Aula 2 – Tópicos Especiais em Computação: Agrupamento de Dados Introdução

Profa. Elaine Faria

UFU - 2020

Aprendizado de Máquina

Será que um computador pode aprender?

Aprendizado de Máquina

- O que computadores capazes de aprender poderiam fazer?

Aprendizado de Máquina

É uma área de IA cujo objetivo é o desenvolvimento de técnicas computacionais sobre o aprendizado bem como a construção de sistemas capazes de adquirir conhecimento de forma automática.

Monard, M. C e Baranauskas J. A. Sistemas Inteligentes, Ed. Manole, 2003.

Aprendizado de Máquina

Qual o cenário atual do aprendizado de máquina?

Mineração de Dados

Processo de automaticamente descobrir informação útil em grandes repositórios de dados

Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

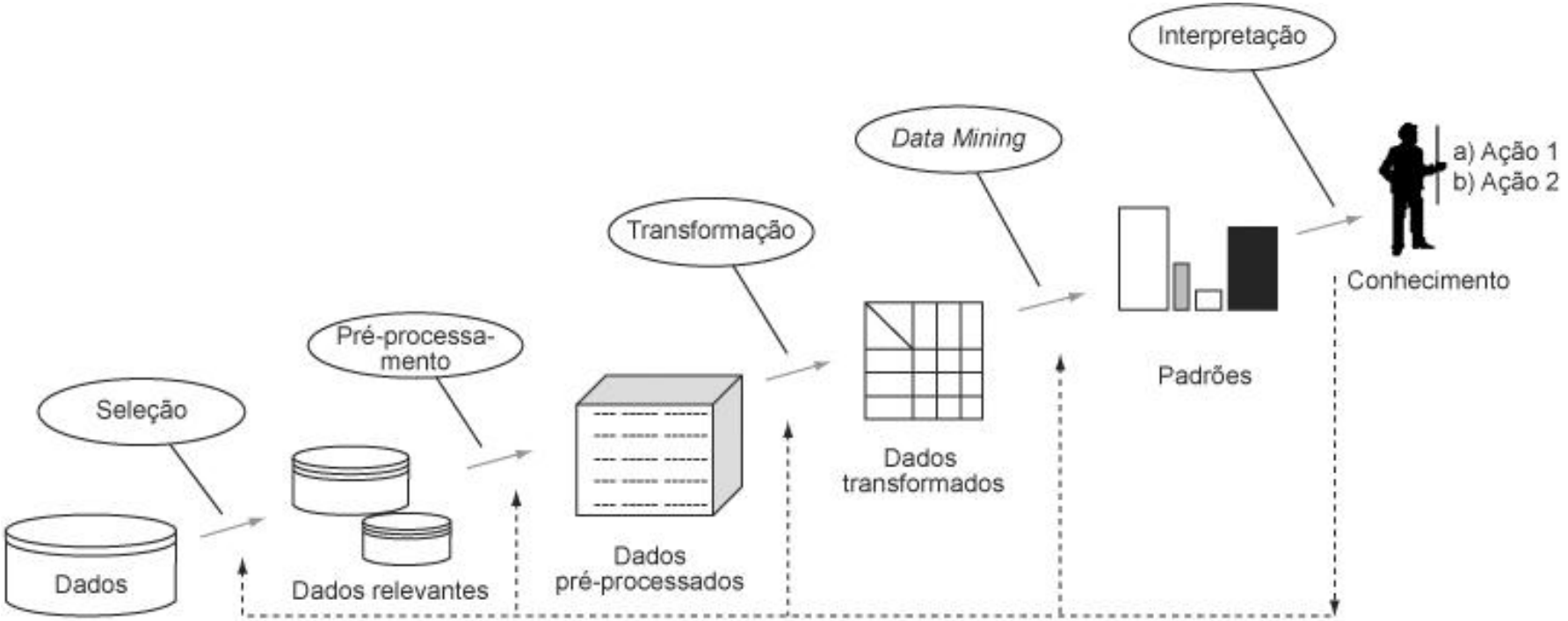
Mineração de Dados

- Motivação
  - Grandes quantidades de dados geradas
  - Dados de alta dimensionalidade
  - Dados heterogêneos e complexos
  - Dados distribuídos
  - Análise não tradicional

Mineração de Dados

- Exemplos de aplicação
  - Negócios
    - Padrões de compra/ligações → Marketing
    - *Logs* da Web → Melhor design de sites
    - Detecção de fraude
    - Agrupamento de clientes
  - Bioinformática
    - Genes importantes
    - Sintomas associados a doenças

Descoberta do conhecimento em bases de dados - KDD



Etapas do processo *KDD*

Fayyad U. et al. Knowledge Discovery and Data Mining: Towards a Unfiying Framework, KDD 96, 1996

Mineração de Dados e Aprendizado de Máquina

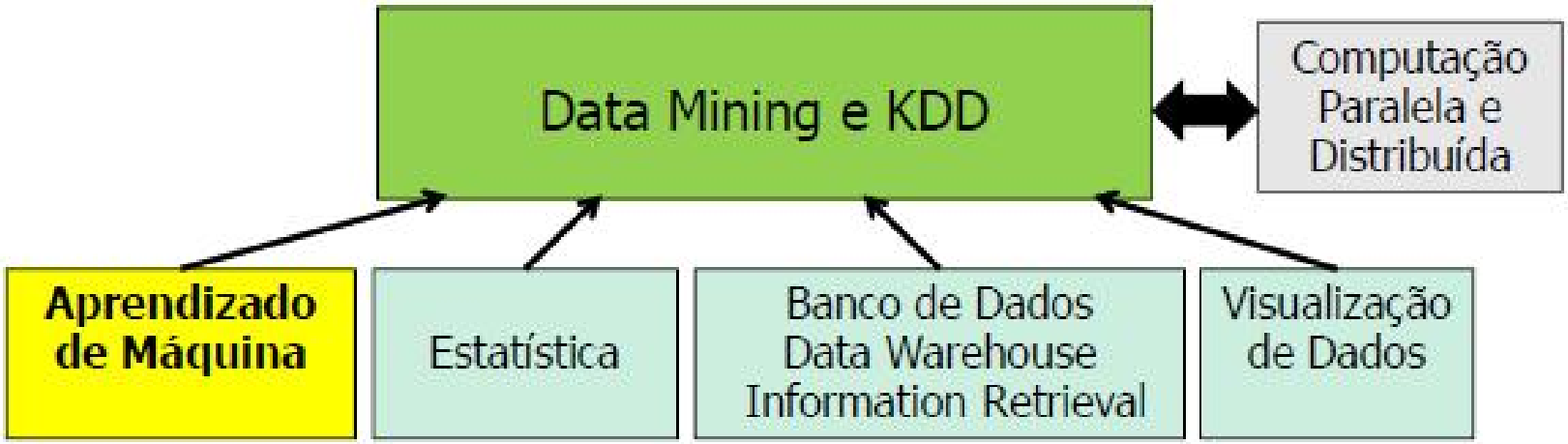
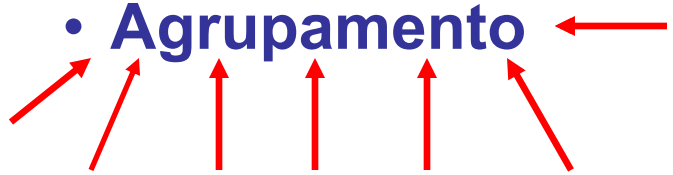


Figura retirada dos slides do Prof. Ricardo Campello – ICMC-USP

Mineração de Dados

- Tarefas de mineração
  - Tarefas Preditivas
    - Classificação
    - Regressão
  - Tarefas Descritivas
    - **Agrupamento**



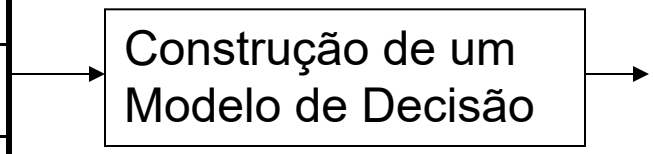
Qual a diferença entre agrupamento e classificação

- Agrupamento
  - Aprendizado não-supervisionado → sem rótulo
- Classificação
  - Aprendizado supervisionado → usa os rótulos

Exemplo de Classificação

Nome	Idade	Renda	Pagador
João	<30	Média	Bom
Ana	41..50	Alta	Bom
Pedro	41..50	Alta	Bom
Maria	41..50	Baixa	Ruim
Paulo	<30	Baixa	Ruim
Aldo	>60	Alta	Ruim

Base de Dados
Treinamento



Algoritmo

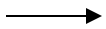
Se idade = 41..50 e
Renda = Alta então
Pagador = Bom

Se renda = baixa
então
Pagador = Ruim

Modelo

Exemplo de Classificação

Nome	Idade	Renda	Pagador
Ivo	31..40	Baixa	**????**



Se idade = 41..50 e
Renda = Alta então
Pagador = Bom

Se renda = baixa
então
Pagador = Ruim

Ruim

Novo Dado

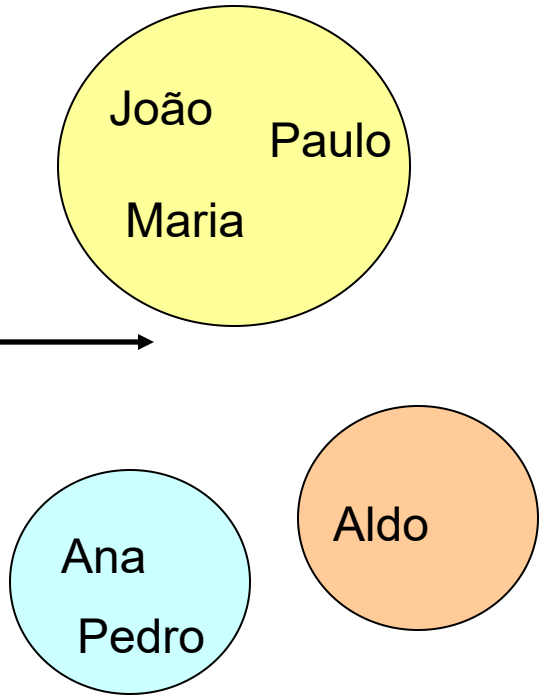
Classificador

Classificação

Exemplo de Agrupamento

Nome	Idade	Peso
João	25	60
Ana	50	75
Pedro	60	90
Maria	22	65
Paulo	18	68
Aldo	15	80

Aplicação de uma técnica agrupamento



Agrupamento - definições

Análise de grupos ou clusters é o estudo de algoritmos e métodos para agrupar objetos de acordo com suas características.

Cluster (grupo) é uma aglomeração de pontos no espaço tal que a distância entre quaisquer dois pontos no cluster é menor que a distância entre qualquer ponto no cluster e qualquer ponto que não está nele.

Jain, A. K.; Dubes, R. C. Algorithms for Clustering Data, Prentice Hall, 1988.

Agrupamento - questões

- O que é um grupo ideal?
- Quantos grupos devem ser formados?
- Há um agrupamento natural dos dados?
- Como podemos definir o que é semelhante?

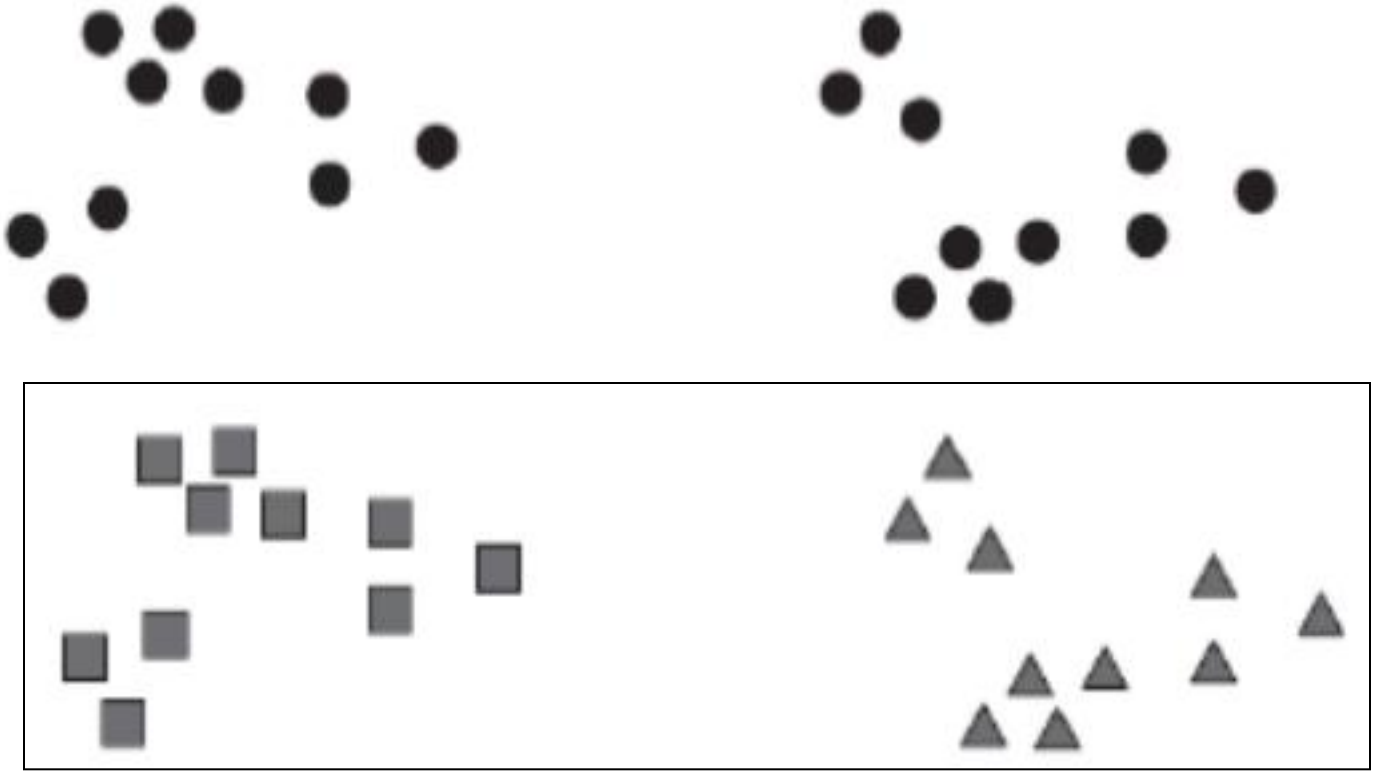
Agrupamento

Quantos grupos?

Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Agrupamento

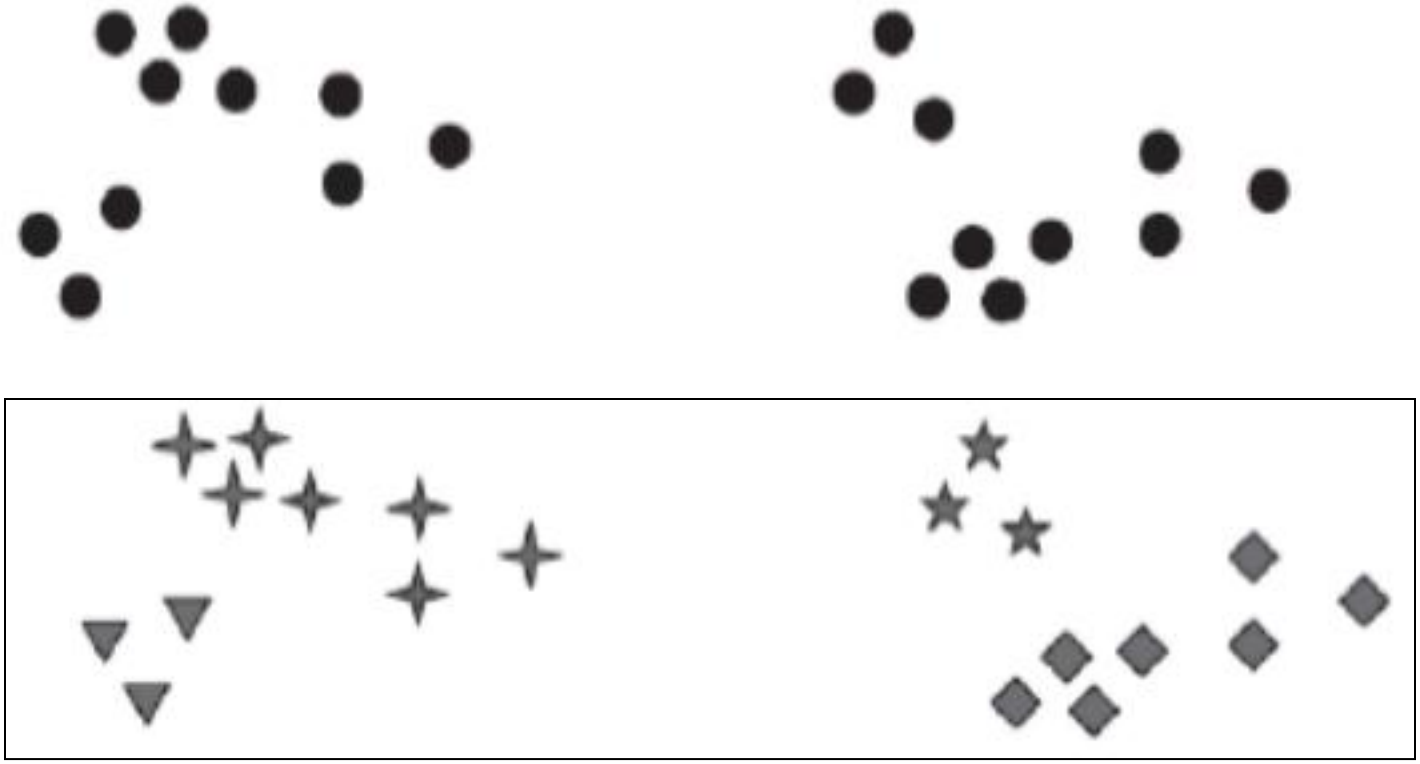
Quantos grupos?



Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Agrupamento

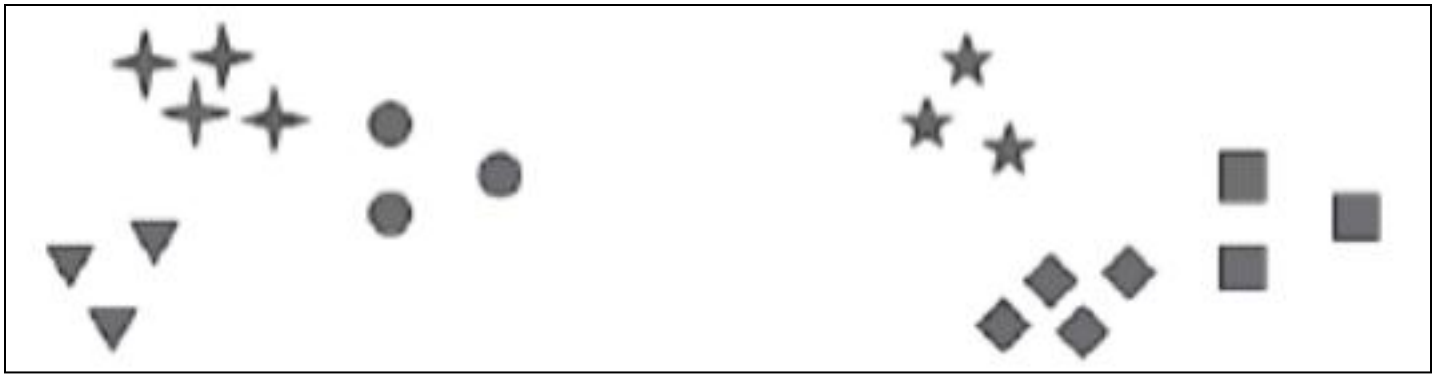
Quantos grupos?



Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Agrupamento

Quantos grupos?

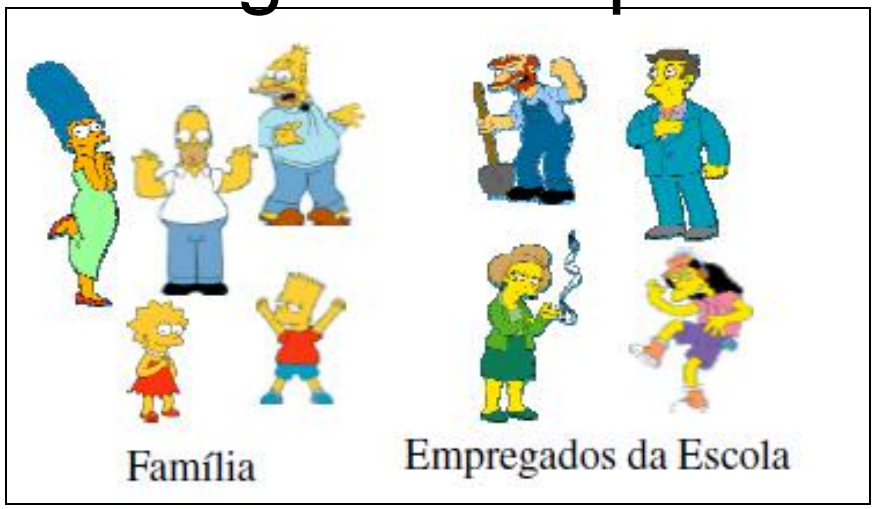


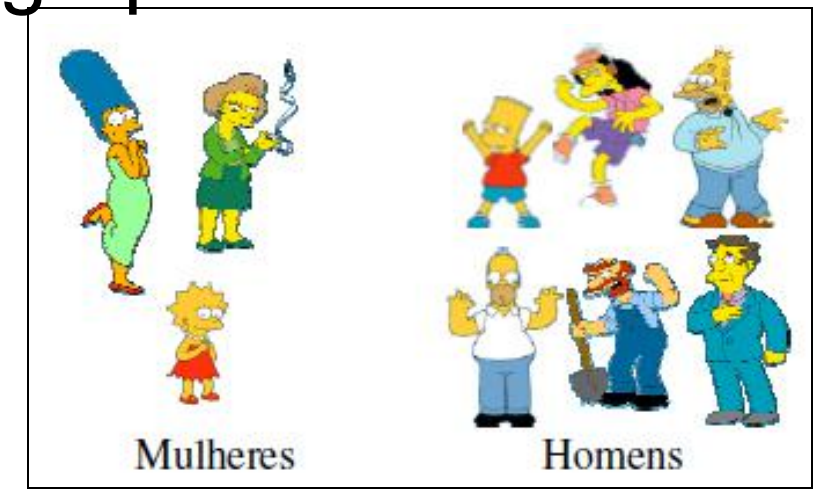
Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Agrupamento

- Como agrupar os objetos??

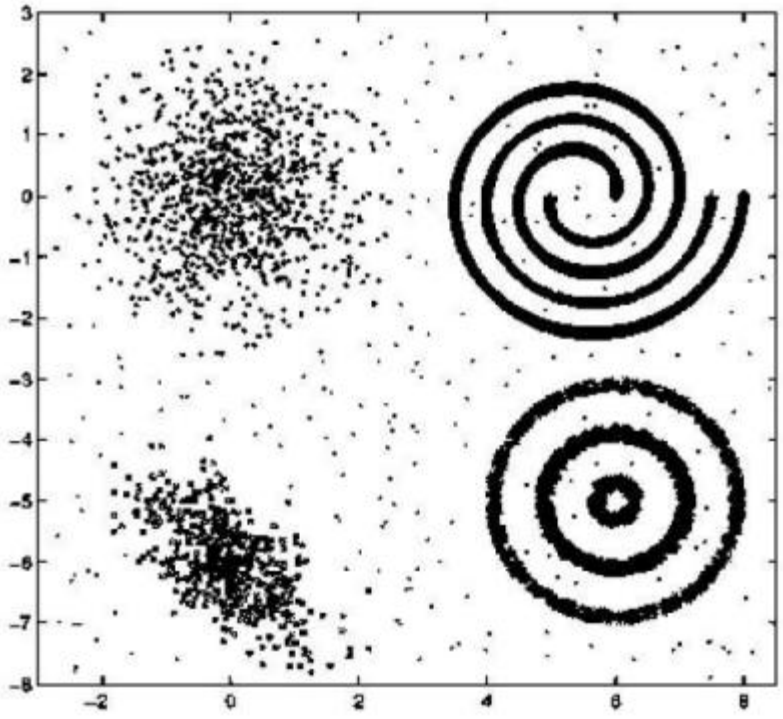
- Alguns dos possíveis agrupamento



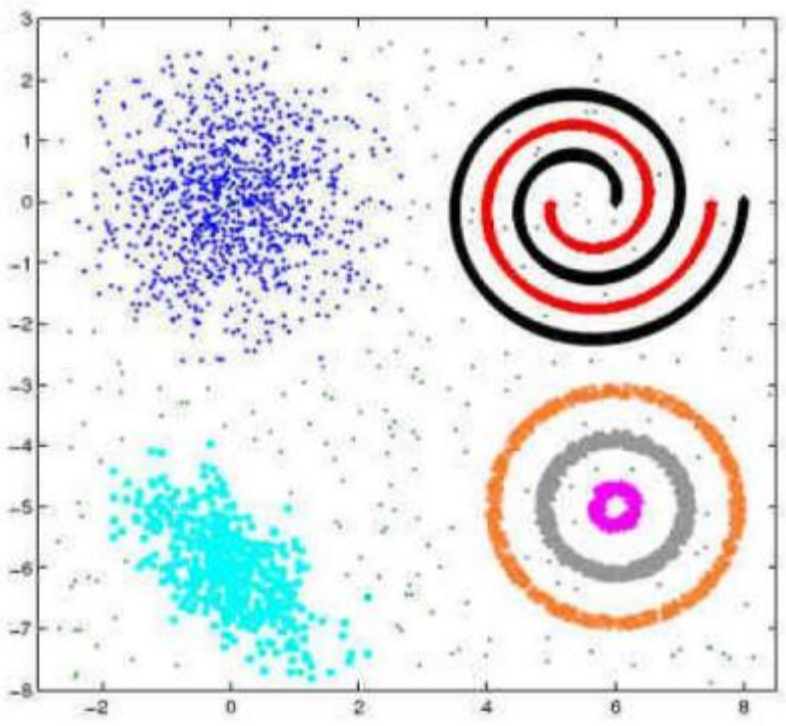


Keogh, E. A g. Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

Agrupamento - questões



Dados de entrada



Agrupamento desejado

JAIN, A. K. 2008. Data Clustering: 50 Years Beyond K-Means, Pattern Recognition Letters, vol. 31, n 8, 2010.

Agrupamento

- Semelhança entre objetos

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

Agrupamento

- Quais as vantagens das técnicas de agrupamento em relação ao agrupamento manual?
  - Um programa pode aplicar um critério objetivo para forma os grupos, de forma consistente
  - Um programa consegue fazer o agrupamento em n-dimensões
  - Um programa consegue fazer o agrupamento em pouco tempo

Agrupamento

Após realizar o agrupamento o problema acabou?

Não, agora é preciso interpretar os resultados obtidos!!!!

Analisar e avaliar o agrupamento!

O que vamos estudar nessa disciplina

- Objetivo
  - Conhecer algumas das técnicas de agrupamento de dados: particionais e hierárquicas
    - Implementar algumas dessas técnicas
  - Usar ferramentas disponíveis na internet que já possuem alguns algoritmos de agrupamento já implementados

O que vamos estudar nessa disciplina

- Mas antes disso precisamos:
  - Conhecer os dados e sua representação
  - Conhecer as principais medidas de (dis)similaridade
    - Implementar essas medidas de (dis)similaridade

O que vamos estudar nessa disciplina

- E depois disso?
  - Avaliar o agrupamento realizado usando diferentes medidas de avaliação
    - Implementar essas medidas de avaliação
  - Aplicar os conceitos aprendidos na solução de problemas do mundo real
  - Conhecer sobre tendências na área de agrupamento de dados → Data streams

Importância do agrupamento

- Será que agrupamento é mesmo importante?
  - Mais de 50 anos de estudo sobre técnicas de agrupamento
  - Milhares de aplicações
  - Dezenas de ferramentas na internet para fazer agrupamento automático
  - Milhares de artigos publicados na área

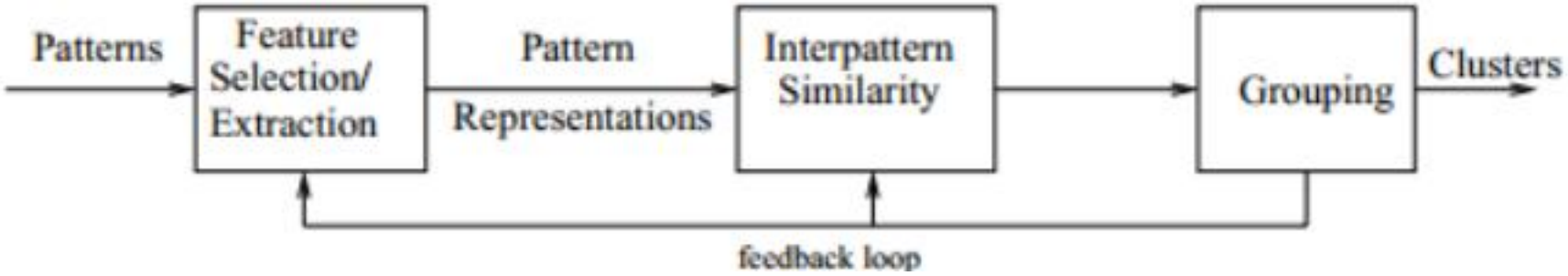
Exemplos de aplicações de agrupamento de dados

- Marketing
  - Ex: Agrupamento de clientes para direcionar campanhas de marketing
- Documentos
  - Ex: agrupar documentos que tratam do mesmo assunto
- Imagens
  - Ex: Segmentar imagens usando técnicas de agrupamento
- Biologia
  - Ex: agrupar animais de acordo com o reino, ramo, classe, ordem, gênero, ...

Processo de Agrupamento

Mas e ai, por onde eu começo?

Vamos estudar cada um dos passos do processo de agrupamento



Jain, A. K., Murty, M. N., Flynn, P. J., Data clustering: a review, ACM Computing Surveys, vol. 31, n 3, 1999.

Tarefa

- Ler o capítulo 1 do livro do Jain e Dubes

Referências

- Jain, A. K., Murty, M. N., Flynn, P. J., Data clustering: a review, ACM Computing Surveys, vol. 31, n 3, 1999.
- Jain, A. K. 2008. Data Clustering: 50 Years Beyond K-Means, Pattern Recognition Letters, vol. 31, n 8, 2010.
- Jain, A. K.; Dubes, R. C. Algorithms for Clustering Data, Prentice Hall, 1988.
- Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006.
- Keogh, E. A g. Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.
- Aggarwal, C. Data Mining: The text book, Springer, 2015.