Governo Federal

Mineração de Dados Aplicada à Engenharia

Profº - Dr. Thales Levi Azevedo Valente

thales.l.a.valente@gmail.com.br

Mineração de Dados

Pré-Processamento de Dados

Sumário

- Dados
- Escalas
- Cardinalidade
- Porque pré-processar dados ?
  - Limpeza de Dados
  - Integração e Transformação
  - Redução de Dados
  - Discretização

Dados

- Medidas
  - O que é possível medir sobre as características: meu carro é azul escuro, 2 portas, 6 cilindros, 5 passageiros
- Variáveis, descritores
  - Uma variável representa uma medida que toma um numero particular de valores, com a possibilidade de valores diferentes para cada observação.

Classificação dos Dados

- Qualitativas: observações não numéricas
  - Nominais: dados sem ordenamento nem hierarquia
    - Ex: estado civil; nome de cidades; etc
  - Ordinais: equivalente aos nominais, porém com hierarquia
    - Ex: cargos; resposta questionário com escala: ótimo, bom, regular, ruim; posição das maiores empresas em volume de vendas; etc.

Classificação dos Dados

- Quantitativas: quantidades medidas numa escala numérica
  - Discretos: valores numéricos inteiros positivos:
    - Ex: Número de filhos, quantidade de peças, número de clientes,etc
  - Contínuos: valores númericos do conjunto dos número reais
    - Ex: Estaturas, valor de vendas mensais

Atributos das variáveis

- Variável qualitativa:
  - nominal
  - ordinal
- Variável quantitativa:
  - intervalar
  - Proporcional ou razão

Escalas

- Escala Nominal
  - Nessa escala os valores são não numéricos e são não ordenados.
  - Duas instâncias apresentam ou não o mesmo valor.
  - Ex: Cor, Modelos de Carro, etc
- Escala Ordinal
  - Nessa escala os valores são não numéricos e ordenados.
  - Uma instância pode apresentar um valor comparativamente maior do que uma outra.
  - Ex: Grau de Instrução

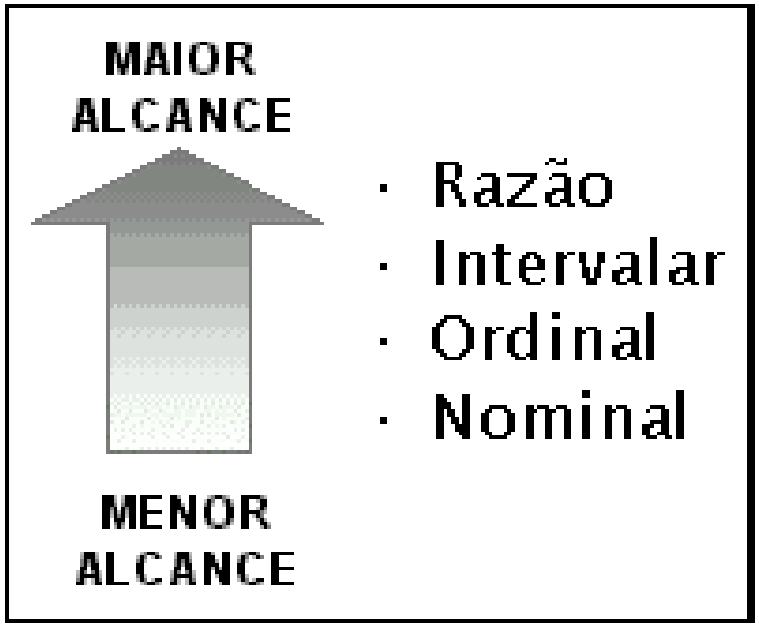
Escalas

- Escala Intervalar
  - Nessa escala (particular) de valores numéricos, não existe apenas uma ordem entre os valores, mas também existe diferença entre esses valores. Não há ponto de nulidade. O zero é relativo
  - Ex: Temperatura em Graus Celsius

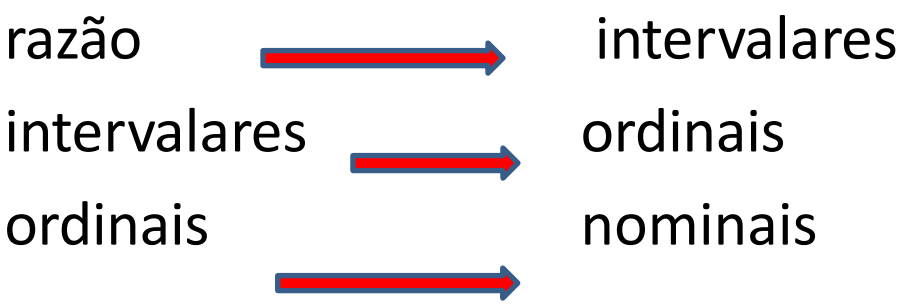
- Escala de Razão ou Proporcional
  - Nessa escala de valores numéricos, além da diferença, tem sentido calcular a proporção entre valores (o zero é absoluto).
  - Ex: idade, salário, preço, volume de vendas, distâncias, etc

Hierarquia Entre as Escalas

- As escalas numéricas apresentam entre si uma clara hierarquia no que concerne à sua sofisticação e à sua capacidade de representar os nuances do que é observado.



Transição de Escalas



- Naturalmente, tais transições envolvem, necessariamente, uma perda de informação.
- Apenas em certas situações muito especiais,é possível se fazer a trajetória no sentido inverso

Cardinalidade: Discreto versus Contínuo

- Variáveis Discretas
  - Qualquer variável que possui um conjunto finito de valores distintos.
  - Ex: Departamentos da Faculdade de Engenharia
- Variáveis Contínuas
  - Podem, em principio, assumir qualquer valor dentro de um intervalo
  - Exemplo: peso, altura, lucros

Cardinalidade: Discreto versus Contínuo

- Variáveis dicotômicas
  - Ex: Sexo (M, F)
  - Ex: Possui celular? (Sim , não)
- Variáveis binárias
  - Em geral são codificadas como “0”, “1”
  - “0” em geral indica ausência de propriedade

Sumário

- Dados
- Escalar
- Cardinalidade
- Porque pré-processar dados ?
  - Limpeza de Dados
  - Integração e Transformação
  - Redução de Dados
  - Discretização

Porque pré-processar dados ?

- Dados no mundo real são “sujos”
  - Incompletos: valores faltantes, atributos faltantes, etc
    - Exemplo: ocupação = “”
  - Ruidosos: contendo erros ou “outliers”
    - Exemplo: salário = “-10”
  - Inconsistentes: contendo discrepâncias
    - Exemplo: Idade=“42”, Data de Nascimento=“03/07/1997”

Porque os dados são “sujos” ?

- Dados incompletos têm origem em:
  - Indisponibilidade durante a coleta
  - Considerações diferentes quando o dado foi coletado e quando foi analisado
  - Falha humana/software/hardware
- Dados ruidosos têm origem no processo de:
  - Coleta
  - Entrada
  - Transmissão
- Dados inconsistentes têm origem em:
  - Fontes diferentes
  - Violação de dependências funcionais

Porque Pré-Processar os Dados é Importante ?

- Técnicas de pré-processamento e transformação de dados são aplicadas para aumentar a qualidade e o poder de expressão dos dados a serem minerados.
- Dados sem Qualidade = Mineração sem Qualidade
  - Decisões de qualidade precisam ser tomadas sobre dados com qualidade
    - Exemplo: dados duplicados ou faltantes podem gerar cálculos estatísticos incorretos
- Esta fase tende a consumir a maior parte do tempo dedicado ao processo de KDD (aproximadamente 70%).

O que define dados de qualidade ?

- Acurácia
- Completos
- Consistentes
- Temporalmente corretos
- Confiáveis
- Agregam valor
- Interpretabilidade
- Acessibilidade

Principais Tarefas de Pré-Processamento

- Limpeza dos Dados
  - Identifica ou remove “outliers” e resolve inconsistências
  - Preenche valores faltantes, suaviza dados ruidosos,
- Integração
  - Dados de origens diferentes devem ser integrados
- Transformação
  - Normalização e agregação
- Redução
  - Tenta reduzir o volume com pouca alteração no resultado final
- Discretização
  - Faz parte do processo de redução, mas tem papel importante, especialmente com dados numéricos

Limpeza de Dados

- Tratamento realizado sobre os dados selecionados de forma a assegurar a qualidade (completude, veracidade e integridade) dos fatos por eles representados.
- Quanto pior for a qualidade dos dados informados ao processo KDD, pior será a qualidade dos modelos e conhecimento gerados.
- A limpeza dos dados objetiva melhorar a qualidade dos mesmos.
- A participação do especialista do domínio, nesta fase, é essencial.

Limpeza de Dados

- Tarefas
  - Identificar “outliers” e suavizar ruídos
  - Preencher valores faltantes
  - Correções de informações errôneas ou inconsistentes.
  - Resolver redundância causada por integração dos dados
    - Ex: Definição de um intervalo para um determinado atributo. Medidas de correção para registros com ocorrência fora do intervalo para o atributo.
    - Ex: Padronização de unidades.

Dados Faltantes

- São os atributos que não possuem valor ou quando o valor dos mesmos está incompleto ou não detalhado.
  - Exemplo: muitos registros podem estar com valores faltantes (renda do cliente em dados de venda)
- Causas:
  - Mal funcionamento do equipamento
  - Remoção por inconsistência
  - Dado não inserido propositalmente
    - Falta de compreensão
    - Falta de importância
- Dados faltantes podem ter que ser inferidos

Como manipular dados faltantes ?

Métodos para tratar os dados faltantes:

- Preenchimento Manual de Valores: com base em pesquisas nas fontes originais dos dados
- Preenchimento com Medidas Estatísticas (*ReplaceMissingValues - Weka)*
  - Usar a média do atributo ou a média relativa do atributo em todos os registros que estiverem na mesma situação;
  - Usar o valor mais provável (moda)
- Usar um valor constante global:
  - substituir todos os valores ausentes de um atributo por um valor padrão (“desconhecido” ou “null”), especificado pelo especialista de domínio.

Como manipular dados faltantes ?

Métodos para tratar os dados faltantes:

- Eliminar a observação:
  - excluir do conjunto de dados as tuplas que possuam pelo menos um atributo não preenchido
- Preenchimento com Métodos de Mineração de Dados:
  - Utilizar modelos preditivos para sugerir os valores mais prováveis a serem utilizados no preenchimento dos valores ausentes.
  - Melhor estratégia (as demais podem ser tendenciosas demais)
- Todos os métodos apresentam vantagens e desvantagens
- A natureza do atributo, a quantidade de registros e o número de faltantes serão determinantes para a escolha do método mais adequado.

Dados Ruidosos

- Ruído: erro aleatório ou variância em medições
- Causas
  - Falha nos instrumentos de coleta
  - Problemas de entrada de dados ou de transmissão
  - Limitação da tecnologia
  - Inconsistência nas convenções de nomes
- Outros tipos de problemas que requerem limpeza de dados
  - Registros duplicados
  - Dados incompletos
- Uma inconsistência pode envolver um único registsro ou um conjunto de registros. Demanda conhecimento especialista.
- □ Um cliente com idade inferior a 21 possui crédito aprovado.

Dados Ruidosos:

Métodos para tratar os dados ruidosos:

- Exclusão de Casos
  - Excluir do conjunto de dados original os registros que possuem pelo menos uma inconsistência.
  - SQL pode ser utilizada para encontrar tais registros (regras de negócio).
- Correção de Erros (*RemoveMissclassified -Weka)*
  - Substituir os valores errôneos / corrigir as inconsistências.

Como manipular dados ruidosos ?

- Compartimentalização (binning)
  - Ordena os dados e particiona em “compartimentos” do mesmo tamanho
  - Feito isto, pode-se suavizar os dados pela média, mediana, pelas fronteiras da partição, etc.
- Clusterização
  - Detecta e remove “outliers”
- Inspeção humana + computadorizada
  - Detecta valores suspeitos que são checados por um ser humano (por exemplo, detecção de outliers)
- Regressão
  - Suavização através do ajuste de dados a uma função de regressão

Compartimentalização

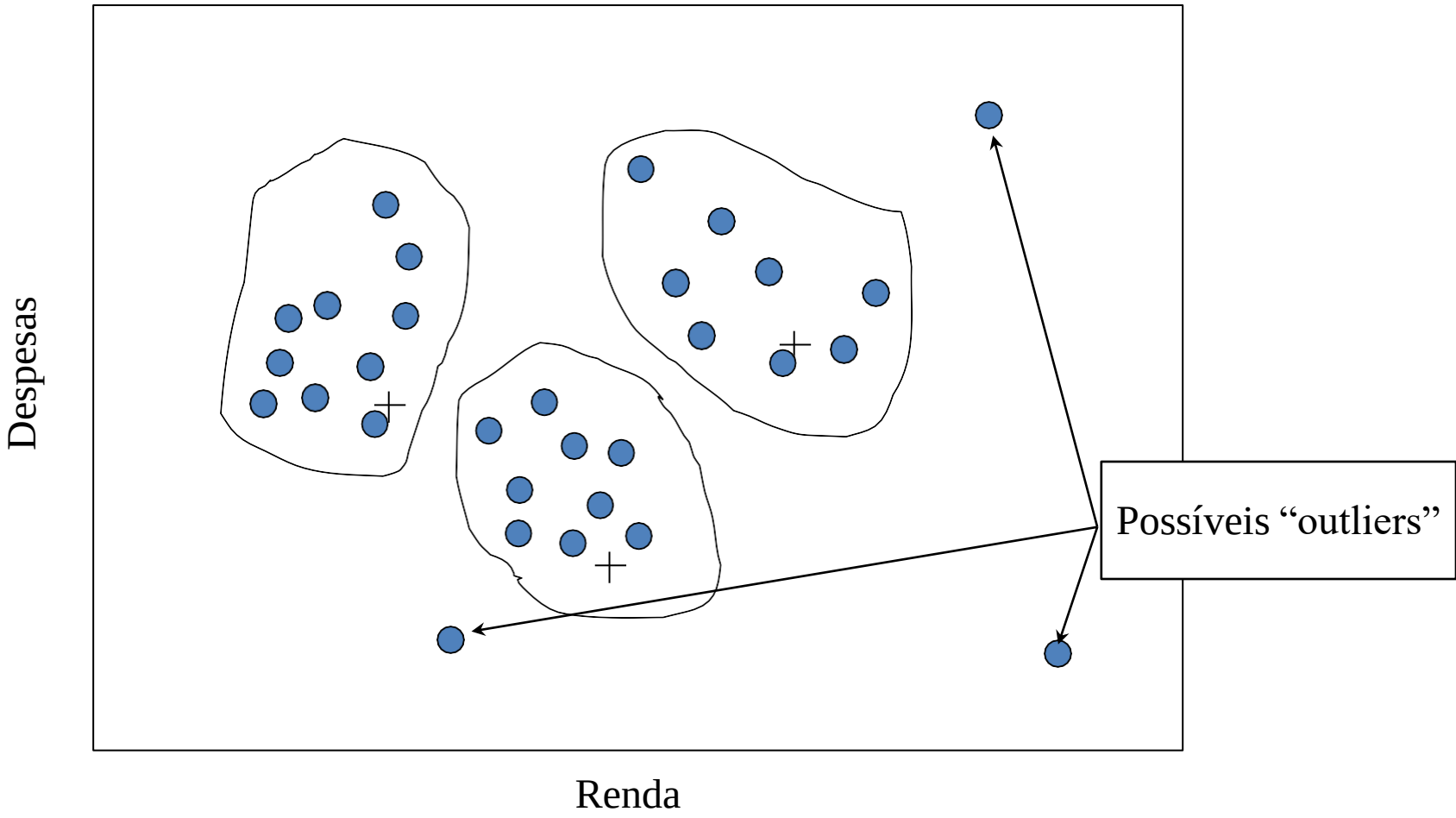
- Particionamento por Distância
  - Divide os dados em N intervalos de mesmo tamanho
  - Se A e B são os valores mínimo e máximo do atributo, a largura dos W intervalos será W=(B-A)/N
  - Esta técnica é a mais direta mas pode ser prejudicada pela presença de “outliers”

- Particionamento por freqüência
  - Divide os dados em N intervalos com o mesmo número de amostras
  - Boa escalabilidade
  - Manipulação de atributos categóricos

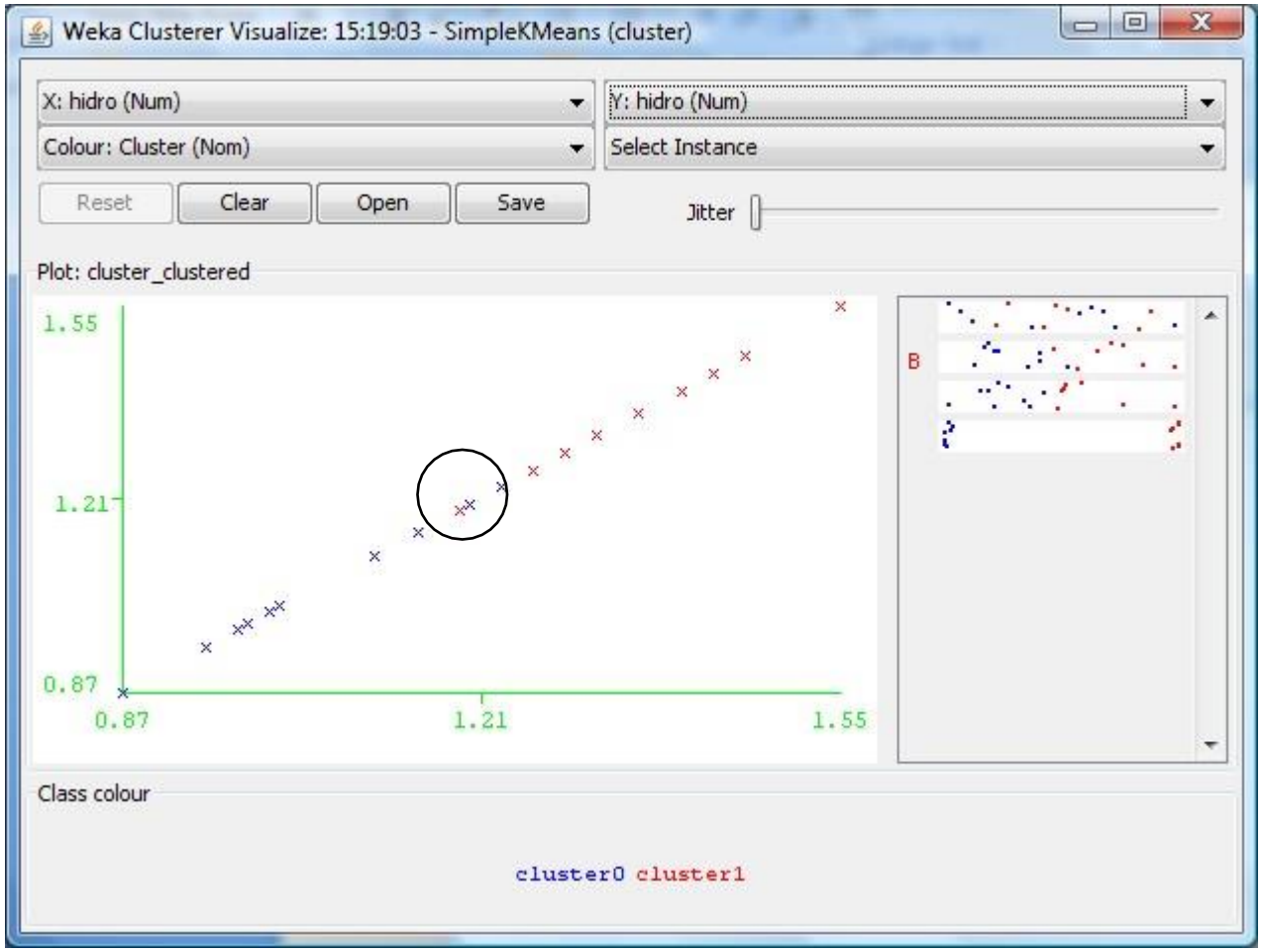
Compartimentalização

- Exemplo
  - Preço de um produto
    - {4,8,9,15,21,21,24,25,26,28,29,34}
  - Particionamento por freqüência
    - Bin 1: 4, 8, 9, 15
    - Bin 2: 21, 21, 24, 25
    - Bin 3: 26, 28, 29, 34
  - Suavização pela média
    - Bin 1: 9, 9, 9, 9
    - Bin 2: 23, 23, 23, 23
    - Bin 3: 29, 29, 29, 29

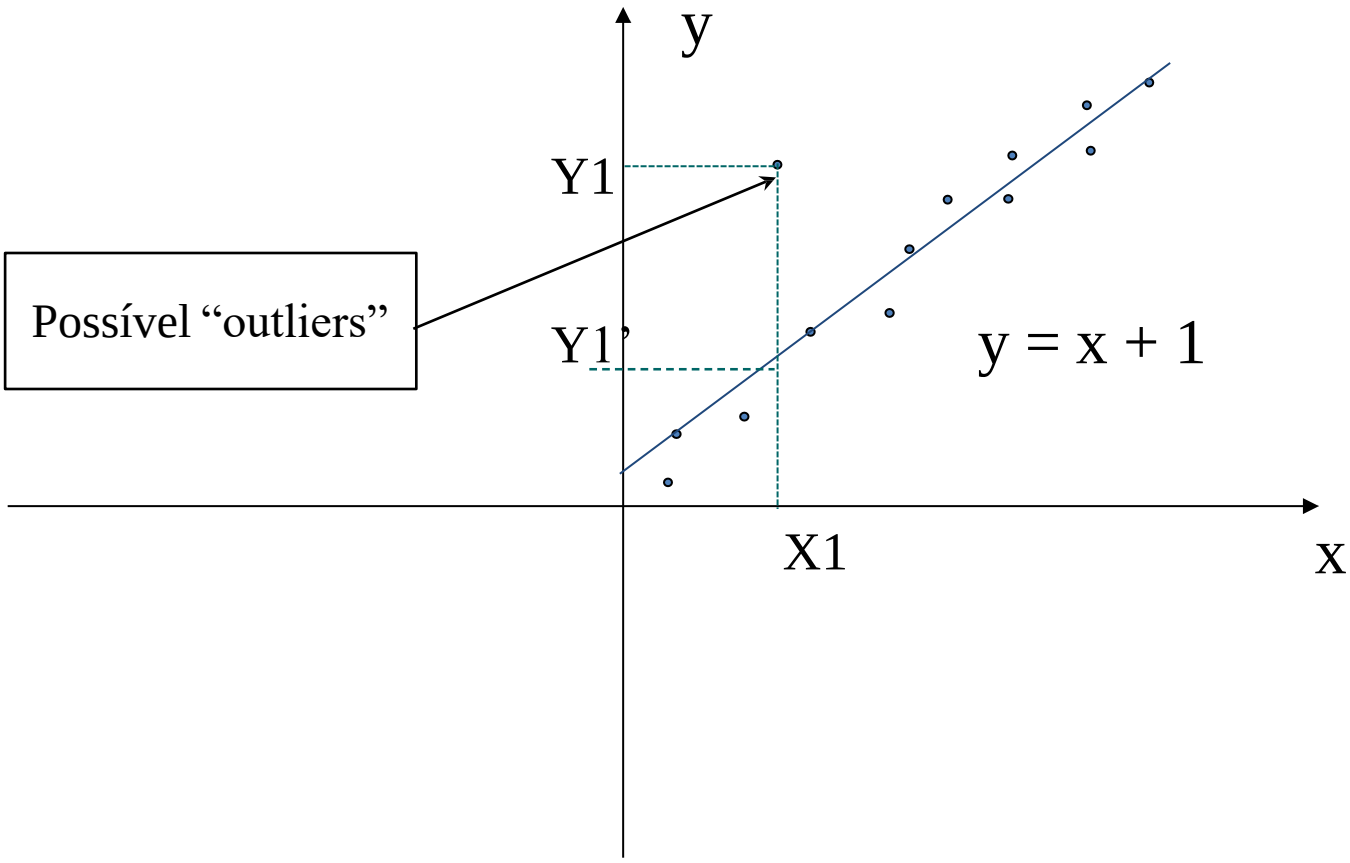
Análise de Cluster

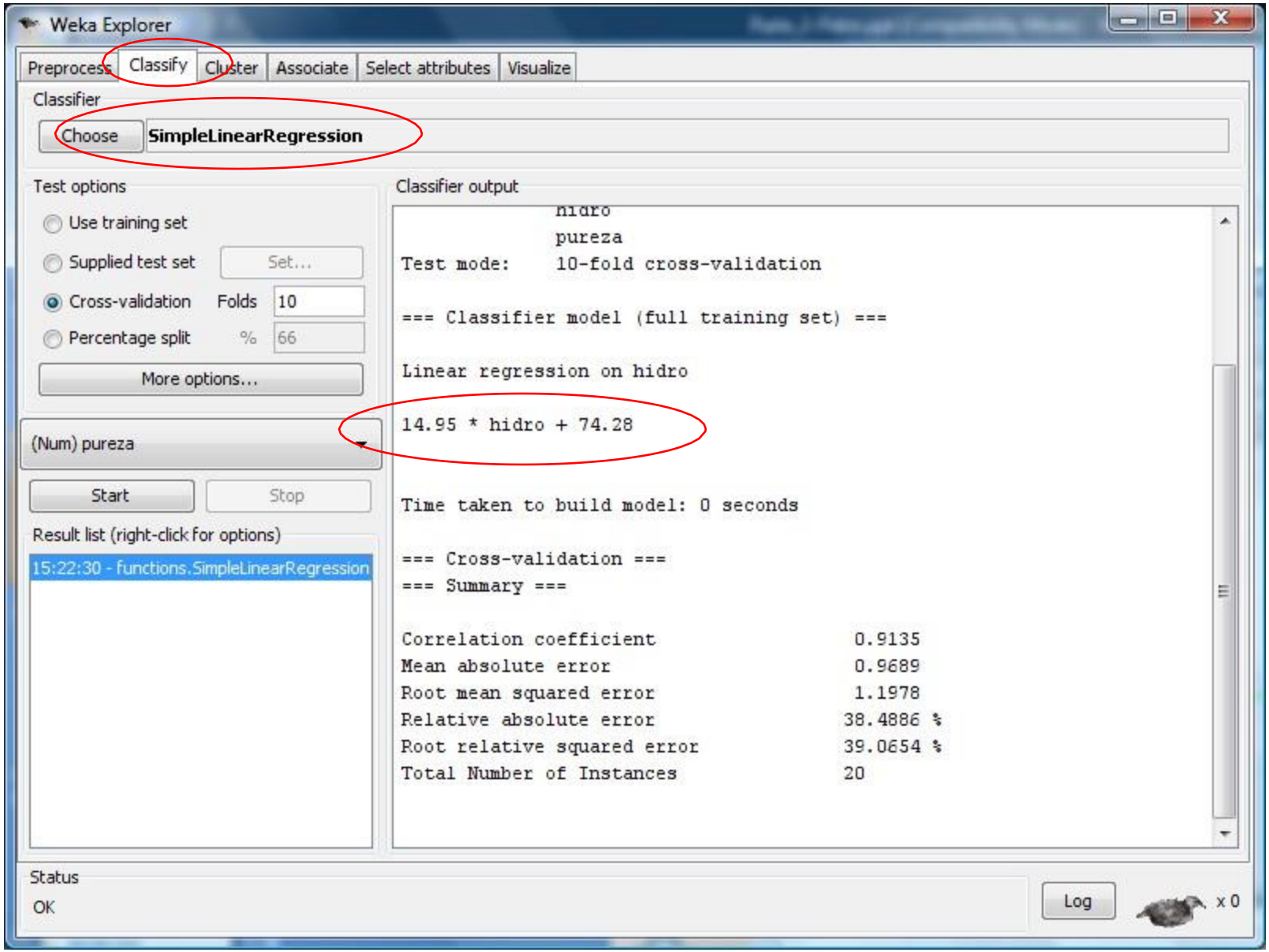


KMeans


Weka Clusterer Visualize: 15:19:03 - SimpleKMeans (cluster)
X: hidro (Num)
Y: hidro (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: cluster_clustered
1.55
1.21
0.87
0.87
1.21
1.55
B
Class colour
cluster0 cluster1

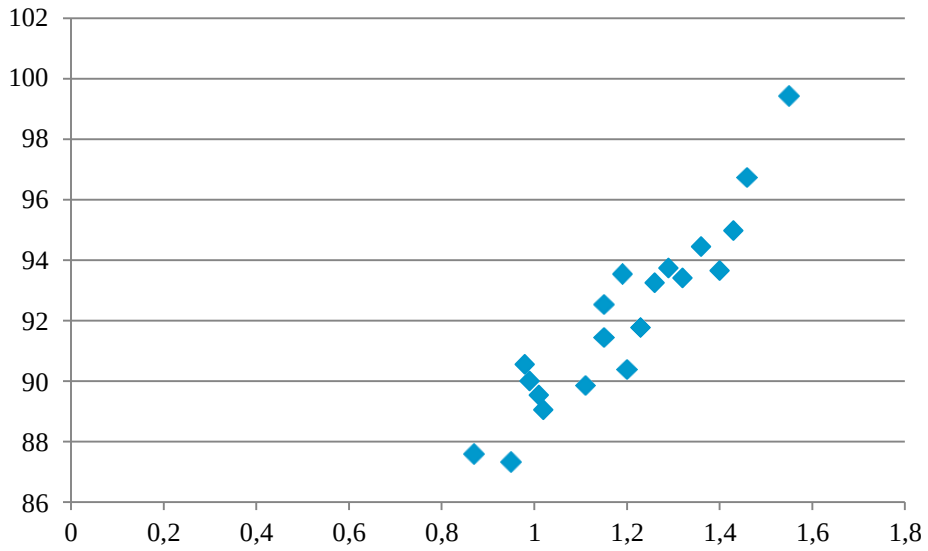
Regressão Linear




Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
SimpleLinearRegression
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Num) pureza
Start
Stop
Result list (right-click for options)
15:22:30 - functions.SimpleLinearRegression
Classifier output
hidro
pureza
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
Linear regression on hidro
14.95 * hidro + 74.28
Time taken to build model: 0 seconds
=== Cross-validation ===
=== Summary ===
Correlation coefficient 0.9135
Mean absolute error 0.9689
Root mean squared error 1.1978
Relative absolute error 38.4886 %
Root relative squared error 39.0654 %
Total Number of Instances 20
Status
OK
Log
x 0

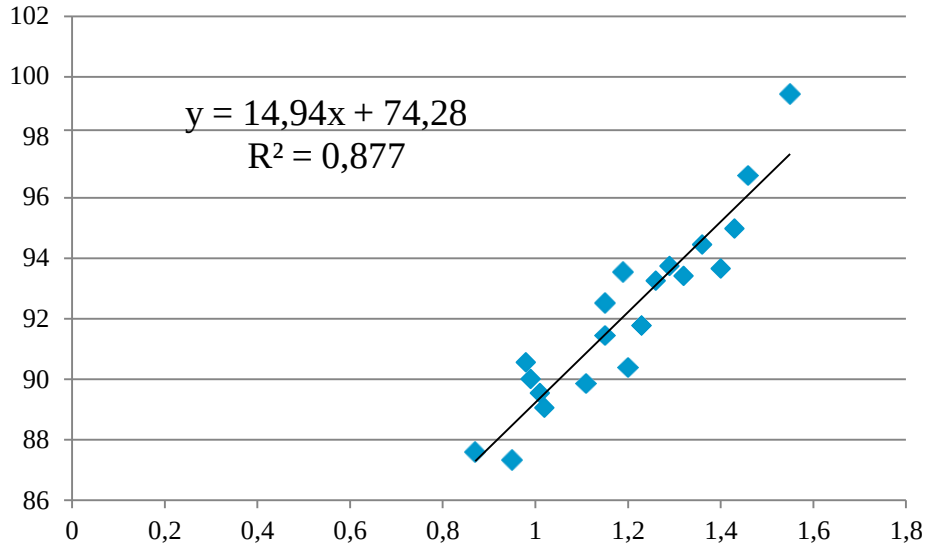
Regrassão Linear

Nível de Hidrocarboneto	Pureza
0,99	90,01
1,02	89,05
1,15	91,43
1,29	93,74
1,46	96,73
1,36	94,45
0,87	87,59
1,23	91,77
1,55	99,42
1,4	93,65
1,19	93,54
1,15	92,52
0,98	90,56
1,01	89,54
1,11	89,85
1,2	90,39
1,26	93,25
1,32	93,41
1,43	94,98
0,95	87,33



Regressão Linear

Nível de Hidrocarboneto	Pureza
0,99	90,01
1,02	89,05
1,15	91,43
1,29	93,74
1,46	96,73
1,36	94,45
0,87	87,59
1,23	91,77
1,55	99,42
1,4	93,65
1,19	93,54
1,15	92,52
0,98	90,56
1,01	89,54
1,11	89,85
1,2	90,39
1,26	93,25
1,32	93,41
1,43	94,98
0,95	87,33



Regression Statistics	
Multiple R	0,93
R Square	0,87
Adjusted R Square	0,87
Standard Error	0,06
Observations	

Principais Tarefas de Pré-Processamento

- Limpeza dos Dados
  - Preenche valores faltantes, suaviza dados ruidosos, identifica ou remove “outliers” e resolve inconsistências
- Integração
  - Dados de origens diferentes devem ser integrados
- Transformação
  - Normalização e agregação
- Redução
  - Tenta reduzir o volume com pouca alteração no resultado final
- Discretização
  - Faz parte do processo de redução, mas tem papel importante, especialmente com dados numéricos

Integração de Dados

- Combina dados de diferentes fontes em uma armazenagem única e coerente
- Detecta e resolve conflitos de valores
  - Para uma mesma entidade do mundo real, valores de atributos oriundos de fontes diferentes podem ter valores diferentes
  - Razões possíveis: representações diferentes, escalas diferentes, etc

Redundância dos Dados

- Redundância geralmente ocorre durante integração
  - Mesmo atributo com nomes diferentes em diferentes bancos de dados
- Dados redundantes podem ser detectados por análise de correlação
- Integração deve ser feita de forma cuidadosa para minimizar redundância e inconsistências nos dados

Principais Tarefas de Pré-Processamento

- Limpeza dos Dados
  - Preenche valores faltantes, suaviza dados ruidosos, identifica ou remove “outliers” e resolve inconsistências
- Integração
  - Dados de origens diferentes devem ser integrados
- Transformação
  - Normalização e agregação
- Redução
  - Tenta reduzir o volume com pouca alteração no resultado final
- Discretização
  - Faz parte do processo de redução, mas tem papel importante, especialmente com dados numéricos

Transformação de Dados

- Suavização: remove ruído dos dados
- Agregação: sumarização
- Normalização: escalona os dados para caírem em uma faixa pequena de valores
  - Normalização min-max
  - Z-Score
  - Escalonamento Decimal
- Construção de Novos Atributos

Transformação de Dados

- Discretização de Variáveis Contínuas/ Transformação de Variáveis Discretas em Contínuas
  - Adequação aos métodos inteligentes a serem utilizados Posteriormente
  - Melhoria de desempenho
- Transformação de Variáveis Contínuas
  - Melhoria na distribuição dos dados
  - Melhoria de desempenho dos métodos inteligentes

Transformação de Dados

A propósito da normalização é minimizar os problemas oriundos do uso de unidades e dispersões distintas entre as variáveis.

- Normalização min-max $$v' = \frac{v - min_A}{max_A - min_A}$$

- Z-Score $$v' = \frac{v - media_A}{desvio_A}$$

- Escalonamento decimal

$$v' = \frac{v}{10^j}$$ Onde j é o menor inteiro tal que $\text{Max}(|v'|)<1$

Normalização Min-Max

CPF Cliente	Despesa
999999999999	1000
111111111111	2000
222222222222	4000

Min=1000
Max=4000

$$v' = \frac{v - min_A}{max_A - min_A}$$

$$v'1 = \frac{1000 - 1000}{4000 - 1000} = 0$$

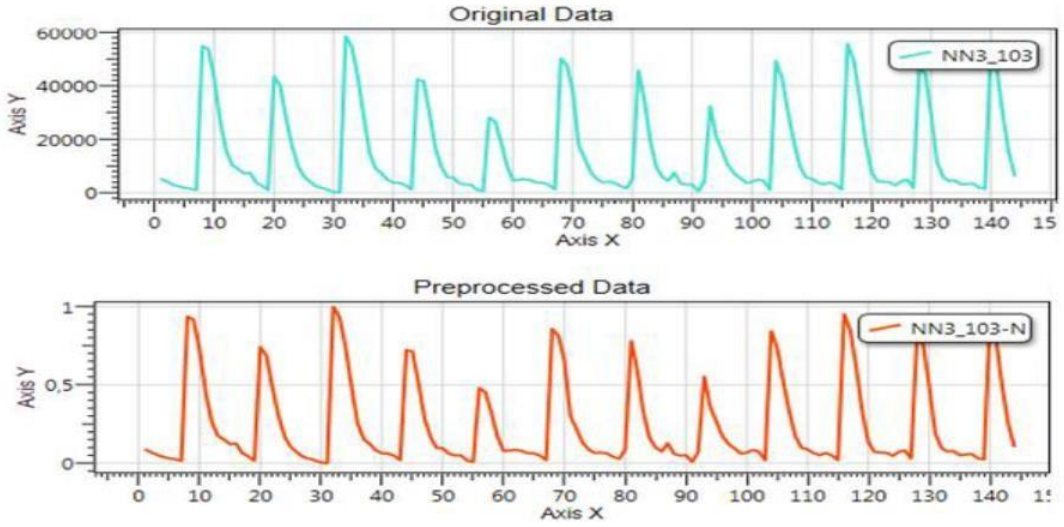
CPF Cliente	Despesa
999999999999	0
111111111111	0.333333
222222222222	1

$$v'2 = \frac{2000 - 1000}{4000 - 1000} = \frac{1000}{3000} = 0.3333$$

$$v'3 = \frac{4000 - 1000}{4000 - 1000} = 1$$

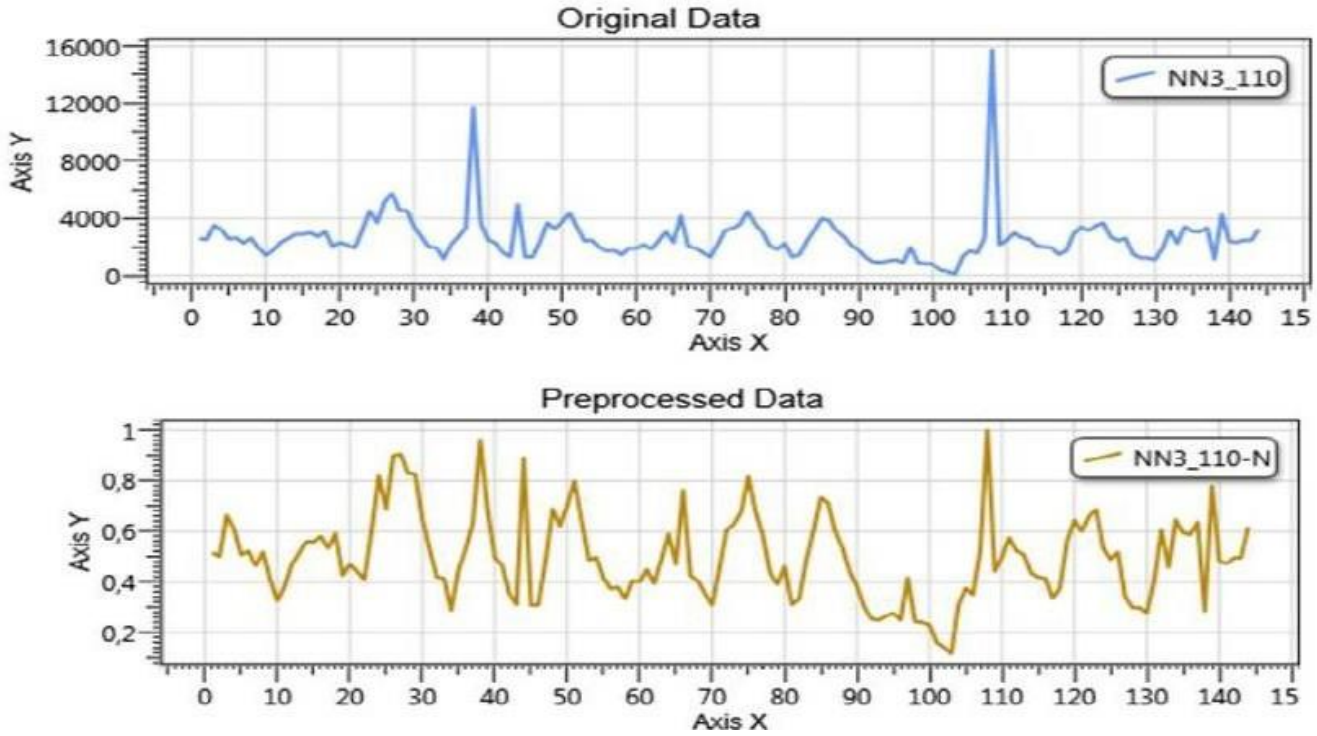
Normalização

- Normalização Padrão (Standard): É feito para mapear diretamente os valores da série do universo original para um novo universo.
- No exemplo, os valores originais estão distribuídos no intervalo [0, 60000] e foram escalados para o universo de [0, 1].



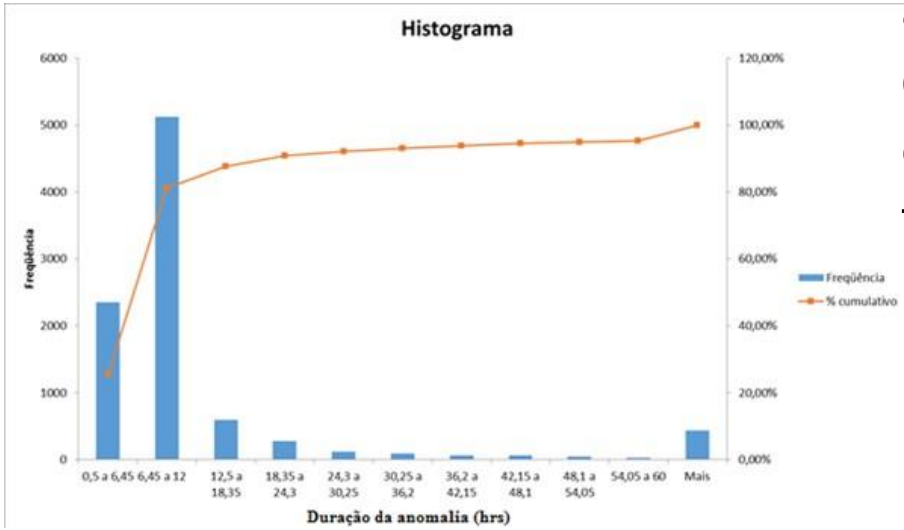
Normalização por Partes

- *Normalização Linear por Partes:* é usado quando o valor de uma variável não é uniformemente distribuído no domínio da mesma.



Normalização por Partes

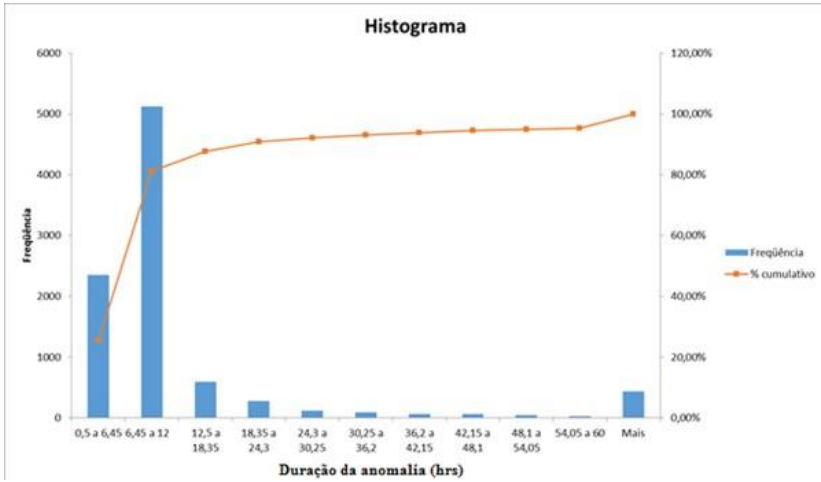
- A normalização por partes, é indicada pois os valores dos dados brutos possuíam uma grande amplitude e com distribuição não uniforme, conforme evidenciado no histograma



Considerando que aproximadamente 80% dos dados numéricos de duração estão entre 0,5 e 12 horas, faz-se a normalização por partes da seguinte forma:

Normalização por Partes

Considerando que aproximadamente 80% dos dados numéricos de duração estão entre 0,5 e 12 horas, faz-se a normalização por partes da seguinte forma:



$$X_{i,\ 0\ a\ 0,8} = \frac{X_i - X_{Min}}{X_{Max} - X_{Min}}$$

Onde X_{min} = 0,5 (mínimo) e X_{max} = 12,5 horas (máximo), incluindo este valor, resultando em valor X_i, de 0 a 0,8 (incluindo 0,8);

$$X_{i,\ 0,8\ a\ 1} = \frac{X_i - X_{Min}}{X_{Max} - X_{Min}}$$

Onde X_{min} = 12,5 (mínimo) (não incluindo esse valor) e X_{max} = máximo horas (máximo), incluindo este valor, resultando em valor X_i, de 0,8 (não incluindo) a 1;

Normalização Z-Score

CPF Cliente	Despesa
999999999999	1000
111111111111	2000
222222222222	4000

$$media = \frac{1000 + 2000 + 4000}{3} = 2333.333$$

$$v' = \frac{v - media_A}{desvio_A}$$

$$desvio = \sqrt{\sum \frac{(1000 - media)^2 + (2000 - media)^2 + (4000 - media)^2}{N - 1}} 1527.5252$$

$$v' = \frac{1000 - 2333.333}{1527.5252} \qquad v' = \frac{2000 - 2333.333}{1527.5252} \qquad v' = \frac{4000 - 2333.333}{1527.5252}$$

CPF Cliente	Despesa
999999999999	-08729
111111111111	-0.2182
222222222222	1.0911

Escalonamento Decimal

CPF Cliente	Despesa
999999999999	1000
111111111111	2000
222222222222	4000

$$v' = \frac{v}{10^{j}}$$

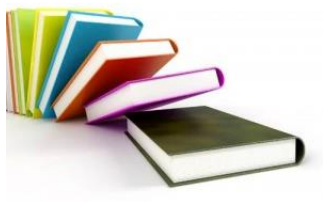
Nesse caso j = 4

CPF Cliente	Despesa
999999999999	0.1
111111111111	0.2
222222222222	0.4

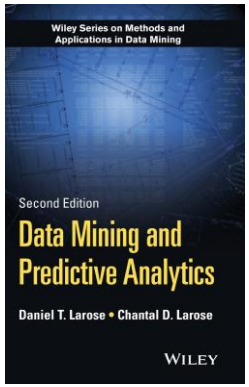
Principais Tarefas de Pré-Processamento

- Limpeza dos Dados
  - Preenche valores faltantes, suaviza dados ruidosos, identifica ou remove “outliers” e resolve inconsistências
- Integração
  - Dados de origens diferentes devem ser integrados
- Transformação
  - Normalização e agregação
- Redução
  - Tenta reduzir o volume com pouca alteração no resultado final
- Discretização
  - Faz parte do processo de redução, mas tem papel importante, especialmente com dados numéricos

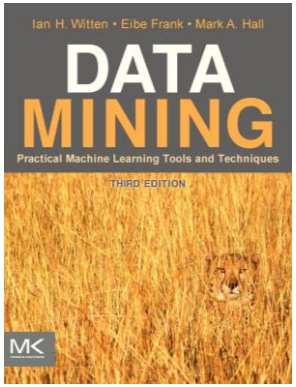
49



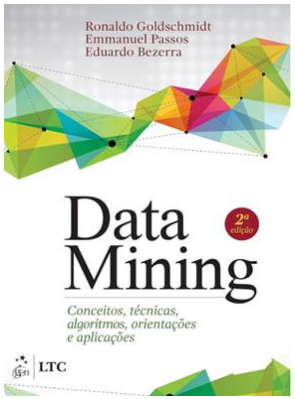
Bibliografia



.



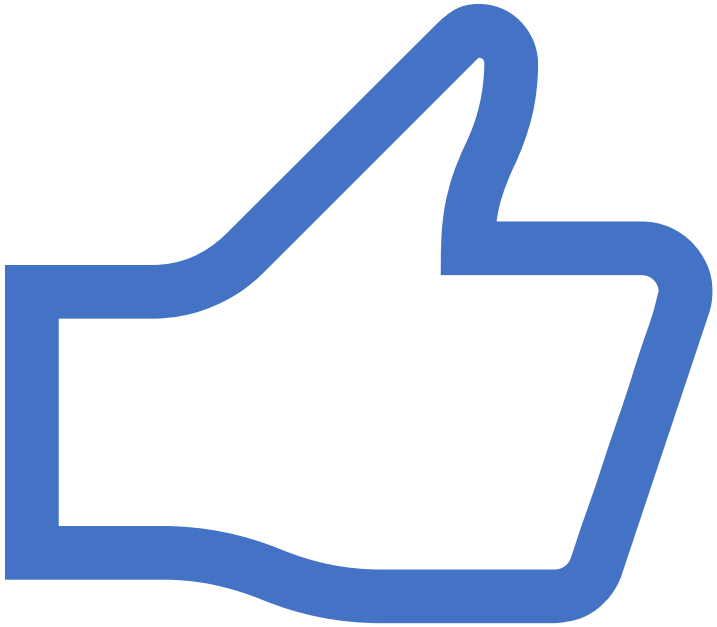
.



.

Dúvidas?

Obrigado !



Apresentador

Thales Levi Azevedo Valente

E-mail:
thales.l.a.valente@gmail.com