

# Aprendizado de Máquina

## Aula 7.1 - Agrupamento de dados


Adriano Rivolli

[rivolli@utfpr.edu.br](mailto:rivolli@utfpr.edu.br)

**Especialização em Inteligência Artificial**

Universidade Tecnológica Federal do Paraná (UTFPR)  
Câmpus Cornélio Procópio  
Departamento de Computação

# Conteúdo

- 
- 1** Aprendizado não supervisionado
  - 2** Agrupamento de dados
  - 3** Medidas de distância

## Aprendizado não supervisionado

# Introdução

- Aprendizado não supervisionado: as instâncias não são rotuladas
- Usado para descobrir propriedades dos dados
- Principais tarefas:
  - ▶ Agrupamento de dados
  - ▶ Redução de dimensionalidade
  - ▶ Detecção de *outliers* e novidades
  - ▶ Regras de associação
  - ▶ Descoberta de subgrupos (*subgroup discovery*)

# Redução de dimensionalidade

- Conhecido como transformação de dados
- Consiste em transformar o espaço dos dados
  - ▶ Reduzir a dimensionalidade
  - ▶ Sem 'perder' a informação
- Exemplos:
  - ▶ *Principal Component Analysis* (PCA)
  - ▶ *Singular Value Decomposition* (SVD)
  - ▶ Autoencoders (redes neurais)

## Detecção de *outliers*

- Um *outlier* consiste em um ponto que se difere dos demais
  - ▶ anormalidades, discordantes, desviantes ou anomalias
- Aplicações:
  - ▶ Limpeza de dados
  - ▶ Fraudes em sistemas
  - ▶ Detecção de intrusão/invasão
- Exemplos:
  - ▶ Valores extremos
  - ▶ Agrupamentos: distância e densidade
  - ▶ Modelos probabilísticos e baseados em teoria da informação

# Regras de associação

- Descoberta de padrões frequentes em dados transacionais
- Aplicações:
  - ▶ Compras de produtos
  - ▶ Mineração de texto
  - ▶ Análise de logs
- Algoritmos:
  - ▶ Apriori
  - ▶ Enumeration-Tree

## Descoberta de subgrupos

- Encontra associações entre diferentes variáveis em relação a uma propriedade específica
- Encontra elementos que consistentemente se diferem da população em decorrência de alguma propriedade
- Exemplo:
  - ▶ Os alunos que reprovaram na disciplina X moram no estado Y e tiraram nota menor do que 7 na disciplina Z

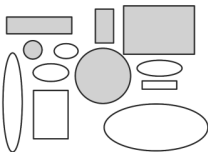


## Agrupamento de dados

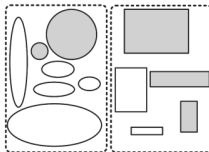
## Visão geral

- Agrupar pontos similares
  - ▶ Proximidade (distância)
  - ▶ Relação espacial
- Tipos de agrupamento
  - ▶ Bem separados
  - ▶ Baseados em centroides
  - ▶ Conectados/ligados
  - ▶ Densidade
  - ▶ Similaridade

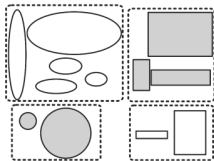
# Critérios de agrupamento



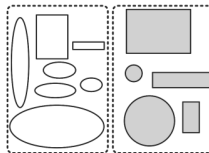
(a) Objetos



(b) Agrupamento pela forma (2 clusters)



(d) Agrupamento pelo preenchimento e pela forma (4 clusters)



(c) Agrupamento pelo preenchimento (2 clusters)

**Fonte:** Faceli K., Lorena A. C., Gama J., Carvalho, A. C. P. L., 2011. Inteligência artificial: uma abordagem de aprendizado de máquina. LTC, 2a Edição.

## Critérios

### ■ Compactação

- ▶ A homogeneidade de um grupo está relacionada a pequenas variações intra-grupo
- ▶ Adequado para formas esféricas e grupos disjuntos
- ▶ Não é adequado para estruturas complexas

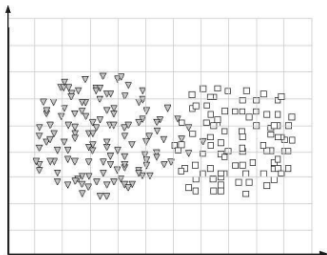
### ■ Conexão/Ligação

- ▶ Cada elemento está conectado aos vizinhos mais próximos nos mesmos clusters
- ▶ Funciona com qualquer formato dos dados
- ▶ Não adequado quando os grupos estão próximos

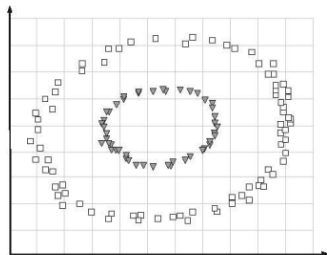
### ■ Separação espacial

- ▶ Não há sobreposição entre os grupos
- ▶ Geralmente associado com outro critério

## Compactação x Ligação



(a) Globular



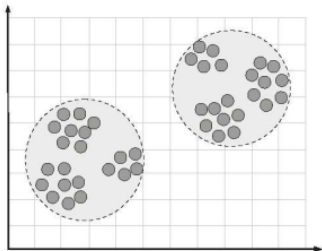
(b) Anel

**Fonte:** Faceli K., Lorena A. C., Gama J., Carvalho, A. C. P. L., 2011. Inteligência artificial: uma abordagem de aprendizado de máquina. LTC, 2a Edição.

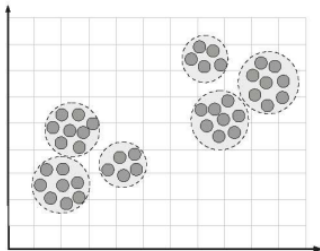
## Questões importantes

- Níveis de refinamento (granularidade dos grupos)
- Estruturas heterogênea
- Validação dos grupos
- Interpretação dos resultados

## Granularidade dos grupos



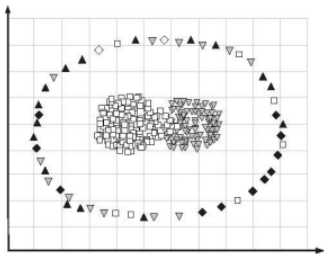
(a) 2 clusters



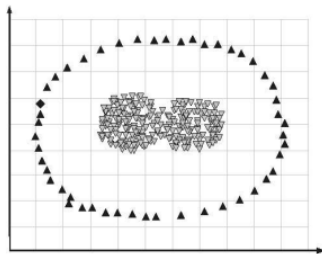
(b) 6 clusters

**Fonte:** Faceli K., Lorena A. C., Gama J., Carvalho, A. C. P. L., 2011. Inteligência artificial: uma abordagem de aprendizado de máquina. LTC, 2a Edição.

## Granularidade dos grupos



(a)  $k$ -médias

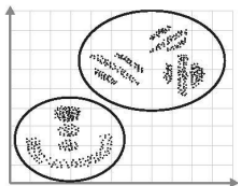


(b) Hierárquico com ligação simples

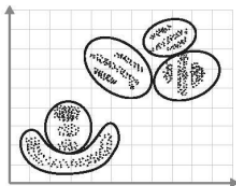
**Fonte:** Faceli K., Lorena A. C., Gama J., Carvalho, A. C. P. L., 2011. Inteligência artificial: uma abordagem de aprendizado de máquina. LTC, 2a Edição.



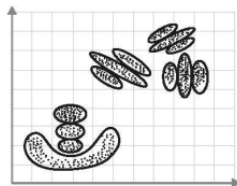
# Estruturas complexas



(a) Estrutura E1



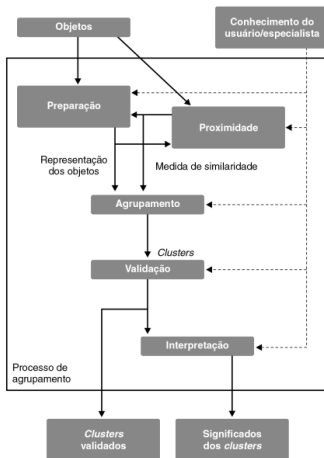
(b) Estrutura E2



(c) Estrutura E3

**Fonte:** Faceli K., Lorena A. C., Gama J., Carvalho, A. C. P. L., 2011. Inteligência artificial: uma abordagem de aprendizado de máquina. LTC, 2a Edição.

# Etapas do processo de agrupamento



## Comparação entre algoritmos

- Complexidade
- Escalabilidade
- Métricas de similaridade
- Robustez a ruído e *outliers*
- Suporte a alta dimensionalidade
- Estabilidade
- Agrupamento incremental

## Outras comparações

### ■ Resultado

- ▶ Formato dos grupos
- ▶ Interpretabilidade

### ■ Dados


- ▶ Tipo de dado suportado
- ▶ Ordem dos dados (sequência)

### ■ Hiperparâmetros

- ▶ Número de grupos
- ▶ Outros (específicos)

## Medidas de distância

## Similaridade e distâncias

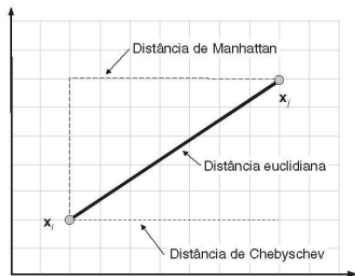
- 
- Há diferentes maneiras de modelar a similaridade entre pontos
    - ▶ Quando  $a = b$  a similaridade é máxima (ex: 1)
    - ▶ Quando  $a$  é muito diferente de  $b$  a similaridade deve ser próximo de 0
  - A similaridade pode ser modelada em função da distância

# Métricas

- Manhattan / *City Block* (L1)
- Euclideana (L2)
- Chebyshev (*supremum*)
  - ▶ Máxima diferença entre todos os atributos
- Similaridade de Cosine
- Similaridade de Jaccard
- Correlação Pearson (similarity)

# Minkowski

$$d(a, b) = \left( \sum_{i=1}^d |a_i - b_i|^p \right)^{\frac{1}{p}}$$





# Propriedades métricas

- $d(x_i, x_i) = 0$ , para todo  $x_i$
- $d(x_i, x_j) = d(x_j, x_i)$  (simetria)
- $d(x_i, x_j) \geq 0$ , para todo  $i$  e  $j$  (positividade)
- $d(x_i, x_j) = 0 \iff x_i = x_j$
- $d(x_i, x_j) \leq d(x_i, x_k) + d(x_k, x_j)$  (desigualdade triangular)