Aula 9 – Mineração de Dados
Agrupamento: Algoritmos Particionais

Profa. Elaine Faria

UFU

Material

- Este material é baseado
  - No livro Tan et al, 2006
  - Nos slides do prof Andre

- Agradecimentos
  - Ao professor André C. P. L. F. Carvalho que gentilmente cedeu seus slides

Agrupamento

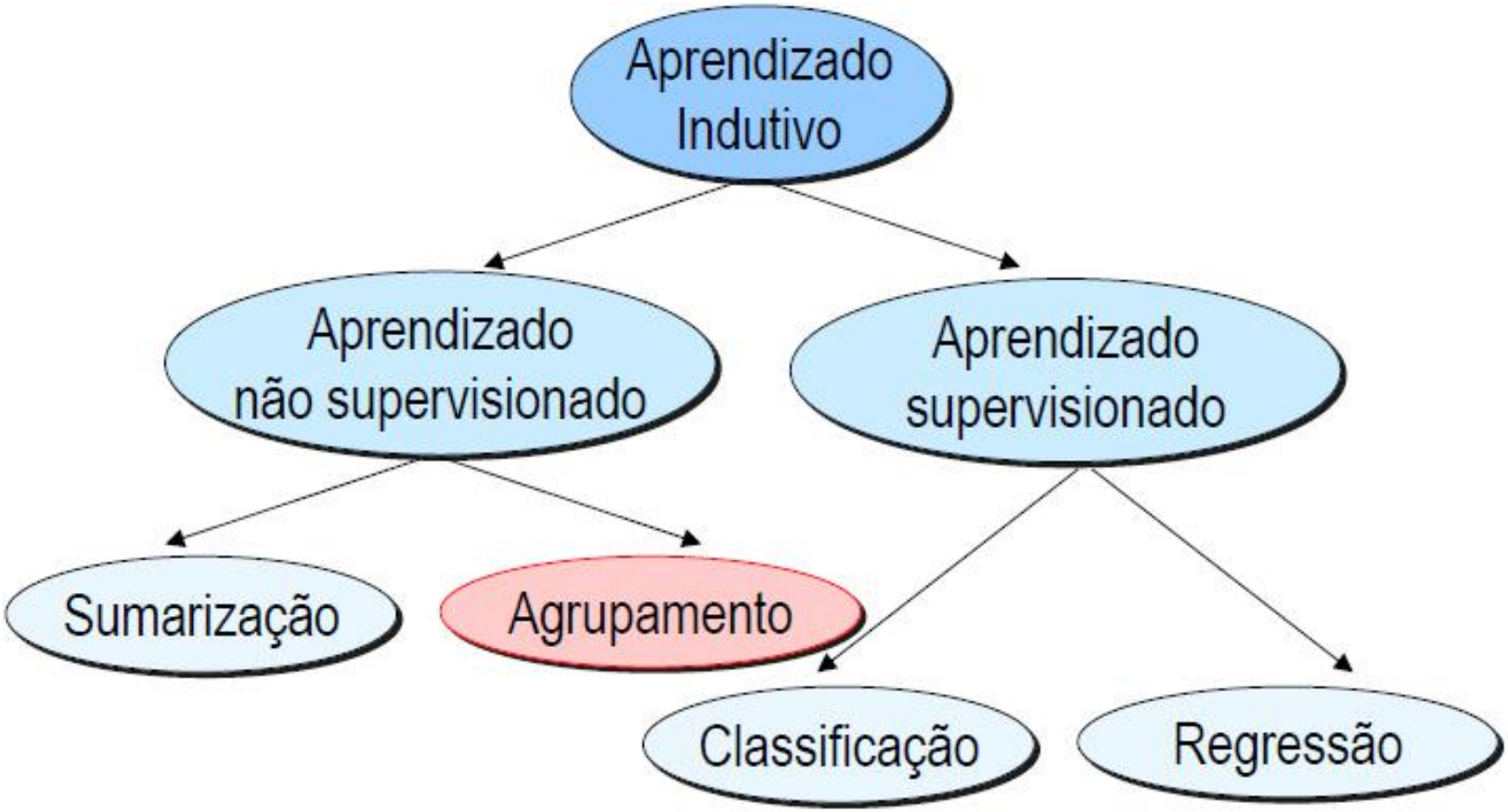


Figura retirada dos slides do prof. André C. P. L. F. Carvalho – disciplina Aprendizado de Máquina – ICMC-USP

Agrupamento

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

Agrupamento

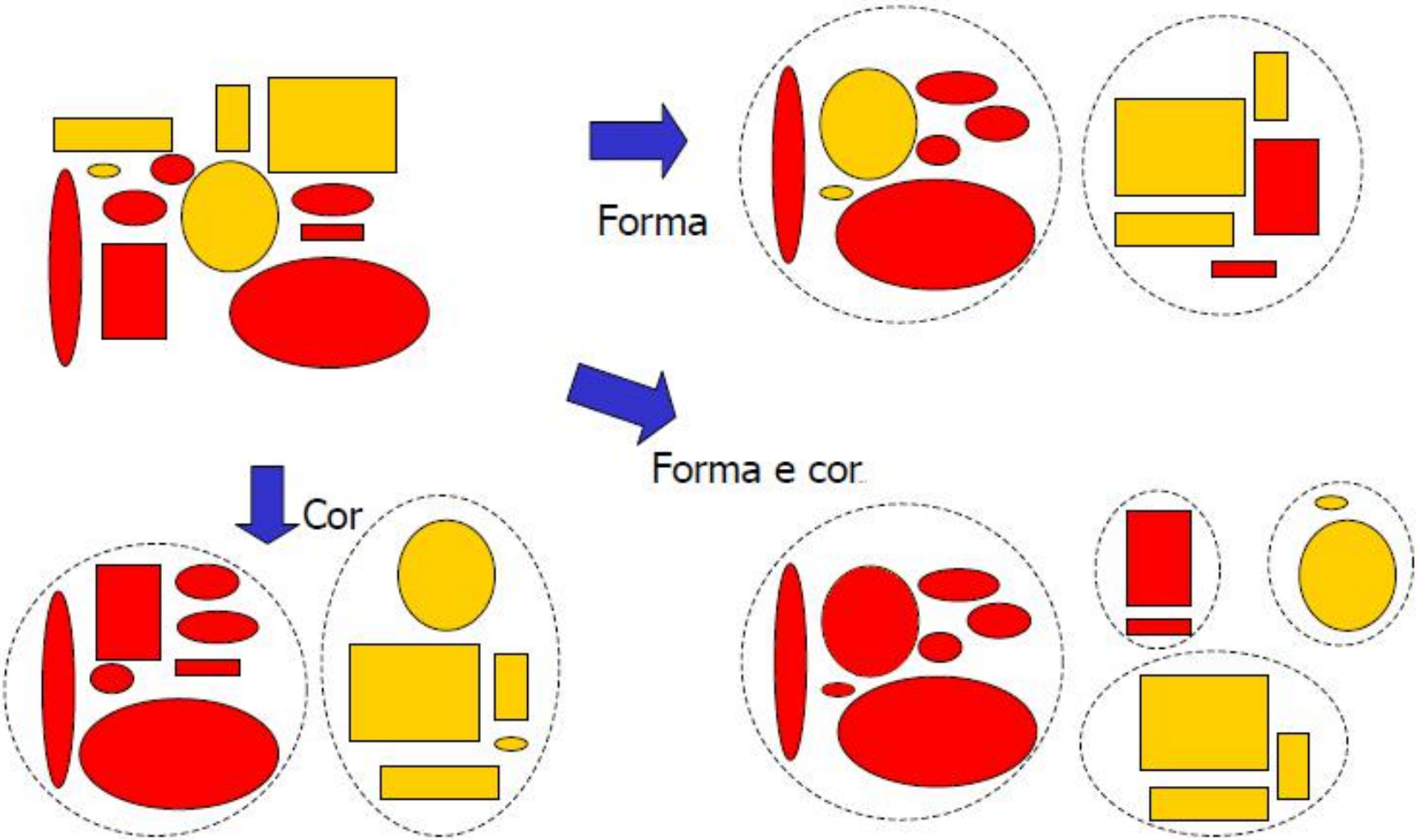


Figura retirada dos slides do prof. André C. P. L. F. Carvalho – disciplina Aprendizado de Máquina – ICMC-USP

Objetivo da tarefa de agrupamento

- Agrupar dados em grupos que
  - Possuem um significado
    - Grupos devem capturar a estrutura natural dos dados
    - Entendimento
  - Sejam úteis
    - Grupos podem ser um passo inicial para outros propósitos
    - Ex. Sumarização de dados, compressão de dados

Entendimento

- Grupos (clusters) são potenciais classes
- Análise de cluster
  - Estudo de técnicas para encontrar classes automaticamente

Análise de Clusters

- Agrupa objetos utilizando apenas informações sobre objetos e seus relacionamentos
- Objetivo
  - Objetos dentro de um grupo são semelhantes e em grupos diferentes são distintos
- Quanto maior a similaridade (homogeneidade) dentro de um grupo e maior a diferença entre grupos → melhor o agrupamento
- Em várias aplicações, noções do que é um cluster não esta bem definida

Análise de Clusters

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

Análise de Clusters

- Algoritmos de agrupamento
  - São uma importante ferramenta de análise de dados
  - Agrupam exemplos semelhantes de acordo com alguma medida de similaridade
  - Chamado de aprendizado de máquina não supervisionado
  - Detectam diferentes estruturas para diferentes valores dos parâmetros

Principais Etapas

1. Pré-processamento
   - Seleção de características
   - Normalização pode garantir que todas as características são tratadas igualmente
2. Definição de medida de similaridade
   - Mede similaridade/dissimilaridade entre:
     - Dois exemplos
     - Exemplo e conjunto de exemplos (grupo)
     - Dois conjuntos de exemplos (grupos)

Principais Etapas

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

Figura retirada dos slides do prof. André C. P. L. F. Carvalho – disciplina Aprendizado de Máquina – ICMC-USP

Principais Etapas

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

Principais Etapas

- Cuidado!!!
  - Cada passo é subjetivo e influenciado pela experiência e conhecimento do especialista

Tipos de Agrupamento

- Seja X = {x_1, x_2,..., x_n} o conjunto de todos os dados
  - Tarefa: colocar cada Xi 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

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

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

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

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

Tipos de Agrupamento

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

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

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

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

m = número de grupos
n = número de objetos

Matriz de Partição

- Matriz com K linhas (nro de grupos) e N colunas (nro de objetos) no qual cada elemento X_{ij} indica o grau de pertinência do j-ésimo elemento ao i-ésimo grupo

- Matriz rígida (sem sobreposição)
  - Se a matriz for binária
  - Se restrição for satisfeita $\sum_i X_{ij} = 1$

Diferentes Tipos de Agrupamento

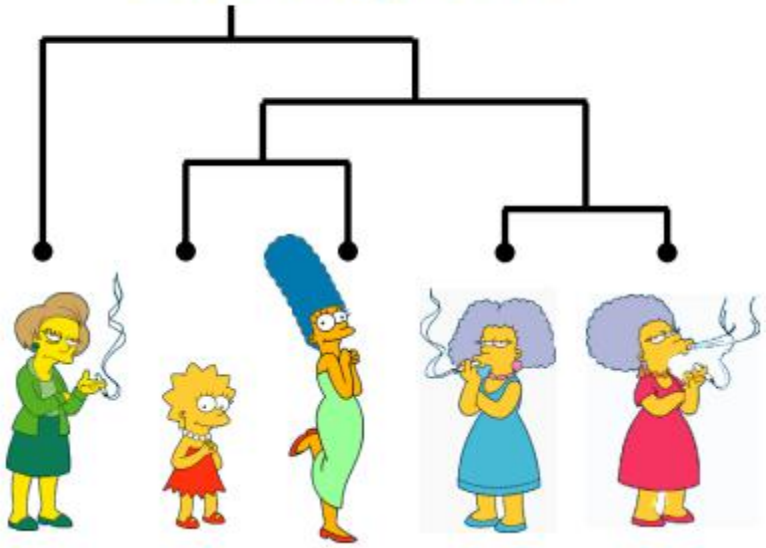
- Agrupamento Particional
  - Divisão do conjunto de dados em grupos (não sobrepostos) tal que cada objeto está em exatamente um grupo

x

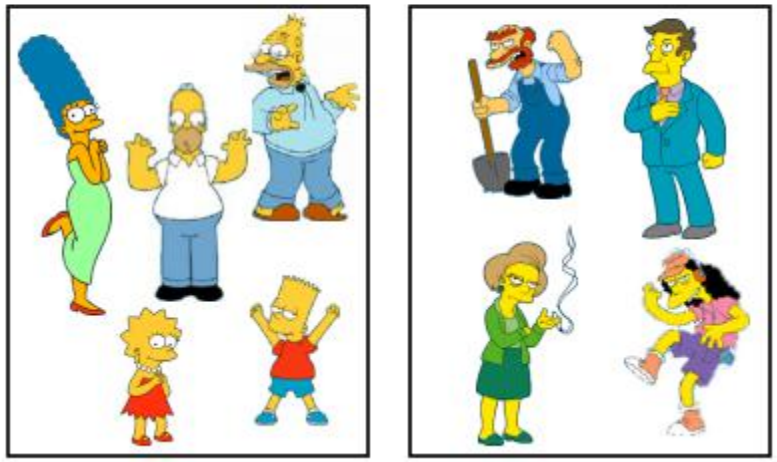
- Agrupamento Hierárquico
  - Conjunto de grupos aninhados que estão organizados como uma árvore
  - Cada nó (grupo) na árvore (exceto as folhas) é a união de dos seus filhos (subgrupos)
  - A raiz contém todos os objetos da base

Diferentes Tipos de Agrupamento

Hierárquicos



Particionais



Keogh, E. A g. Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Diferentes Tipos de Agrupamento

- Exclusivo
  - Associa cada objeto a um único cluster

X

- Sobreposição (não exclusivo)
  - Um objeto pode pertencer simultaneamente a mais que um grupo

X

- *Fuzzy*
  - Cada objeto pertence a cada grupo com um grau de pertinência entre 0 e 1

Diferentes Tipos de Agrupamento

- Completo
  - Associa cada objeto a um cluster

X

- Parcial
  - Não associa cada objeto a um cluster
  - Motivação: alguns objetos no conjunto de dados podem não pertencer a grupos bem definidos
    - Ex: ruídos ou *outliers*

Diferentes Tipos de Grupos

- Bem separados
  - Conjunto de objetos no qual cada objeto está mais próximo a todo objeto no seu grupo do que a qualquer objeto que não está no seu grupo
    - Definição é satisfeita quando os dados contém uma estrutura natural de grupos
- Baseado em Protótipo
  - Conjunto de objetos no qual cada objeto está mais próximo ao protótipo que define um grupo do que ao protótipo de qualquer outro grupo
    - Protótipo: por exemplo, o centróide

Diferentes Tipos de Grupos

- Baseado em Grafos
  - Dados são representados como grafos
  - Os nós são os objetos e as arestas são as conexões entre os objetos
  - Cluster: é um componente conectado
- Baseado em Densidade
  - Região densa de objetos que é circundada por uma região de baixa densidade

Algoritmos Particionais

- Características
  - São baseados na minimização de uma função de custo
  - Objetos agrupados em um número K de grupos
  - Cada objeto é agrupado no grupo que minimiza a função de custo
  - Uma única partição é obtida
- Vantagem
  - Um objeto pode mudar de grupo ao longo do agrupamento

Algoritmos Particionais

Como evitar que o agrupamento vire um problema combinatorial?

Algoritmos Particionais

- Exemplos
  - K-means
  - SOM
  - DENCLUE
  - CLICK
  - CAST

k-Means

- Algoritmo Básico
  - Escolher K centróides
    - K é um parâmetro especificado pelo usuário
    - K representa o nro de grupos
  - Cada objeto é associado ao seu centróide mais próximo
    - Cada coleção de objetos associados a um centróide forma um grupo
  - Recalcule os centros dos grupos
  - Repetir os dois passos anteriores até que não haja mudança nos grupos ou equivalentemente, até que os centróides permaneçam o mesmo

k-Means

Selecione K objetos como centróides

Repita

Forme K grupos associando cada objeto ao seu centróide mais próximo

Recalcule os centróides de cada grupo

Até que Convergência seja obtida

k-Means

- Videos no youtube sobre K-Means
  - https://www.youtube.com/watch?v=IuRb3y8qKX4
  - https://www.youtube.com/watch?v=zHbxbb2ye3E

k-Means

- Centróides iniciais
  - Ex de Técnica: Escolher aleatoriamente objetos do conjunto de dados
- Associar um objeto ao seu grupo mais próximo
  - Usar uma medida de proximidade que quantifica a noção de mais próximo
  - Ex: usar distância Euclidiana

k-Means

- Critérios de Convergência
  - Número máximo de iterações é obtido
  - Limiar mínimo de mudanças nos centróides

k-Means

- Função Objetivo
  - Objetivo do agrupamento

    Minimizar a distância quadrada de cada objeto ao seu centróide mais próximo

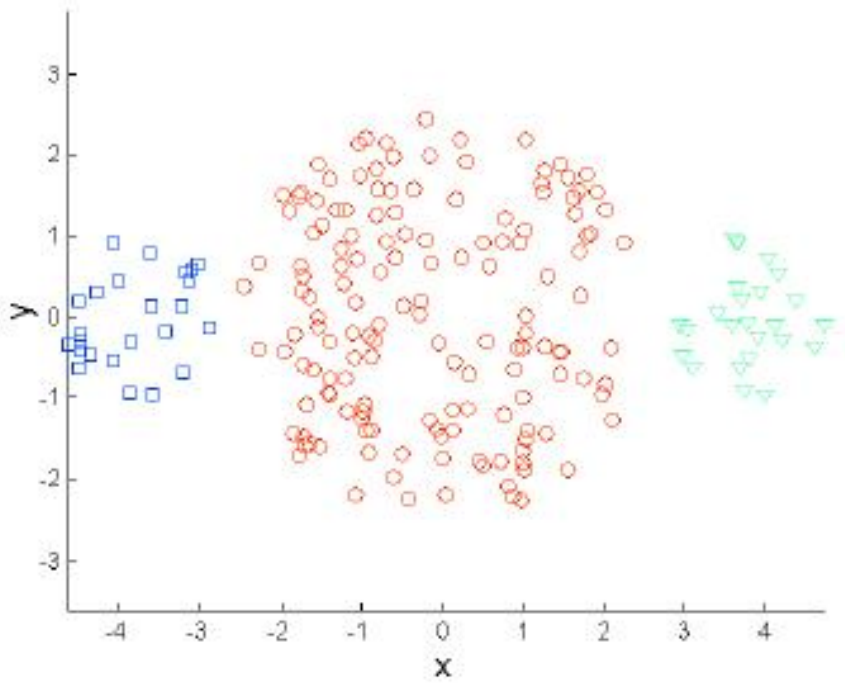
$$J = \sum_{c=1}^{k} \sum_{x_j \in C_c} d(x_j, \bar{x}_c)^2$$

d: distância Euclidiana

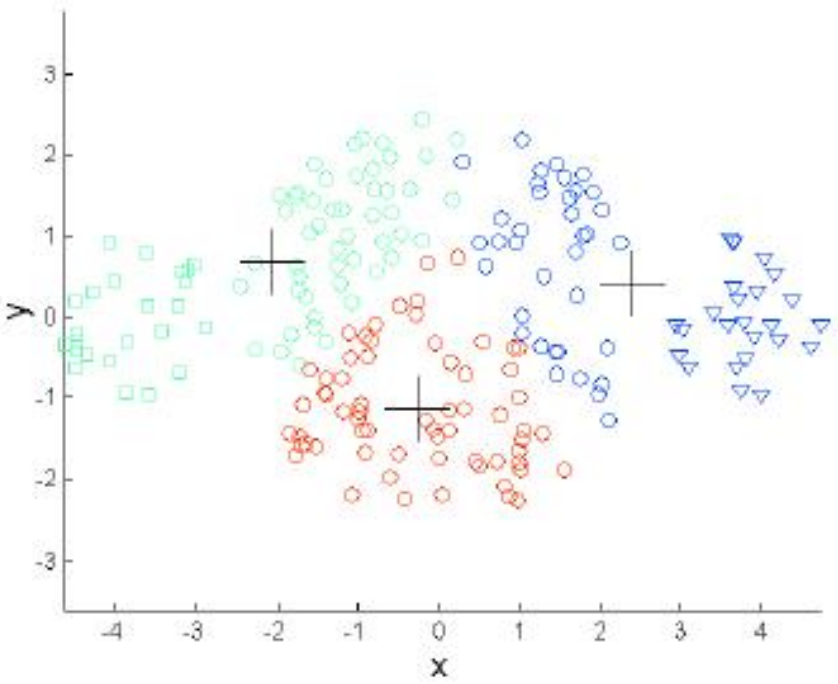
k-Means

- Limitações
  - Escolha do valor de K
  - Problemas quando os grupos têm
    - Diferentes densidades
    - Formatos não hiper-esféricos
  - Problemas quando os dados possuem *outliers*

k-Means – Grupos com diferentes tamanhos



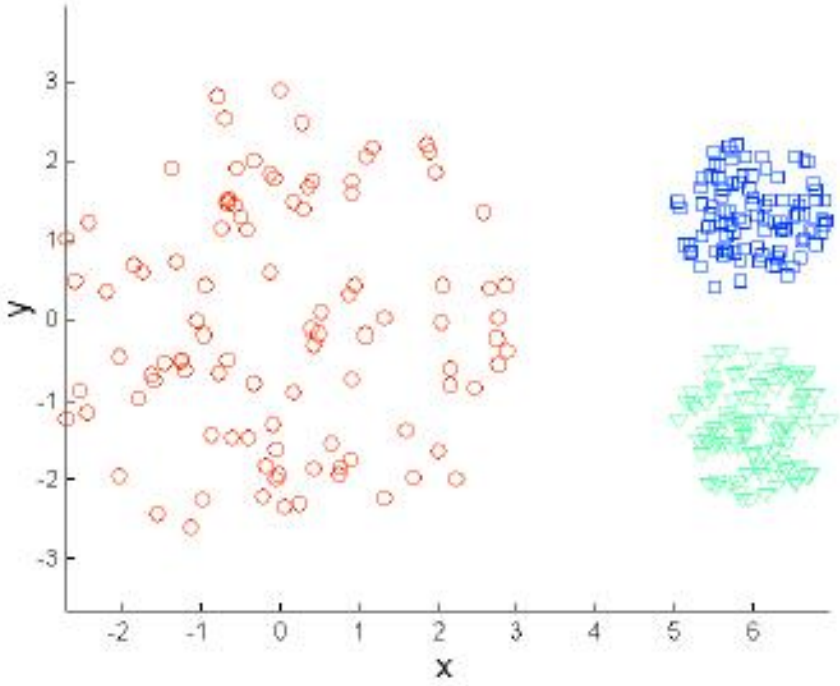
Dados originais



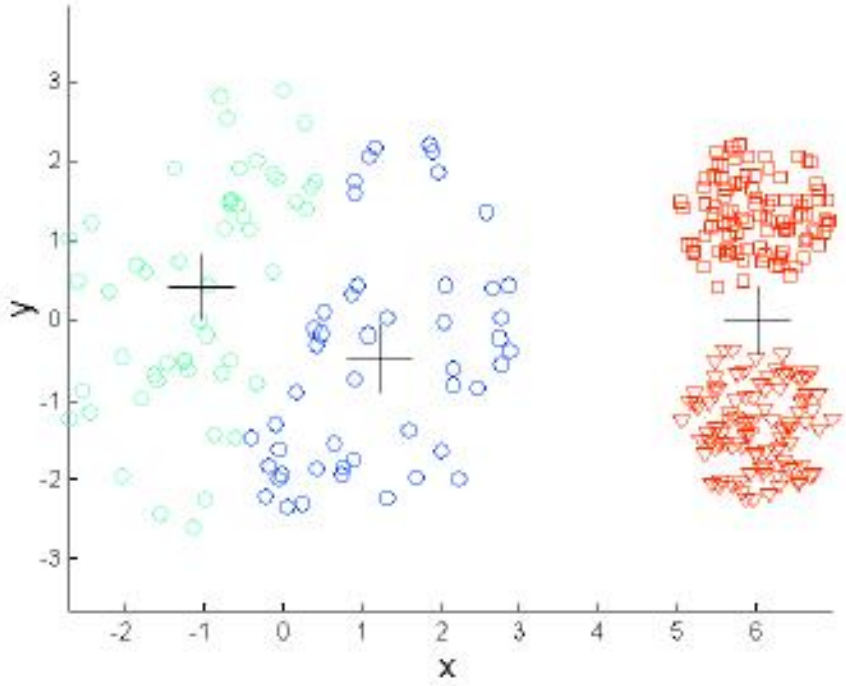
K-médias (3 Clusters)

Figura retirada dos slides do prof. André C. P. L. F. Carvalho – disciplina Aprendizado de Máquina – ICMC-USP

k-Means – Grupos com diferentes densidades



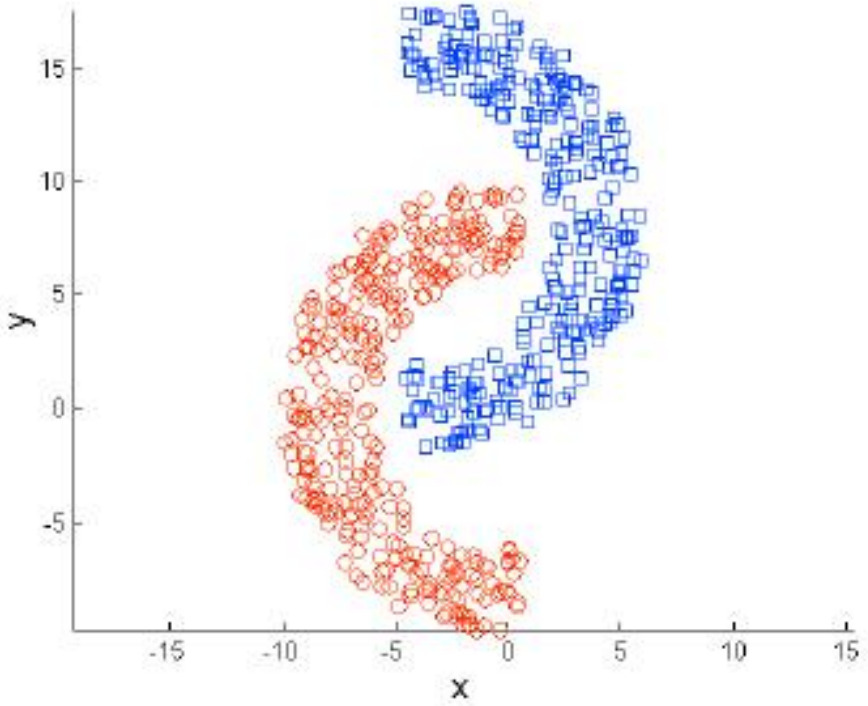
Dados originais



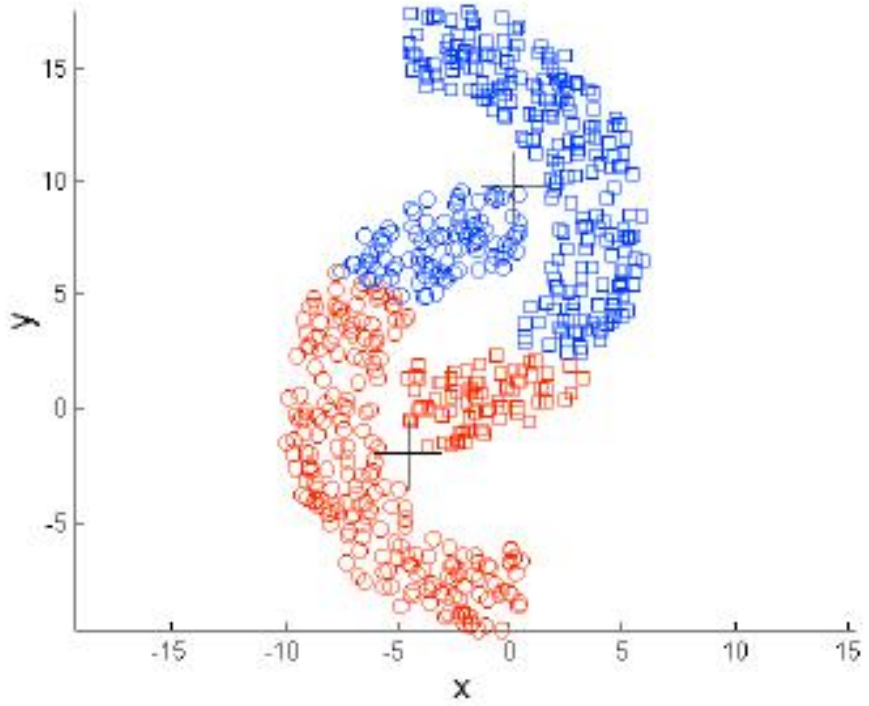
K-médias (3 Clusters)

Figura retirada dos slides do prof. André C. P. L. F. Carvalho – disciplina Aprendizado de Máquina – ICMC-USP

k-Means – Grupos não globulares



Dados originais



K-médias (2 Clusters)

Figura retirada dos slides do prof. André C. P. L. F. Carvalho – disciplina Aprendizado de Máquina – ICMC-USP

K-Means

Qual a complexidade do K-Means?

O(N*d*K*I)

N → nro de elementos da base de dados

d → dimensão

K → nro de grupos

I → nro de iterações

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

Figura retirada dos slides do prof. André C. P. L. F. Carvalho – disciplina Aprendizado de Máquina – ICMC-USP

Exercício

- Agrupar os dados em dois grupos usando o algoritmo K-means
  - Usar k = 2
  - Informação sobre a classe não é usada
- Em que grupos seriam colocados os novos casos?
  - (Luis, não, não, pequenas, sim)
  - (Laura, sim, sim, grandes, sim)

K-medóides

- Ao invés de calcular a média dos elementos em um grupo, um elemento representativo, medóide, é escolhido a cada iteração para cada grupo
- Medóides são calculados encontrando o objeto *i* em um grupo que minimiza

$$\sum_{j \in C_i} d(i, j)$$

K-medóides

- Vantagem
  - É menos sensível a *outliers*
  - Pode ser aplicado a bases com atributos categóricos

Obs.: Não há necessidade de calcular repetidamente as distâncias a cada iteração

- Basta apenas obter as distâncias a partir da matriz de distâncias

K-medóides

1. Escolha k objetos aleatoriamente para serem os medóides dos grupos
2. Associe cada objeto ao grupo como medóide mais próximo
3. Recalcule as posições dos K-medóides
4. Repita os passos 2 e 3 até que os medóides não mudem

K-medóides

- Como calcular o passo 3
  - Para cada objeto do grupo calcular a soma das distâncias dele a todos os outros objetos do grupo
    - Escolher o objeto com a menor soma
  - Como muitos objetos permanecem no mesmo grupo de uma iteração para outra
    - Ajustar as somas sempre que um objeto entra ou sai do grupo

Tarefa

- Leitura do Capítulo 8 (seções 8.1, 8.2.1 e 8.2.2, 8.2.3, 8.2.4, 8.2.5) do livro Tan et al, 2006.
  - Está disponível em: http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf
- Leitura do Capítulo 3 (seções 3.1 e 3.3.1 e 3.3.2) do livro do Jain e Dubes, 1999.

Referências

- Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006.
- Jain, A. K.; Dubes, R. C. Algorithms for Clustering Data, Prentice Hall, 1988.
- Keogh, E. A g. Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.