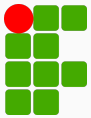
Mineração de Dados

02 - Dados

Marcos Roberto Ribeiro


Instituto Federal Minas Gerais - Campus Bambuí

2018

Introdução

- Uma mineração de dados bem sucedida deve levar em consideração o tipo e a qualidade dos dados
- O tipo de dado diz respeito ao formato dos dados e isto pode definir quais técnicas de mineração podem ser usadas
- A qualidade é importante porque, embora muitas das técnicas de mineração de dados toleram certo nível de ruídos, a melhora da qualidade dos dados garante melhores resultados
- Além disto, em determinadas situações, é preciso modificar os dados brutos para que se tornem mais apropriados para certas técnicas de mineração
- Por exemplo, atributos contínuos podem ser discretizados ou o número de atributos pode ser reduzido

Tipos de Dados

- Um conjunto de dados pode ser visto como uma coleção de objetos de dados
- Os objetos de dados também podem ser chamados de registros, linhas, ponteiros, vetores, eventos, casos, exemplos, observações ou entidades
- Os objetos de dados são descritos por um número de atributos
- Outros nomes para atributos são variáveis, características, campos ou dimensões
- Exemplo: qualquer tabela ou resultado de uma consulta em um banco de dados é um conjunto de dados

Atributos e Medidas

- Um atributo é uma característica de um objeto que pode variar com o tempo ou de um objeto para outro
- Ex.: cor dos olhos e peso de pessoas
- A medição é o processo de dar um valor a um atributo
- Além disto, as propriedades de números são úteis para definir os tipos de atributos

Propriedades de Números

Distinção: $=$ e $\neq$
Ordenação: $<$, $\leq$, $>$ e $\geq$
Adição: $+$ e $-$
Multiplicação: $\times$ e $\div$

Tipos de Atributos

Tipo	Descrição	Exemplos	Operações
Nominal	Valores nominais ($=$ e $\neq$)	CEP, ID, Sexo	Modo, entropia
Ordinal	É possível ordenar os valores ($>$ e $<$)	Notas em letras	Mediana, porcentagens
Intervalar	As diferenças entre os valores são significativas ($+$ e $-$)	Datas, temperatura (Celsius)	Média, desvio padrão
Proporcional	Existem diferenças e proporções entre os valores ($\times$ e $\div$)	Idade, temperatura (Kelvin)	Média, geométrica

- Os atributos nominais e ordinais são chamados de *categorizados* ou *qualitativos*. Já os atributos intervalares e proporcionais são chamados de *numéricos* ou *categóricos*
- Os atributos também podem ser discretos (contáveis) ou contínuos (valores do tipo real)
- Um atributo é *assimétrico* se apenas a presença de um valor é importante (diferente de zero)

Tipos de Conjuntos de Dados

- Podemos classificar os tipos de conjuntos de dados em três grandes grupos:
  - Dados em registros
  - Dados baseados em grafos
  - Dados ordenados
- Além disto é importante analisar as seguintes características dos dados:

  Dimensão: Número de atributos

  Dispersão: Poucos registros com valores diferentes de zero

  Resolução: Nível de detalhes

Dados em Registros

- Este é o tipo de conjunto de dados mais comum
- Consiste em uma coleção de registros sendo que cada registro tem um número fixo de atributos

Dados de Transações

- Dados de transações são um tipo especial de dados em registros, onde cada registro (transação) é um conjunto de itens
- Um exemplo típicos são os carrinhos de compras de clientes de supermercados

Exemplo de Carrinho de Compras

TID	Itens
1	pão, refrigerante, leite
2	cerveja, pão
3	cerveja, refrigerante, fralda, leite
4	cerveja, pão, fralda, leite
5	refrigerante, fralda, leite

Dados em Registros - Matriz de Dados

Matriz de Dados

- Se todos os atributos de um conjunto de dados são numéricos, então tal conjunto pode ser interpretado como uma matriz $m \times n$ (m linhas e n colunas)
- As matrizes de dados são interessantes porque podemos aplicar operações de matrizes para transformar os dados

Exemplo de Matriz de Dados

Carga X	Carga Y	Distância	Espessura
10.23	5.27	15.22	1.2
12.65	6.25	16.22	1.1
13.54	7.23	17.34	1.2
14.27	8.43	18.45	0.9

Dados em Registros - Matriz de Dados Dispersos

Matriz de Dados

- A matriz de dados dispersos é um caso especial de matriz de dados onde os atributos são do mesmo tipo e dispersos
- Um exemplo interessante é a matriz de termos de documentos contendo o número de vezes que o termo aparece no documento

Exemplo de Matriz de Termos de Documentos

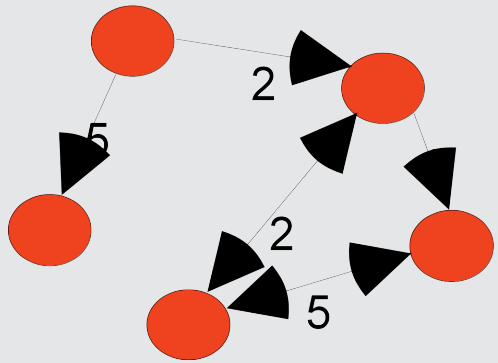
	equipe	treinador	jogo	bola	placar	partida	vitória	derrota
Doc 1	3	0	5	0	2	6	0	2
Doc 2	0	7	0	2	1	0	0	0
Doc 3	0	1	0	0	1	2	2	3

Dados Baseados em Grafos

- Os grafos são representações poderosas para dados
- Existem basicamente dois tipos de conjuntos de dados baseados em grafos:

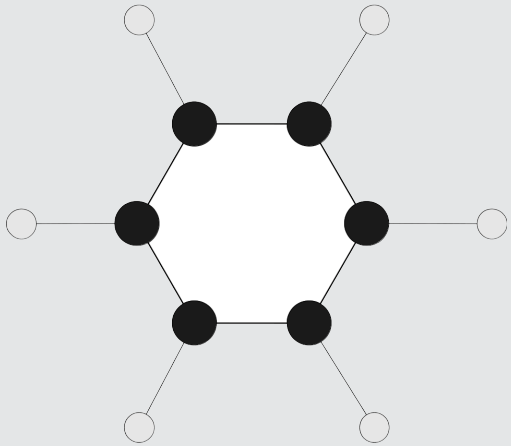
Grafos que representam os relacionamentos entre objetos

- Por exemplo, os links entre páginas na Internet



Objetos de Dados que São Grafos

- Por exemplo, estruturas químicas



Dados Ordenados

- No caso dos dados ordenados, os atributos têm relacionamentos que envolvem ordenação no tempo ou espaço

- Basicamente existem os seguintes tipos:
  - Dados sequenciais
  - Dados de sequência
  - Dados de séries temporais
  - Dados espaciais

Dados Ordenados - Dados Sequenciais

- Os dados sequenciais basicamente são dados de registros com um tempo associado a cada registro
- Exemplo: Transações sequenciais de clientes

Cliente	Sequência de compras
c1	(t1: A, B) (t2: C,D) (t3: A, E)
c2	(t3: A, D) (t4: E)
c3	(t2: A, C)

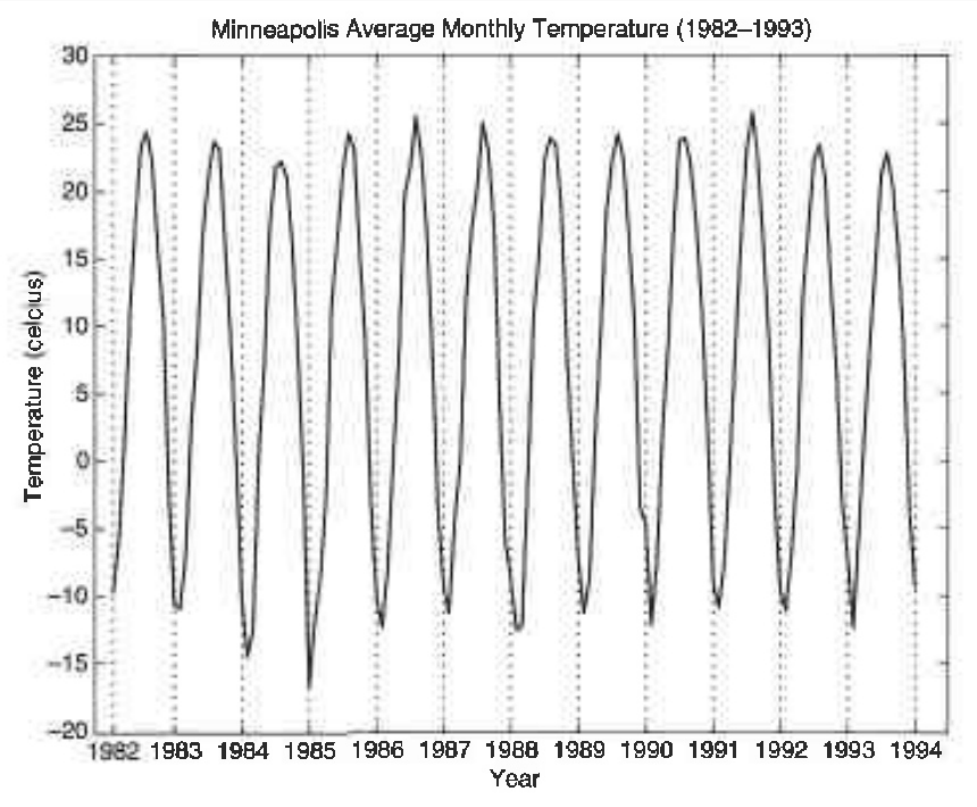
Dados Ordenados - Dados de Sequência

- Os dados de sequência são sequências de dados individuais, como palavras ou letras
- Os dados de sequência são parecidos com os dados sequenciais, mas não possuem marcação de tempo
- Exemplo: Sequências de DNA

```
GGTTCCGCCTTCAGCCCCGC
CGCAGGGCCCGCCCCGCGCC
GAGAAGGGCCCGCCTGGCGG
GGGGGAGGCGGGGCCGCCCG
CCAACCGAGTECGACCAGGT
CCCTCTGCTCGGCCTAGACC
GCTCATTAGGCGGCAGCGGA
GCCAAGTAGAAEAEGCGAAG
TGGGCTGCCTGCTGCGACCA
```

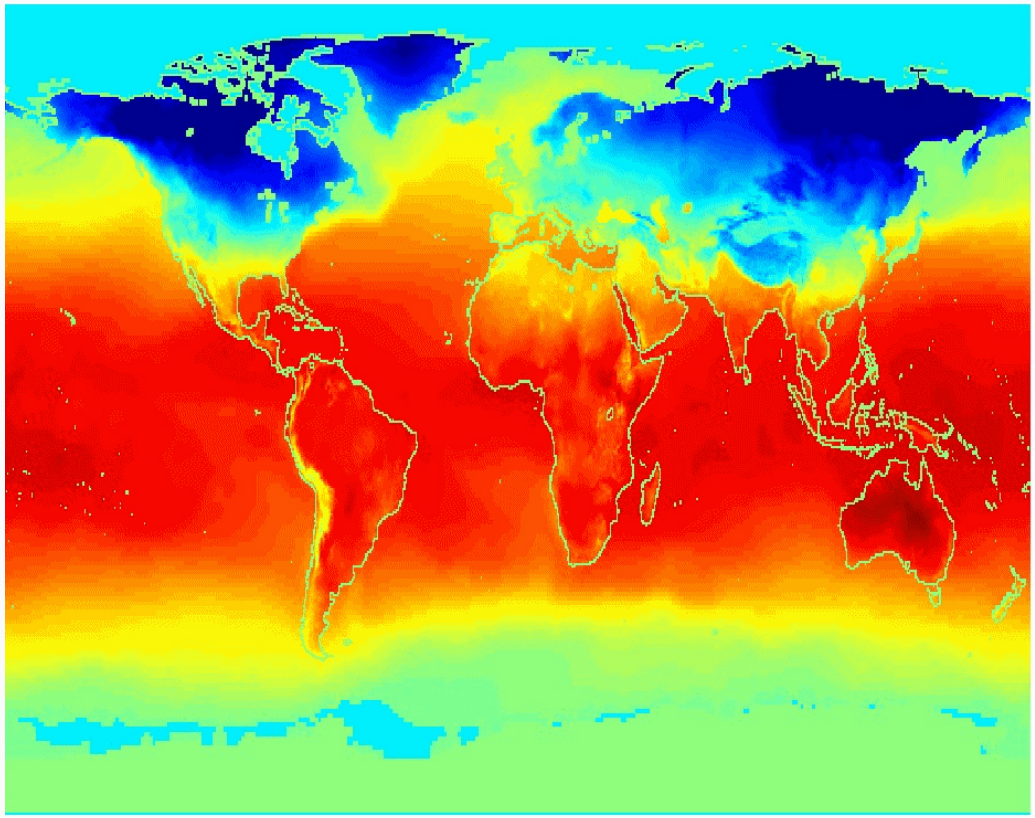
Dados Ordenados - Dados de Séries Temporais

- Os dados de séries temporais são um tipo especial de dados sequenciais que representa uma série de medições feitas no decorrer do tempo
- Exemplo: Sequência de medições de temperatura



Dados Ordenados - Dados Espaciais

- Os dados espaciais possuem informações geográficas
- Normalmente os dados que estão fisicamente próximos tendem a ter outras características semelhantes
- Exemplo: temperatura de pontos geográficos



Pré-Processamento de Dados

- As principais técnicas de pré-processamento de dados são:
  - Agregação
  - Amostragem
  - Redução de dimensionalidade
  - Seleção de subconjuntos de recursos
  - Criação de recursos
  - Discretização e binarização
  - Transformação de variáveis

Agregação

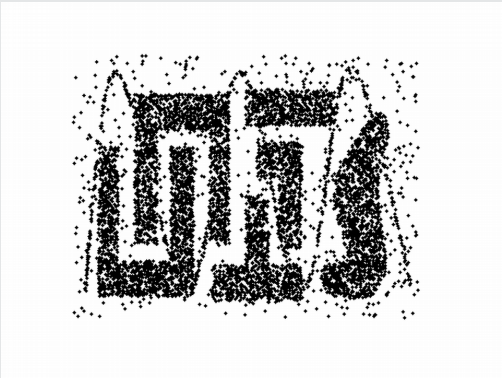
- “Menos é mais!”
- A agregação consiste em combinar vários registros em um único registro
- Os conjuntos de dados agregados são menores e com isto consumem menos memória e menos tempo de processamento
- A agregação serve também para se ter uma visão de mais alto nível dos dados
- Uma desvantagem da agregação é a potencial perda de detalhes

Amostragem I

- O tipo mais simples de amostragem é a *amostragem aleatória simples* onde há uma probabilidade igual de selecionar qualquer item
- Durante a seleção dos elementos podemos proceder de duas maneiras:
  - Remover o item selecionado da população (*amostragem sem substituição*)
  - Não remover os itens selecionados da população. Neste caso pode ocorrer repetições (*amostragem com substituição*)
- A amostragem aleatório simples pode não ser adequado quando a população possui poucos objetos de um determinado tipo
- Uma solução para este problema é utilizar a *amostragem estratificada* que traz objetos de todos os grupos da população

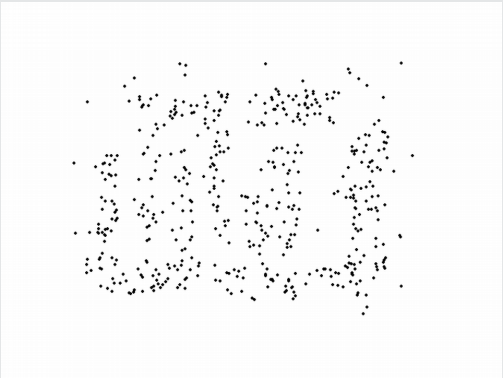
Amostragem II

Exemplo



8000 pontos

2000 pontos



500 pontos

- Um dos problemas relacionados a amostragem é definir o tamanho da amostra
- Observe as figuras com diferentes números de pontos
- O tamanho da amostra deve ser representativo e pequeno

Redução da Dimensionalidade

- Conjuntos de dados com um grande número de dimensões pode acarretar problemas às técnicas de mineração de dados tais como modelos de baixa precisão e alto consumo de memória e tempo para execução da mineração (*maldição da dimensionalidade*)
- Desta maneira, foram criados métodos para reduzir a dimensionalidade de conjuntos de dados sem que os elementos percam sua representatividade
- uma das técnicas mais usadas para a redução de dimensionalidade é a *Análise de Componentes Principais (PCA)*

Seleção de Subconjunto de Características

- Outra maneira de reduzir a dimensionalidade é usar apenas um subconjunto de características
- Os seguintes atributos podem ser descartados:
  Atributos redundantes: são atributos que duplicam informações de outros atributos. Exemplo: o valor total pago por um produto que é calculado multiplicando-se a quantidade e o valor unitário
  Atributos irrelevantes: são atributos que não possuem informações úteis. Exemplo: ID e CPF de alunos
- Alguns atributos podem ser eliminados de forma mais trivial, mas obter o melhor subconjunto de atributos não é uma tarefa fácil
- O ideal seria testar todas as combinações possíveis, mas muitas vezes isto não é viável na prática

Abordagens para Seleção de Características

- Existem três abordagens padrões:

  Abordagem Interna: A seleção de atributos ocorre naturalmente como parte do algoritmo de mineração

  Abordagem de Filtro: Os atributos são selecionados antes da mineração usando alguma técnica independente do algoritmo de mineração, por exemplo, correlação

  Abordagem de Envoltório: Este método usa o algoritmo de mineração com caixa preta para encontrar o melhor subconjunto de atributos. Semelhante ao teste ideal, mas sem testar todas as cominações de atributos

Criação de Recursos

- A criação de recursos visa criar novos atributos a partir dos atributos existentes
- A ideia é que o número de atributos seja menor para termos as mesmas vantagens da redução de dimensionalidade
- As principais metodologias são as seguintes:
  - Extração de características
  - Mapeamento de dados para um novo espaço
  - Construção de características

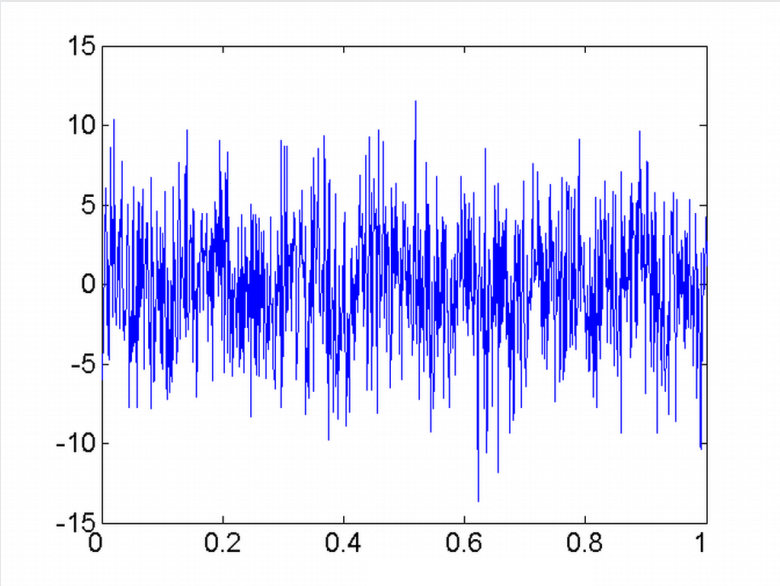
Extração de Características

- A extração de características tem como objetivo criar um novo conjunto de atributos para representar os atributos originais dos objetos
- Esta metodologia é amplamente utilizada quando precisamos trabalhar com imagens
- Desta maneira, uma imagens contendo milhares de pixeis pode ser representa de outras formas, por exemplo, histograma de cores

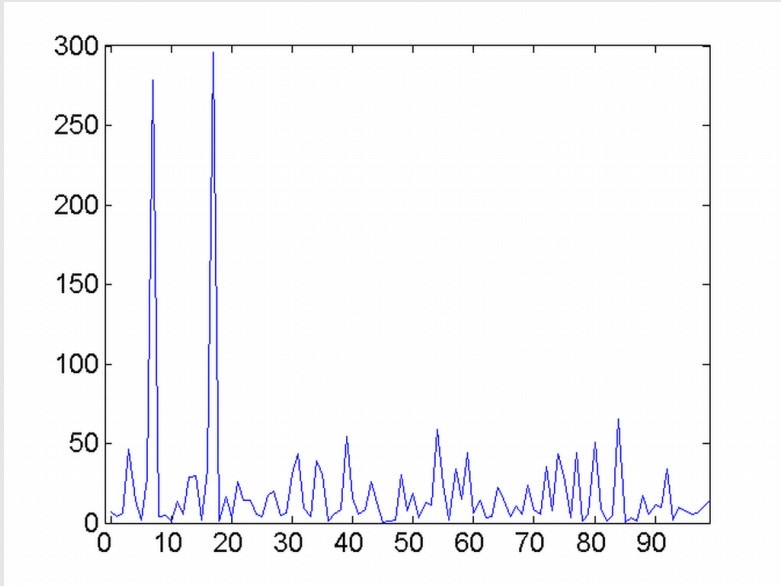
Mapeamento de Dados para um Novo Espaço

- Padrões periódicos com quantidades significativa de ruídos são difíceis de detectar
- Porém, a aplicação de transformações como a *transformação de Fourier* podem tornar tais padrões mais explícitos

Exemplo



Série com ruído



Transformação de Fourier

Construção de Recursos

- Em determinadas situações, a criação de novos atributos pode facilitar o processo de mineração de dados

Exemplo

- Considere um banco de dados de itens históricos contendo volume e massa
- O objetivo é classificar os itens de acordo com o material que são feitos (madeira, barro, bronze, etc.)
- um atributo de densidade (massa / volume) pode facilitar muito esta tarefa
- A construção de recurso normalmente é feita a partir do conhecimento do domínio do banco de dados

Discretização e Binarização

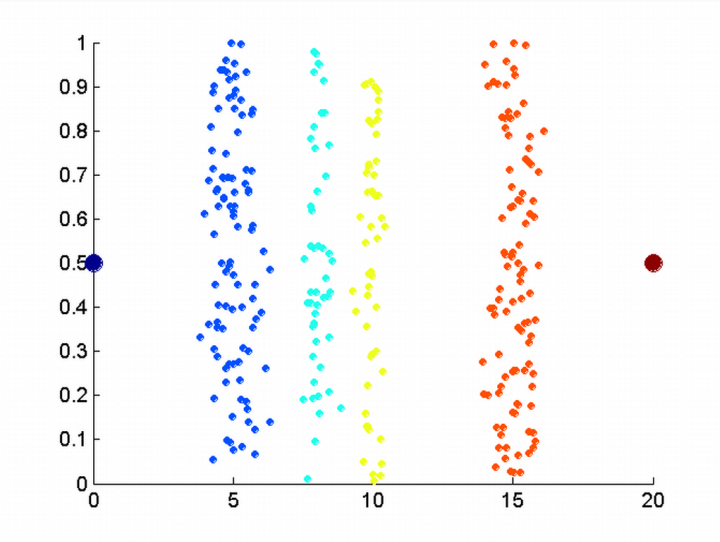
- Alguns algoritmos de mineração requerem que os atributos estejam categorizados ou ou sejam atributos binários
- Desta forma, pode ser necessário realizar *discretização* ou *binarização*
- A discretização transforma atributos contínuos em atributos categóricos
- Já a binarização transforma atributos numéricos em atributos binários

Exemplo de Binarização

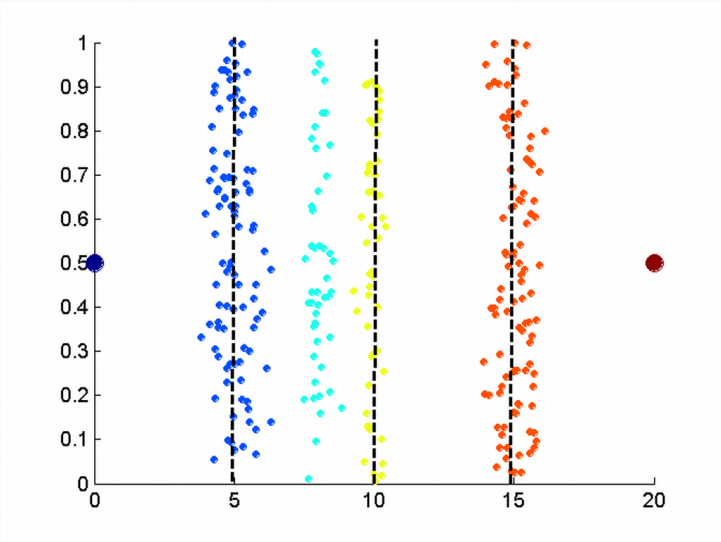
- Usando $\log_2(m)$ bits (m é o número de valores)

Valor Original	Valor Inteiro	x_1	x_2	x_3
terrível	0	0	0	0
fraco	1	0	0	1
satisfatório	2	0	1	0
bom	3	0	1	1
excelente	4	1	0	0

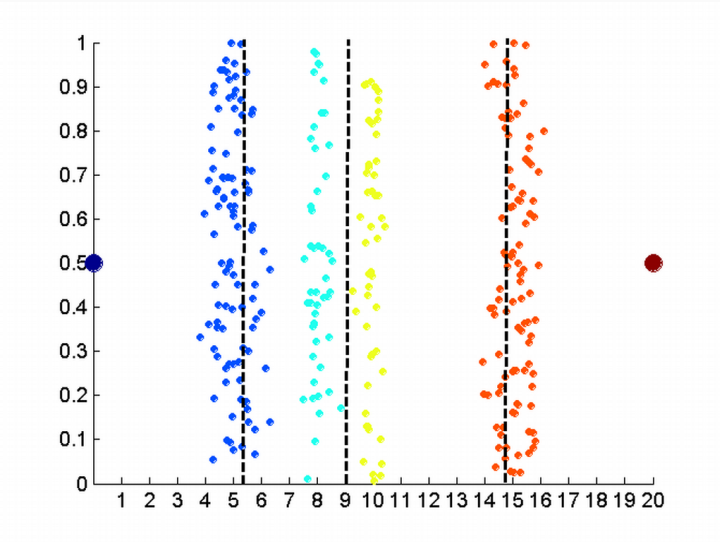
Exemplo de Discretização



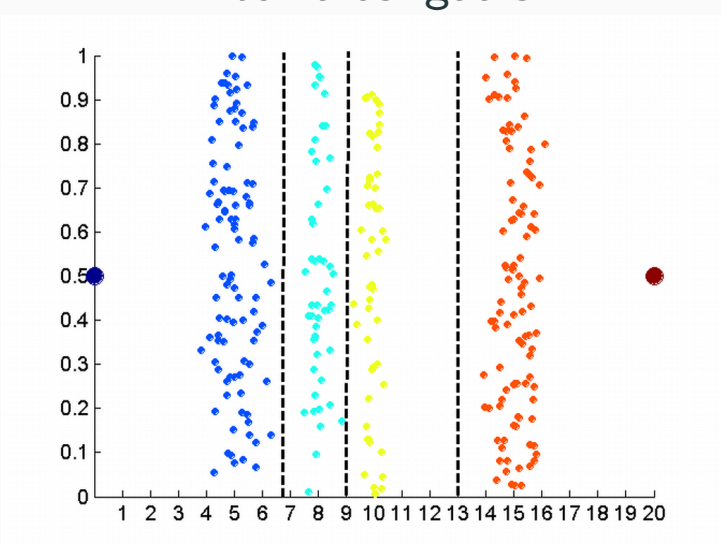
Dados



Intervalos iguais



Discretização de frequência igual



Discretização K-means

Transformação de Variáveis

- A transformação de uma variável faz uma modificação em todos os valores desta variável dentro da base de dados
- Estas transformações podem ser necessárias para evitar discrepâncias entre valores de duas variáveis distintas
- Por exemplo, no caso de pessoas com atributos de idade e renda
  - A diferença entre duas idades pode ser muito menor do que a diferença entre duas rendas
  - Mesmo que os dois atributos tenham a mesma importância, alguns cálculos podem ser mais afetados por esta diferença de magnitude

Medidas de Semelhança e Diferença

- As medidas de semelhança e diferença são importantes para diversas técnicas de mineração de dados
- Em muitos casos, o conjunto inicial de dados é dispensado e depois do cálculo destas medidas
- Muitas técnicas usam alguma medida de distância para representar o qual próximos ou distantes estão dois objetos

Semelhanças e Diferenças para Atributos Simples

Tipo de Atributo	Diferença	Semelhança
Nominal	$d = \begin{cases} 0, & \text{se } x = y \\ 1, & \text{se } x \neq y \end{cases}$	$s = \begin{cases} 1, & \text{se } x = y \\ 0, & \text{se } x \neq y \end{cases}$
Ordinal[1]	$d = \frac{\lvert x - y \rvert}{n - 1}$	$s = d - 1$
Intervalar ou proporcional	$d = \lvert x - y \rvert$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, etc.

[1] Valores entre 0 e $n - 1$, onde n é o número de termos

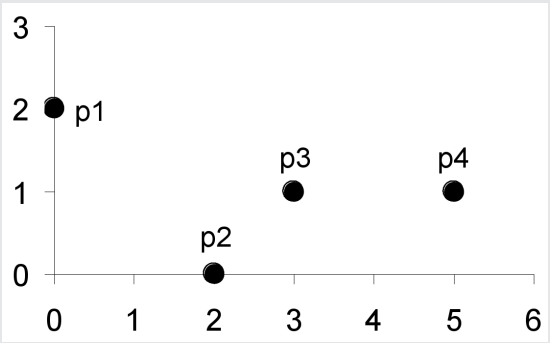
Distâncias Entre Objetos de Dados - Distâncias

- Uma das medidas de distâncias mais usadas em técnicas de mineração de dados é a **distância euclidiana**:

$$d(x, y) = \sqrt{\Sigma_{k=1}^{n}(x_k - y_k)^2}$$

- Em geral, as medidas de distância são usadas para computar a **matriz de distância** contendo as distâncias entre os objetos do conjunto de dados
- Exemplo:

Pontos



	x	y
p_1	0	2
p_2	2	0
p_3	3	1
p_4	5	1

Matriz de Distância

	p_1	p_2	p_3	p_4
p_1	0.0	2.8	3.2	5.1
p_2	2.8	0.0	1.4	3.2
p_3	3.2	1.4	0.0	2.0
p_4	5.1	3.2	2.0	0.0

Propriedades de Distâncias

- As medidas de distância que satisfazem as propriedades a seguir são chamadas de **métricas**

Positividade

- $d(x, y) \geq 0$ para todo x e y
- $d(x, x) = 0$ para todo x

Simetria

- $d(x, y) = d(y, x)$ para todo x e y

Diferença Triangular

- $d(x, z) \geq d(x, y) + d(y, z)$ para todo x, y e z

Referências I


BRAGA, L. P. V. (2005).
***Introdução à mineração de dados*.**
E-Papers, Rio de Janeiro, 2 edition.


TAN, P.-N., STEINBACH, M., and KUMAR, V. (2009).
***Introdução ao data mining: mineração de dados*.**
Ciência Moderna, Rio de Janeiro.