Data Mining
Análise de Grupos II

Prof. Dr. Joaquim Assunção

DEPARTAMENTO DE COMPUTAÇÃO APLICADA
CENTRO DE TECNOLOGIA
UFSM
2019

www.inf.ufsm.br/~joaquim

Fair user agreement

Este material foi criado para a disciplina de Mineração de Dados - Centro de Tecnologia da UFSM.

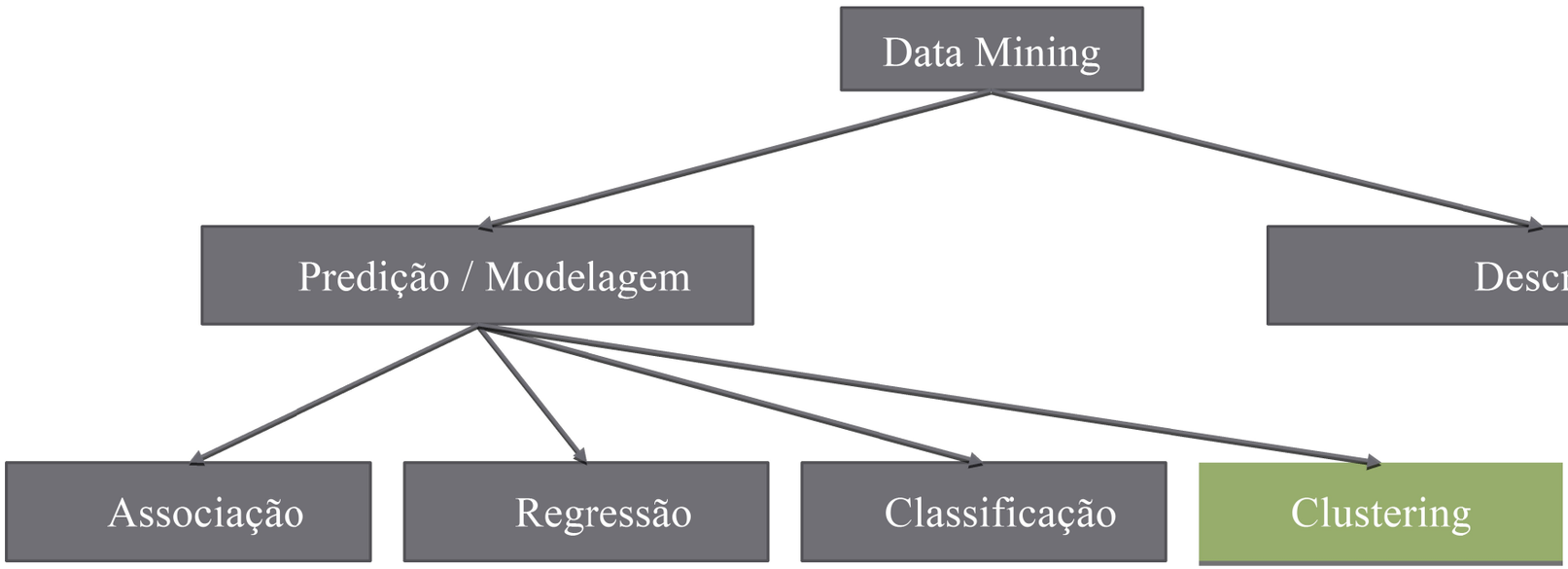
Você pode usar este material livremente*; porém, caso seja usado em outra instituição, **me envie um e-mail** avisando o nome da instituição e a disciplina.

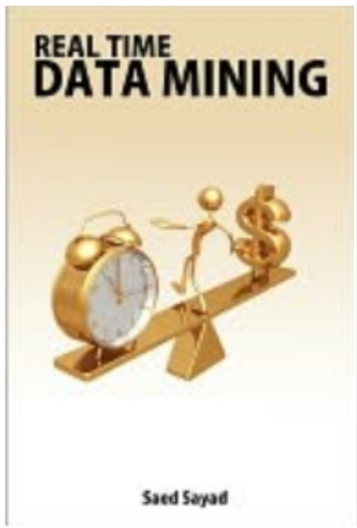
*Caso você queira usar algo desse material em alguma publicação, envie-me um e-mail com antecedência.

Prof. Dr. Joaquim Assunção.

joaquim@inf.ufsm.br

Mapa para Mineração de Dados*





***http://www.saedsayad.com/data_mining_map.htm**

Tipos de agrupamentos

- Bem separados
- Baseados em centro
- Contínuos
- Baseados em densidade
- Conceitual ou baseados em propriedades
- Descritos por uma função objetiva

Tipos de grupos

Dado um agrupamento com grupos Q e P. Onde cada grupo contém dois centróides, pontos de referência q_r e p_r.

Bem separados

$$\forall i, j \mid i \neq j \quad dist(Q_i, Q_j) < dist(Q_i, P_j)$$

Um grupo bem separado é um conjunto de pontos de modo que qualquer ponto do grupo seja mais próximo de todos os outros pontos do grupo do que de qualquer ponto que não esteja no grupo.

Tipos de grupos

Dado um agrupamento com grupos Q e P. Onde cada grupo contém dois centróides, pontos de referência q_r e p_r.

Baseados em centro

$$\forall k, i, j \mid i \neq j \quad dist(Q_i, q_r) < dist(Q_i, p_r^k)$$

Um grupo bem separado é um conjunto de pontos de modo que qualquer ponto do grupo seja mais próximo do centro de seu grupo do que do centro de qualquer outro grupo.

Tipos de grupos

Dado um agrupamento com grupos Q e P. Onde cada grupo contém dois centróides, pontos de referência q_r e p_r.

Contínuos

$$\forall i, j, w \in Q \mid i \neq j \neq w \quad \exists w | dist(Q_i, Q_w) < dist(Q_i, Q_j)$$

Um grupo contínuo é um grupo em que para cada ponto do grupo existe outro ponto cuja distância é menor do que a distância deste ponto para todos os outros pontos do outro grupo.

Tipos de grupos

Dado um agrupamento com grupos Q e P. Onde cada grupo contém dois centróides, pontos de referência q_r e p_r.

Baseados em densidade

Há casos em que um determinado grupo é formado pela densidade de objetos no local. Casos assim, tendem a separar pontos próximos a um centroide devido ao vaco de objetos entre este ponto e este centroide.

Tipos de grupos

Dado um agrupamento com grupos Q e P. Onde cada grupo contém dois centróides, pontos de referência q_r e p_r.

Propriedade ou Conceitual

Grupos que compartilham de uma mesma propriedade ou representam um conceito. Em poucas palavras, a proximidade de algumas variáveis podem pesar mais que outras, assim o grupo estaria sendo formado por propriedades específicas.

Tipos de grupos

Dado um agrupamento com grupos Q e P. Onde cada grupo contém dois centróides, pontos de referência q_r e p_r.

Descritos por função

Exemplos similares ao anterior podem ser aplicados a esta categoria. Imagine, objetos que seguem uma determinada fórmula podem ser agrupados em um mesmo grupo. Por exemplo, um cluster azul pode ser formado por $y = 2x$ e o vermelho por $y = -2x$

Coesão dos Grupos

A medida mais comum para definir o quão bom é um cluster, é a Soma do erro quadrático (Sum of squared error (**SSE**)). Para cada ponto, o “erro” é a distância euclidiana até o centroide mais próximo. Então, quanto menor o SSE, mais próximos o conjunto de pontos está do centroide. Logo, melhor o centroide para definir esse conjunto de pontos

Coesão dos Grupos

Seja x um dado no cluster C_i ; e c_i , o ponto representativo para o cluster C_i .

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} dist^2(c_i, x)$$

Coesão dos Grupos

De maneira simplificada, se usarmos a distância euclidiana, podemos ver que o centroide que minimiza o SSE de um grupo é a média. Assim, para um determinado grupo C_i com n objetos, o centroide é dado por:

$$C_i = \frac{1}{n} \sum_{x \in C_i} x$$

Exercícios

1) Pode um tipo de cluster ter a característica de um modelo de regressão? Qual(is)? Explique sua resposta.

2) Procure na web algum conjunto de dados em que a análise de grupamento seja útil. Formate esse conjunto de dados e rode o algoritmo k-means. Crie um gráfico para visualizar os grupamentos gerados.