Data Mining
Transformação/Integração - III

Prof. Dr. Joaquim Assunção

DEPARTAMENTO DE COMPUTAÇÃO APLICADA
CENTRO DE TECNOLOGIA
UFSM
2022

www.inf.ufsm.br/~joaquim

Fair user agreement

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

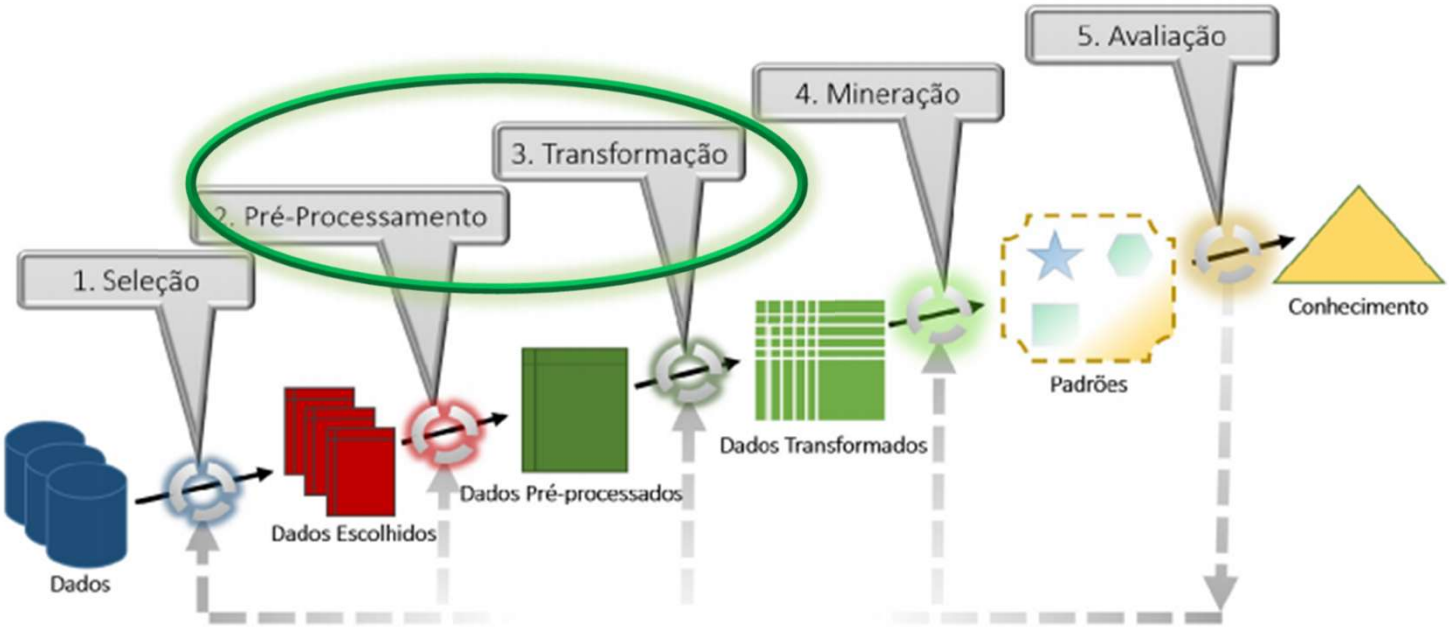
Você pode usar este material livremente*; porém, caso seja usado em outra instituição, **me envie um e-mail** avisando o nome da instituição e a disciplina.

*A maior parte deste material foi retirado do livro: "Joaquim V. C. Assunção. **Uma Breve Introdução à Mineração de Dados: Bases Para a Ciência de Dados, com Exemplos em R**. 192 páginas. Novatec. 2021. ISBN-10 : 6586057507."

Prof. Dr. Joaquim Assunção.

joaquim@inf.ufsm.br

Transformação



Problemas Comum

- Muitos algoritmos não tratam bem com diferentes escalas.
  - Quando falamos de algoritmos de aprendizagem para mineração, escalas as vezes são necessárias; outras, impactam em desempenho e, raramente, não fazem diferença.

Normalização

Abordagens

- Há diferentes tipos de normalização. Dentre as mais usadas, temos...
- Mínimo e Máximo
- *Z-score*
- Escalonamento decimal
- Abrangência do quartil

Abordagens

1. **Min-Max**

Transformação linear que mapeia valores ***X*** em valores ***X'***.

Os novos valores de *min* e *max* são escolhidos pelo usuário, mas a combinação mais frequente é 0 e 1.

$$x' = \frac{x - min_x}{max_x - min_x}(novoMax_x - novoMin_x) + novoMin_x$$

Abordagens

1. **Min-Max**

Transformação linear que mapeia valores ***X*** em valores ***X'***.

Os novos valores de *min* e *max* são escolhidos pelo usuário, mas a combinação mais frequente é 0 e 1.

$$x' = \frac{x - min}{max_x - min_x}(novoMax_x - novoMin_x) + novoMin_x$$

Abordagens

2. ***Z-Score***

A normalização Z-Score se baseia na média e no desvio padrão do conjunto para criar os novos valores.

$$x' = (x - \bar{x})/std_x$$

Abordagens

3. **Escalonamento decimal**

Na normalização por escalonamento decimal move a casa dos atributos originais com base no valor máximo de x, onde *j* é o menor inteiro de modo que *max(x')* seja menor do que 1.

$$x' = x/10^j$$

Abordagens

4. **Abrangência interquartil**

Na normalização por abrangência dos quartis, os três quartis são usados para normalizar os dados. Q_2 é a mediana.

$$x' = (x - Q_2)/IQR$$
$$|IQR = Q_3 - Q_1$$

Hands on!

1. Leia o arquivo '*vendas_lucro.csv*'. Salve em um data frame chamado `DF_vendasLucro`. Crie mais 3 data frames, onde cada um deve ser os valores de `DF_vendasLucro` normalizados por uma técnica de normalização (escolha 3). Finalmente, crie um data frame para agregar os anteriores (composto pelos 4 data frames antigos lado a lado).