que P_t é encaixado em P_t' ($P_t \subset P_t'$) se:
    - Cada conjunto em P_t é um subconjunto de um cluster em P_t' e
    - Pelo menos um cluster em P_t é um subconjunto próprio de algum cluster em P_t'

Exemplo

- Sejam:
  - $P_A = \{\{x_1, x_3\}, \{x_4\}, \{x_2, x_5\}\}$
  - $P_B = \{\{x_1, x_3, x_4\}, \{x_2, x_5\}\}$
  - $P_C = \{\{x_1, x_4\}, \{x_3\}, \{x_2, x_5\}\}$
  - Pode-se dizer que:
    - $P_A \quad P_B$
    - $P_A \quad P_C$
    - $P_A \quad P_A$

Exemplo (2)

- Sejam:
  - $P_A = \{\{x_1, x_3\}, \{x_4\}, \{x_2, x_5\}\}$
  - $P_B = \{\{x_1, x_3, x_4\}, \{x_2, x_5\}\}$
  - $P_C = \{\{x_1, x_4\}, \{x_3\}, \{x_2, x_5\}\}$
  - Pode-se dizer que:
    - $P_A \subset P_B$
    - $P_A \not\subset P_C$
    - $P_A \not\subset P_A$

Algoritmos Hierárquicos

- **Algoritmos aglomerativos**
  - Começam com $P_0 = \{\{\hat{x}_1\}, ..., \{\hat{x}_n\}\}$
  - A cada passo t, combinam dois clusters em um, produzindo:
    - $|P_{t+1}| = |P_t| - 1$ e $P_t \subset P_{t+1}$
  - No passo final (passo n-1) tem-se a hierarquia:
    - $P_{n-1} = \{\{\hat{x}_1, ..., \hat{x}_n\}\}$

Algoritmos Hierárquicos (2)

- **Algoritmos divisivos**
  - Começam com $P_0 = \{\{x_1, ..., x_n\}\}$
  - A cada passo t, dividem um cluster em dois, produzindo:
    - $|P_{t+1}| = |P_t| + 1$ e $P_{t+1} \subset P_t$
  - No passo final (passo n-1) tem-se a hierarquia:
    - $P_{n-1} = \{\{x_1\}, ..., \{x_n\}\}$

Esquema Aglomerativo Generalizado (EAG)

1 Inicializar $P_0 = \{\{x_1\}, \ldots, \{x_n\}\}$, $t = 0$

2 Para t = 1 até n – 1 faça

Encontrar o par de clusters mais próximos (C_i, C_j)

$P_t = (P_{t-1} - \{C_i\} - \{C_j\}) \cup \{\{C_i \cup C_j\}\}$ // atualizar centros

// se medida de similaridade for usada,
// trocar mais próximos por mais distantes
// Número de chamadas a $d(C_i, C_j)$ é $O(n^3)$
// Esse número pode ser reduzido

Esquema Aglomerativo Generalizado (EAG) (2)

- Dois métodos de implementação comuns são baseados em:
  - Matrizes (**foco**)
  - Teoria dos grafos
- Uma matriz de proximidade *(n-t) x (n–t)*, P_t, fornece a proximidade entre todos os pares de clusters em um nível *t*

Exemplo

- Sejam os dados
- $X_1 = [1, 1]^t$,
- $X_2 = [2, 1]^t$,
- $X_3 = [5, 4]^t$,
- $X_4 = [5, 5]^t$,
- $X_5 = [6.5, 6]^t$,

$$p_0^{SM} = \begin{bmatrix} 1 & 0.75 & 0.26 & 0.21 & 0.18 \\ 0.75 & 1 & 0.44 & 0.35 & 0.20 \\ 0.26 & 0.44 & 1 & 0.96 & 0.90 \\ 0.21 & 0.35 & 0.96 & 1 & 0.98 \\ 0.18 & 0.20 & 0.90 & 0.98 & 1 \end{bmatrix}$$

SM: medida de similaridade

Exemplo (2)

- Sejam os dados
- $X_1 = [1, 1]^t$,
- $X_2 = [2, 1]^t$,
- $X_3 = [5, 4]^t$,
- $X_4 = [5, 5]^t$,
- $X_5 = [6.5, 6]^t$,

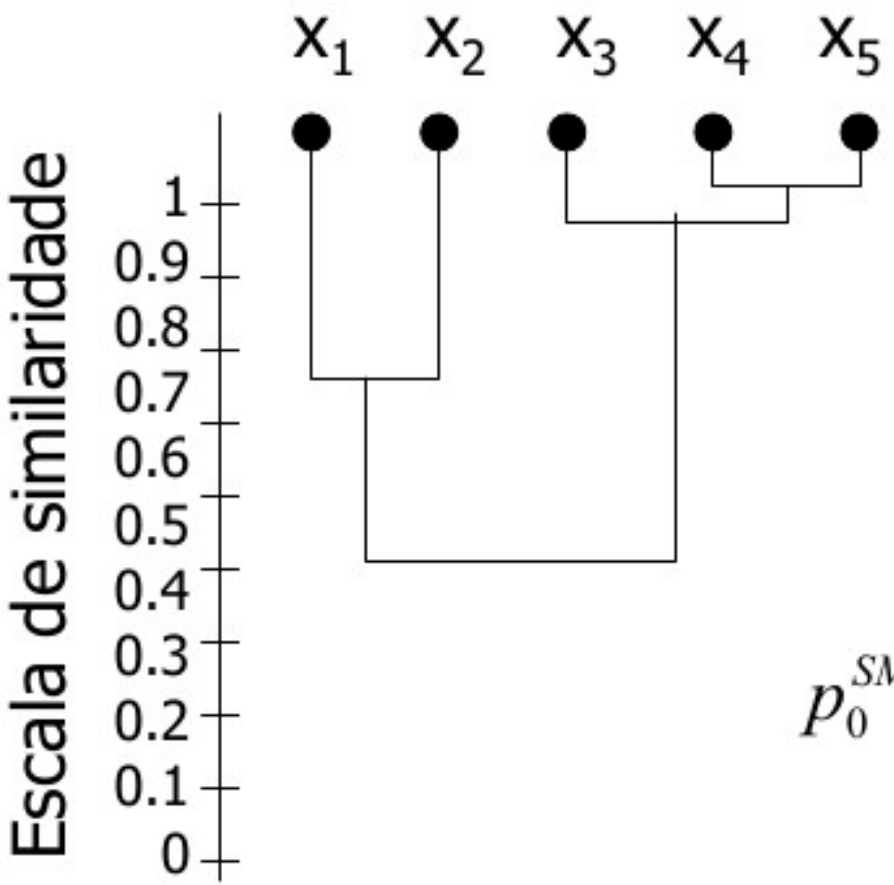
$$p_0^{DM} = \begin{bmatrix} 0 & 1 & 5 & 6,4 & 7,4 \\ 1 & 0 & 4,2 & 5,7 & 6,7 \\ 5 & 4,2 & 0 & 1,4 & 2,5 \\ 6,4 & 5,7 & 1,4 & 0 & 1,1 \\ 7,4 & 6,7 & 2,5 & 1,1 & 0 \end{bmatrix}$$

DM: medida de dissimilaridade

Algoritmos Hierárquicos

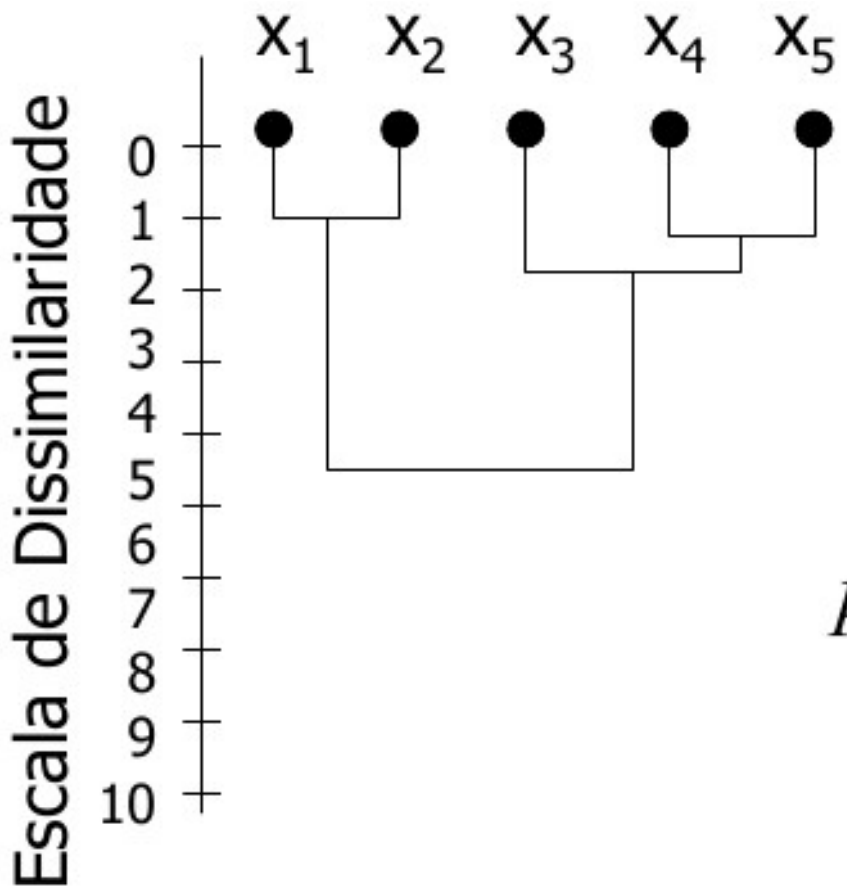
- Dendograma de proximidade: árvore que indica hierarquia de partições
  - Incluindo a proximidade entre dois clusters e quando eles são combinados
  - O corte de um dendograma em qualquer nível produz uma simples partição

Exemplo



$$p_0^{SM} = \begin{bmatrix} 1 & 0.75 & 0.26 & 0.21 & 0.18 \\ 0.75 & 1 & 0.44 & 0.35 & 0.20 \\ 0.26 & 0.44 & 1 & 0.96 & 0.90 \\ 0.21 & 0.35 & 0.96 & 1 & 0.98 \\ 0.18 & 0.20 & 0.90 & 0.98 & 1 \end{bmatrix}$$

Exemplo (2)

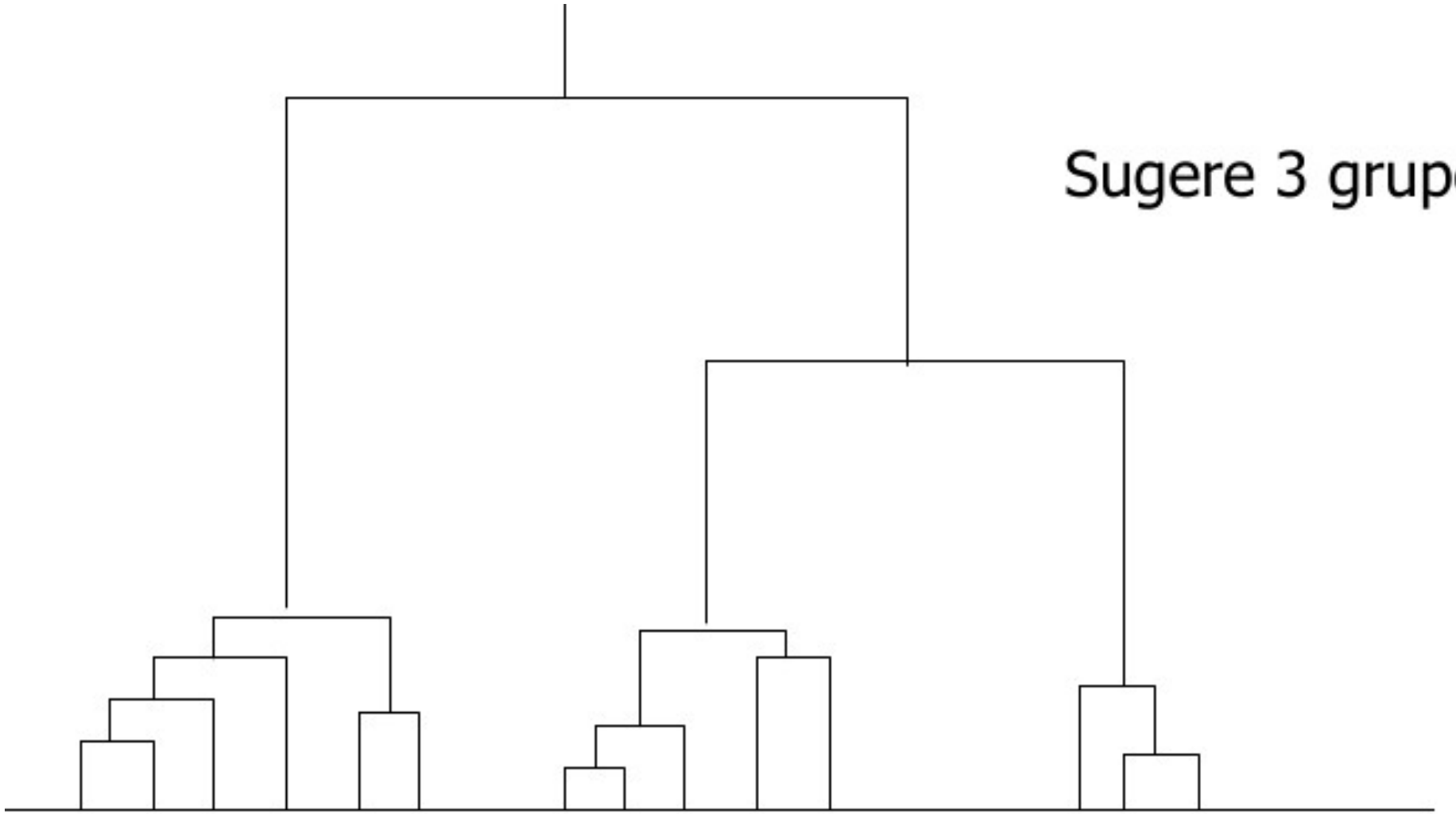


$$p_0^{DM} = \begin{bmatrix} 0 & 1 & 5 & 6,4 & 7,4 \\ 1 & 0 & 4,2 & 5,7 & 6,7 \\ 5 & 4,2 & 0 & 1,4 & 2,5 \\ 6,4 & 5,7 & 1,4 & 0 & 1,1 \\ 7,4 & 6,7 & 2,5 & 1,1 & 0 \end{bmatrix}$$

Algoritmos Hierárquicos

- Como escolher uma partição?
  - **Partição com n clusters**
    - Selecionando partição com *n* clusters na sequência de agrupamentos da hierarquia
  - **Partição que melhor se encaixa nos dados**
    - Procurar no dendograma grandes mudanças em níveis adjacentes
    - Nesse caso, uma mudança de *j* para *j-1* grupos pode indicar que *j* é o melhor número de grupos
    - Existem outros procedimentos, alguns mais objetivos

Exemplo



Algoritmos Hierárquicos

- Outra alternativa
- Usar medida de auto-similaridade de um cluster C_t
  - Interromper processo quando a distância entre as instâncias em algum dos clusters for maior que um valor θ

Algoritmos Hierárquicos (2)

- Existe uma grande variedade de algoritmos hierárquicos
- Geralmente diferem na forma de calcular distância inter-clusters

$$d_{AB} = \min_{\substack{i \in A \\ j \in B}} (d_{ij})$$ Por ligação simples (single-link)

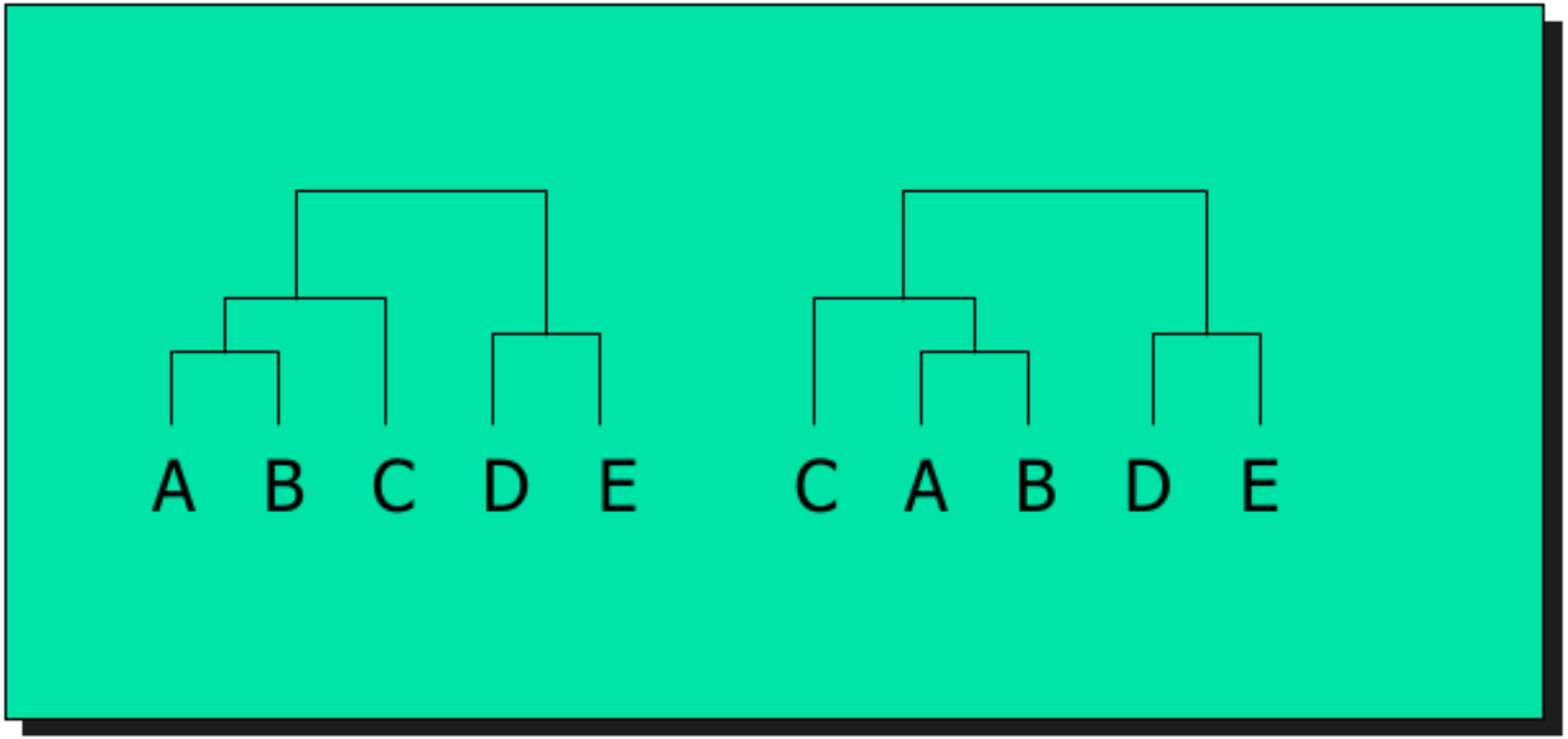
$$d_{AB} = \max_{\substack{i \in A \\ j \in B}} (d_{ij})$$ Por ligação completa (complete-link)

$$d_{AB} = \frac{1}{n_A n_B} \sum_{i \in A} \sum_{j \in B} d_{ij}$$ Pela média do grupo (average-link)

Algoritmos Hierárquicos (3)

- Deve ser observado que o desenho do dendograma é arbitrário
- Clusters podem ser rotacionados no ponto de bifurcação
  - Afeta a proximidade aparente entre fronteiras de clusters adjacentes
  - Mas a informação importante está contida no conteúdo do cluster e na sua similaridade

Algoritmos Hierárquicos (4)



Algoritmos Hierárquicos (5)

- E para calcular a distância?
  - Existem várias métricas
    - Distância Euclidiana
    - Distância Manhattan (bloco-cidade)
    - Distância quadrática
    - Distância de Mahalanobis
    - ...

Avançado: algoritmos baseados em otimização de função de custo

- Família de algoritmos crescentemente popular
- Definir uma função de custo *f* (Φ) que mede a "qualidade" da partição
  - Ex.: Vetor de parâmetros $\Phi = [v_1^t, ..., v_m^t]t$
  - Buscar por valores de Φ que minimizam / maximizam *f* (Φ)
- Em geral, os algoritmos assumem que m é conhecido

Avançado: algoritmos baseados em otimização de função de custo (2)

- Algoritmos
  - Algoritmo c-médias crisp (mais famoso)
  - Algoritmo c-médias fuzzy
  - Isodata
  - Otimização iterativa
  - Algoritmo probabilístico

Tendência e validação

- Tendência de agrupamento (**antes de agrupar**):
  - Testes estatísticos ajudam a verificar que existe nos dados uma estrutura significativa (não aleatória)
- Validação de agrupamento (**após agrupar**):
  - Estima o desempenho do algoritmo
  - Utiliza testes estatísticos (objetivo) e percepção do especialista (subjetivo)

Validação de Agrupamentos

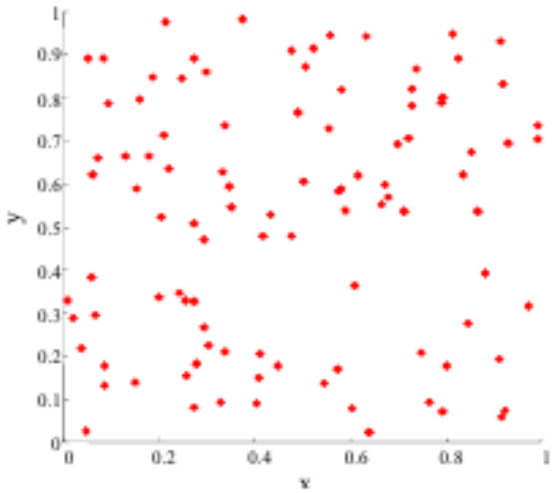
- Existem várias medidas para avaliar qualidade de classificadores
  - Acurácia, precisão, revocação, F1
- Como avaliar os clusters gerados por um algoritmo de agrupamento?

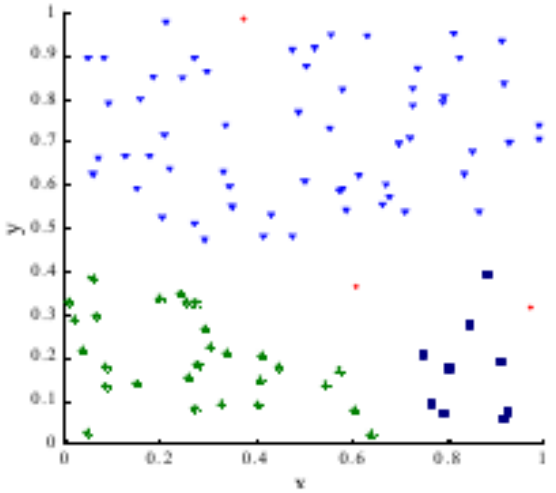
Validação de Agrupamentos (2)

- Por que avaliar agrupamentos?
  - Para evitar encontrar padrões em ruídos
  - Para comparar algoritmos de agrupamento
  - Para comparar duas partições
  - Para comparar dois grupos

Partições de Dados Aleatórios

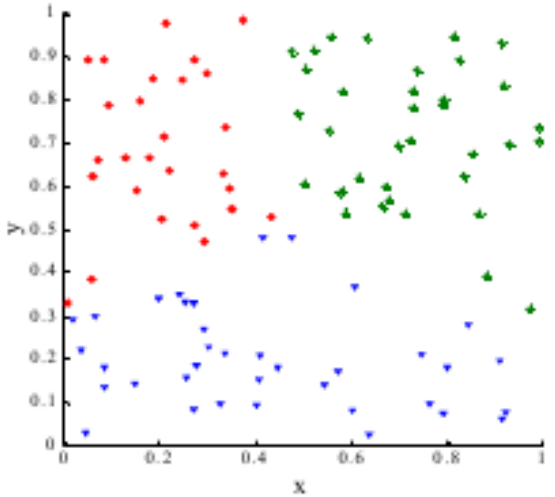
Pontos Aleatórios

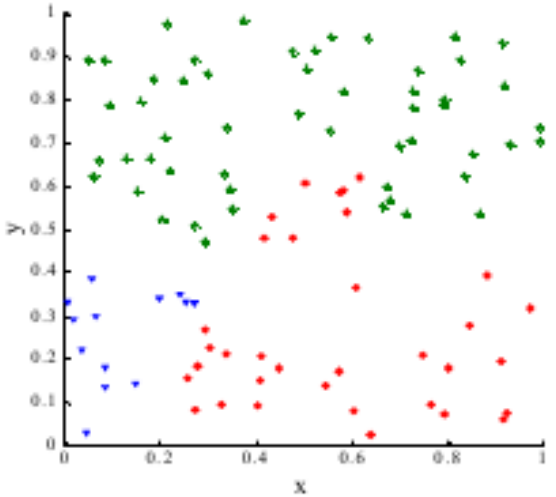




DBSCAN

K-médias





Complete-Link

Medidas de Validação

- As medidas julgam aspectos diferentes, podendo ser divididas em três grupos:
  - **Índices ou critérios externos**: medem o quanto os rótulos dos grupos casam com a classe verdadeira
    - Veremos alguns índices na aula de avaliação de classificadores
  - **Índices ou critérios internos**: medem a qualidade da partição obtida e pode ser usado para comparar duas partições

Medidas externas

- Medidas orientadas a similaridade: comparam clusters com classes
  - Casamento Simples (índice Rand)
  - Jackard

Medidas externas (2)

- Sejam
  - f_{00} = número de pares de instâncias com classes e clusters diferentes
  - f_{01} = número de pares de instâncias com classes diferentes e mesmo cluster
  - f_{10} = número de pares de instâncias com mesma classe e clusters diferentes
  - f_{11} = número de pares de instâncias com mesmas classes e clusters

$$Rand = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}} \qquad Jac = \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

Medidas externas (3)

- Índice Rand: similar a coeficiente de casamentos simples
  - Similaridade entre vetores binários
- Rand Corrigido (CR)
  - Leva aleatoriedade em consideração
  - Normaliza índice rand
    - 0 quando as partições são selecionadas ao acaso
    - 1 quando um casamento perfeito é obtido
    - Pode ser negativo

Medidas externas (4)

- Rand Corrigido
- Seja G = $\{g_1, g_2, \ldots, g_N\}$ a partição gerada
- Seja V = $\{v_1, v_2, \ldots, v_M\}$ a partição verdadeira

$$CR = \frac{\sum_i^N \sum_j^M \binom{n_{ij}}{2} - \binom{n}{2}^{-1} \sum_i^N \binom{n_{i.}}{2} \sum_j^M \binom{n_{.j}}{2}}{\frac{1}{2}\left[\sum_i^N \binom{n_{i.}}{2} + \sum_j^M \binom{n_{.j}}{2}\right] - \binom{n}{2}^{-1} \sum_i^N \binom{n_{i.}}{2} \sum_j^M \binom{n_{.j}}{2}}$$

n_{ij} = número de objetos nos clusters g_i e v_j

n = número total de objetos

Medidas Internas

- Coesão de clusters
  - Mede o quão relacionados estão as instâncias dentro de um cluster
- Separação de clusters
  - Mede quão distintos ou separados um cluster é dos demais clusters

Exemplo

- Usando soma dos erros quadráticos (SSE)
  - Coesão é medida pelo SSE dentro dos clusters

$$WSS = \sum_{i} \sum_{x \in C_i} (x - m_i)^2$$

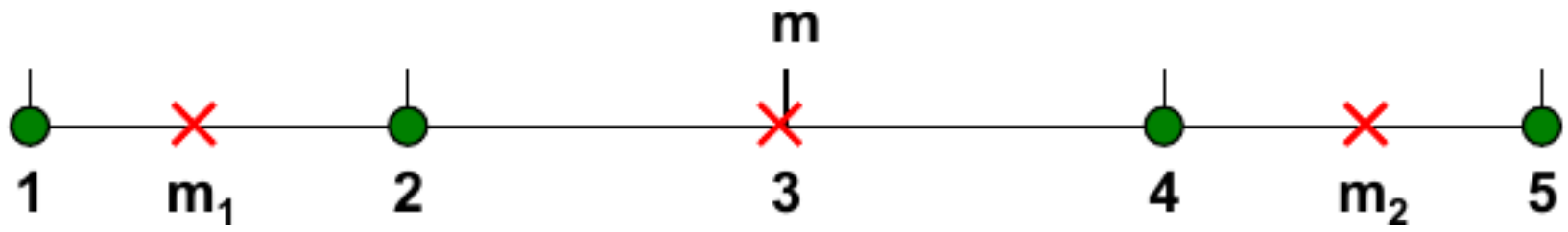
  - Separação é medida pelo SSE entre os clusters

$$BSS = \sum_{i} |C_i| (m - m_i)^2$$

$|C_i|$ é o tamanho do cluster C_i

- SSC + BCC = constante

Exemplo (2)



K=1 cluster:

$$WSS = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$
$$BSS = 4 \times (3-3)^2 = 0$$
$$Total = 10 + 0 = 10$$

K=2 cluster:

$$WSS = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$
$$BSS = 2 \times (3-1.5)^2 + 2x(4.5-3)^2 = 9$$
$$Total = 1 + 9 = 10$$

Medidas Internas

- Silhueta
  - Combina coesão com separação
  - Calculada para cada instância que faz parte de um agrupamento
  - Baseada na proximidade entre as instâncias de um cluster e na distância das instâncias de um cluster ao cluster mais próximo
  - Mostra quais instâncias estão bem situados dentro dos seus clusters e quais estão fora de um cluster apropriado

Medidas Internas (2)

- Silhueta: para cada instância **i**:
  - **a** = distância média de **i** aos outras instâncias de seu cluster
  - **b** = min (distância média de **i** às instâncias do cluster mais próximo, que não o seu próprio)
  - **s** = 1 – a/b se a < b
    = 0 se a = b
    = b/a – 1 se a > b
  - Valor entre -1 e 1 (quanto mais próximo de 1, melhor)
  - Largura média da silhueta: média sobre todos as instâncias do conjunto de dados

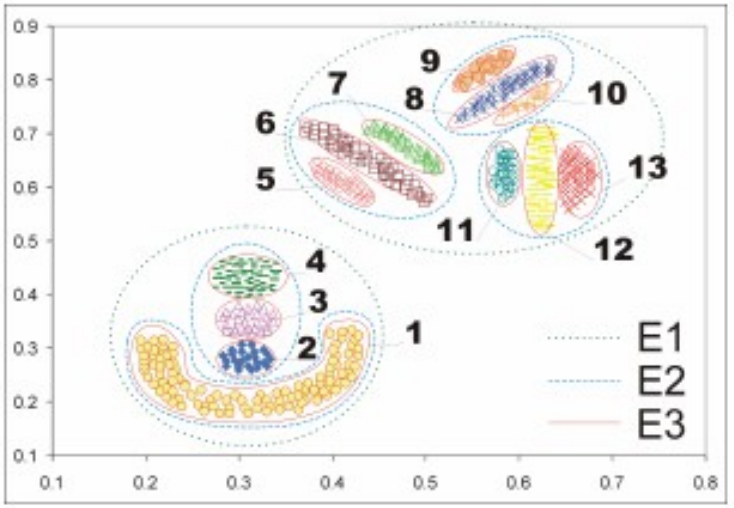
Exercício

- Considere um cluster para João e outro para Pedro em que cada qual é a média de seu cluster
  - Verifique a qual cluster Leila pertence
  - Calcule a silhueta para Leila

Nome	Febre	Enjôo	Manc.	Dores	Diagnóstico
João	sim	sim	peq.	Sim	doente
Pedro	não	não	gran.	não	saudável
Maria	sim	sim	peq.	não	saudável
José	sim	não	gran.	sim	doente
Ana	sim	não	peq.	sim	saudável
Leila	não	não	gran.	sim	doente

Dificuldades

- Um mesmo conjunto de dados pode ter mais de uma estrutura relevante
  - Análise de agrupamento tradicional busca por uma única estrutura dos dados
    - Limita a quantidade de conhecimento que poderia ser obtido



Combinação de Agrupamentos

- Objetivo: obter partições de melhor qualidade
- Medidas de qualidade:
  - Robustez frente a diferentes conformações dos dados
  - Novidade: partições novas que não poderiam ser obtida com nenhum algoritmo, individualmente
  - Estabilidade: obtém partições com menor sensibilidade a ruídos, outliers, variações de amostragem ou variabilidade dos algoritmos

Combinação de Agrupamentos (2)

- Vantagens:
  - Consistência com conjunto de partições iniciais
  - Computação distribuída, paralelismo e escalabilidade
  - Desempenho e custo: Uso de técnicas mais simples para construir as partições base
- Reuso de conhecimento

Aplicações

- Compressão (redução) de dados
  - Representa cada cluster como um único dado
- Formulação de hipóteses sobre a natureza dos dados
- Teste de hipóteses sobre os dados
  - Que características são correlacionadas
  - Que características são independentes
- Predição baseada em grupos

Pontos chaves

- Agrupamento particional e k-médias
- Agrupamento hierárquico aglomerativo e esquema aglomerativo generalizado
- Agrupamento hierárquico aglomerativo
- Critérios de avaliação internos, externos e relativos
- Dendograma e matrizes de (dis)similaridade

Agradecimentos/referências

- Notas de aula do Prof. André de Carvalho (USP)