Mineração de Dados
Aula 1 – parte 2

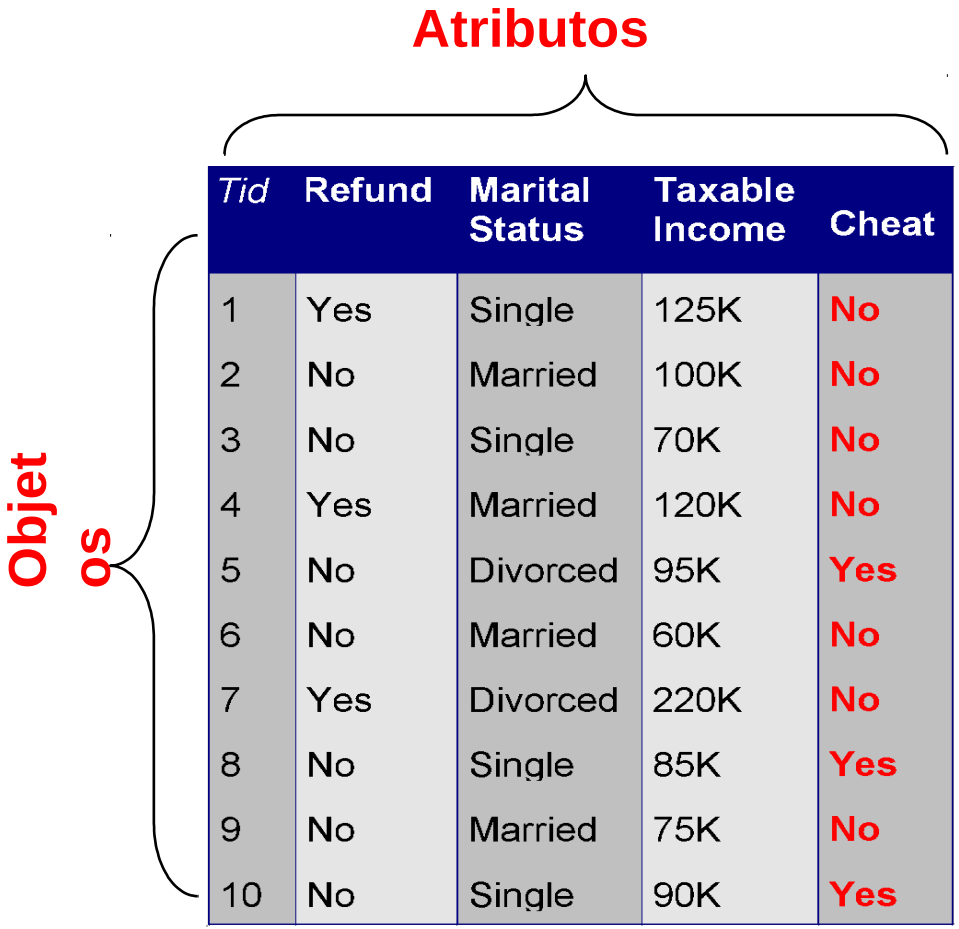
Especialização em Ciência de Dados e suas Aplicações

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Atributos

- Um atributo é uma propriedade de um objeto

- Exemplos: cor de olho, temperatura, etc.
  - Atributo é também conhecido como **variável, campo, característica, *feature***



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Medidas de similaridade e dissimilaridade

- Medida de similaridade
  - Medida numérica de quão parecidos dois objetos de dados são.
  - É maior quando são mais parecidos
  - Geralmente estão no intervalo [0,1]

- Medida de dissimilaridade
  - Medida numérica de quão diferente dois objetos são
  - Dissimilaridade mínima é geralmente 0

- Proximidade refere a similaridade ou dissimilaridade

Distância Euclideana

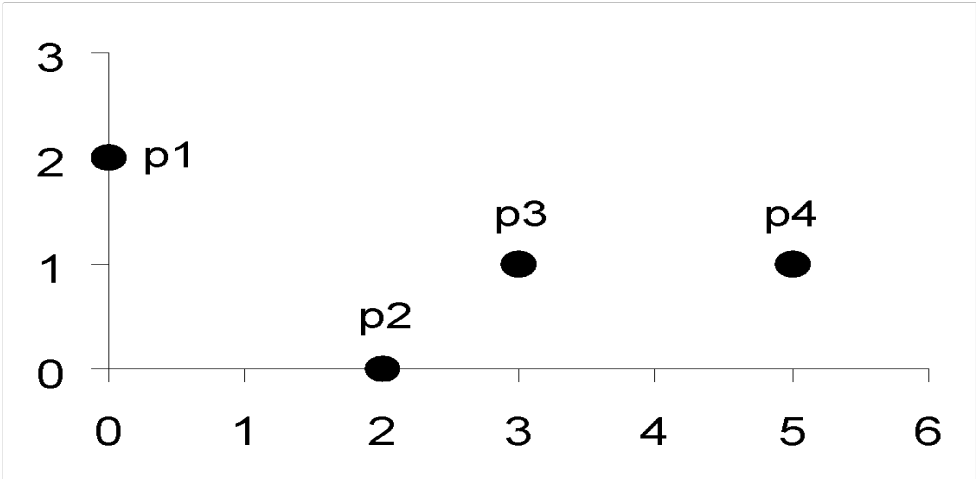
- Distância Euclideana

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

onde *n* é o número de dimensões (atributos) e x_k e y_k são os k^{th} atributos ou objetos de dados $\mathbf{x}$ e $\mathbf{y}$.

- Padronização é necessária, se a escala difere

Distância Euclideana



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Matriz de distância

Distância Manhattan

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

r = 1. Distância Manhattan (City block, taxicab, L_1 norm).

n é o número de dimensões (atributos) e x_k e y_k são os $k^{éssimos}$ atributos ou objetos de dados $\mathbf{x}$ e $\mathbf{y}$.

Comparação

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

Manhattan

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

Euclideana

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Matriz de distância

Similaridade entre vetores binários

- Situação comum: objetos *p* e *q* possuem apenas atributos binários
- Computar a similaridade assim:
  f_{01} = # de atributos onde *p* era 0 e *q* é 1
  f_{10} = # de atributos onde *p* era 1 e *q* é 0
  f_{00} = # de atributos onde *p* era 0 e *q* é 0
  f_{11} = # de atributos onde *p* era 1 e *q* é 1

Simple Matching (**SMC**) and Coeficiente de Jaccard (**J**)

SMC = número de correspondências “11” e “00” / número de atributos
= $(f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$

J = número de correspondências “11” / número de atributos não zero
= $(f_{11}) / (f_{01} + f_{10} + f_{11})$

SMC vs Jaccard: Exemplo

x = 1 0 0 0 0 0 0 0 0 0
y = 0 0 0 0 0 0 1 0 0 1

$f_{01} = 2$
$f_{10} = 1$
$f_{00} = 7$
$f_{11} = 0$

SMC $= (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$
$= (0+7) / (2+1+0+7) = 0.7$

J $= (f_{11}) / (f_{01} + f_{10} + f_{11}) = 0 / (2 + 1 + 0) = 0$

Similaridade do Cosseno

- Se $\mathbf{d}_1$ e $\mathbf{d}_2$ vetores numéricos, então

$$\cos(\mathbf{d}_1, \mathbf{d}_2) = \langle \mathbf{d}_1, \mathbf{d}_2 \rangle / \|\mathbf{d}_1\| \|\mathbf{d}_2\| ,$$

onde $\langle \mathbf{d}_1, \mathbf{d}_2 \rangle$ indica o produto escalar dos vetores, $\mathbf{d}_1$ e $\mathbf{d}_2$, e $\| \mathbf{d} \|$ é a magnitude do vetor $\mathbf{d}$.

- Ex:

$\mathbf{d}_1 = 3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0$

$\mathbf{d}_2 = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2$

$\langle \mathbf{d}_1, \mathbf{d2} \rangle = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

$\| \mathbf{d}_1 \| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$\| \mathbf{d}_2 \| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.449$

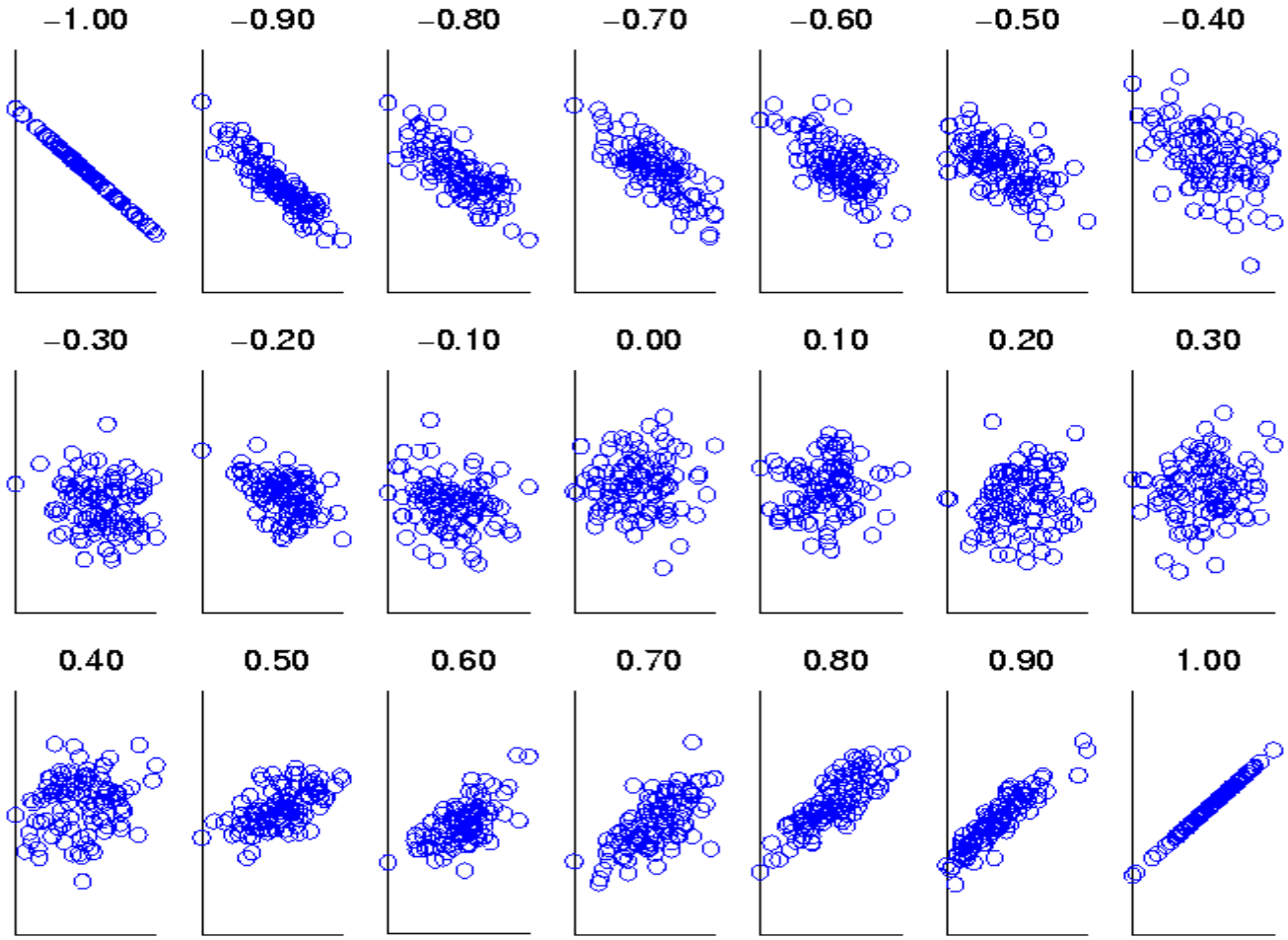
$\cos(\mathbf{d}_1, \mathbf{d}_2) = 0.3150$

Medida de correlação linear

$$\mathrm{corr}(\mathbf{x}, \mathbf{y}) = \frac{\mathrm{covariance}(\mathbf{x}, \mathbf{y})}{\mathrm{standard\_deviation}(\mathbf{x}) * \mathrm{standard\_deviation}(\mathbf{y})}$$

p=1 Significa uma correlação perfeita positiva entre as duas variáveis.
p=-1 Significa uma correlação negativa perfeita entre as duas variáveis - Isto é, se uma aumenta, a outra sempre diminui.
p=0 Significa que as duas variáveis não dependem linearmente uma da outra. No entanto, pode existir uma dependência não linear.

Avaliando correlação visualmente



Scatter plots mostrando a similaridade de –1 a 1.

Desvantagem de correlação

- **x** = (-3, -2, -1, 0, 1, 2, 3)
- **y** = (9, 4, 1, 0, 1, 4, 9)

$y_i = x_i^2$

- corr = (-3)(5)+(-2)(0)+(-1)(-3)+(0)(-4)+(1)(-3)+(2)(0)+3(5) / (6 * 2.16 * 3.74)
  = 0

Tipos de conjunto de dados

Registro

- Matrizes
- Documentos
- Transações

Grafos

- WWW

Ordenados

- Séries temporais
- Dados sequenciais
- ...

Registros

Em uma coleção de registros todos os objetos possuem um conjunto fixo de atributos

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Matrizes

- Se os objetos de dados possuem o mesmo conjunto fixo de atributos numéricos, então eles podem ser considerados pontos em um espaço multidimensional, onde cada dimensão representa um atributo distinto

- Podem ser representados por uma matriz **m** x **n**, onde **m** são as linhas, uma para cada objeto, e **n** são colunas, uma para cada atributo

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Matriz esparsa

- Caso especial de Matriz. Somente não zeros é importante.
- E.g., dados de documento: Cada documento vira um “vetor de termos”.

- Ex: número de vezes que um termo aparece no documento

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Representação conhecida como

Document-term matrix

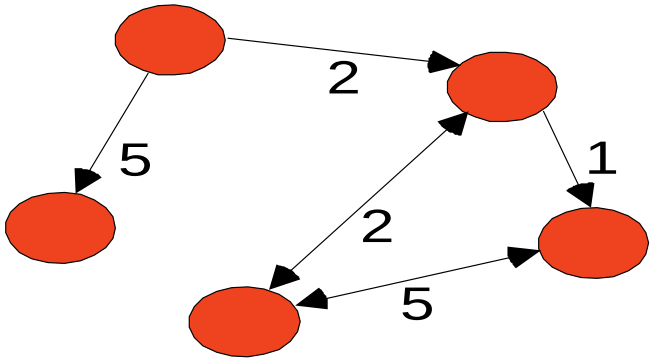
Transações

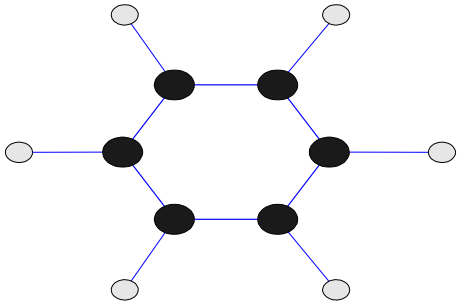
- Um tipo especial de registro, onde:
  - Cada registro (transação) envolve um conjunto de itens
  - Ex: supermercado

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Grafos

- Ex: Interações humanas, uma molécula e páginas Web





Benzene Molecule: C6H6

Useful Links:

- Bibliography
- Other Useful Web sites
  - ACM SIGKDD
  - KDnuggets
  - The Data Mine

Knowledge Discovery and Data Mining Bibliography
(Gets updated frequently, so visit often!)

- Books
- General Data Mining

Book References in Data Mining and Knowledge Discovery

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993. Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support), John Wiley & Sons, 1997.

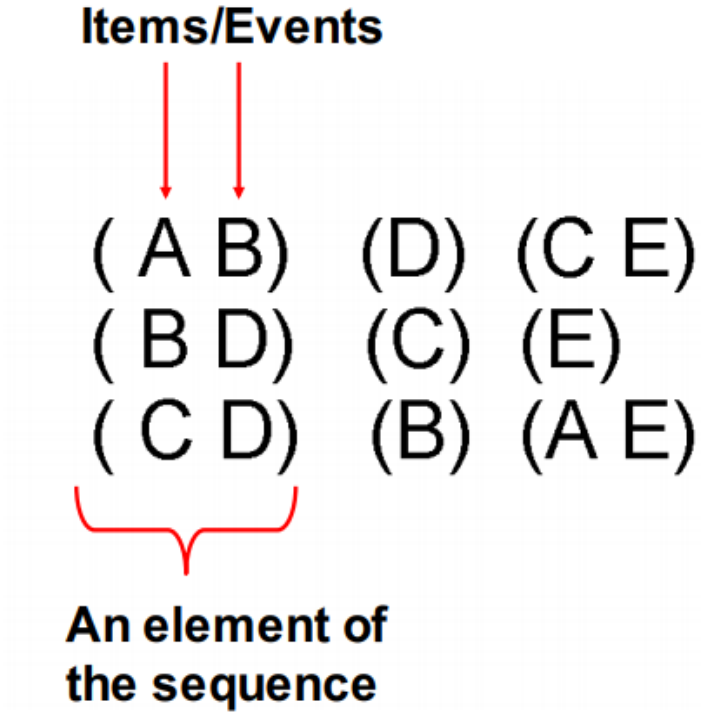
General Data Mining

Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

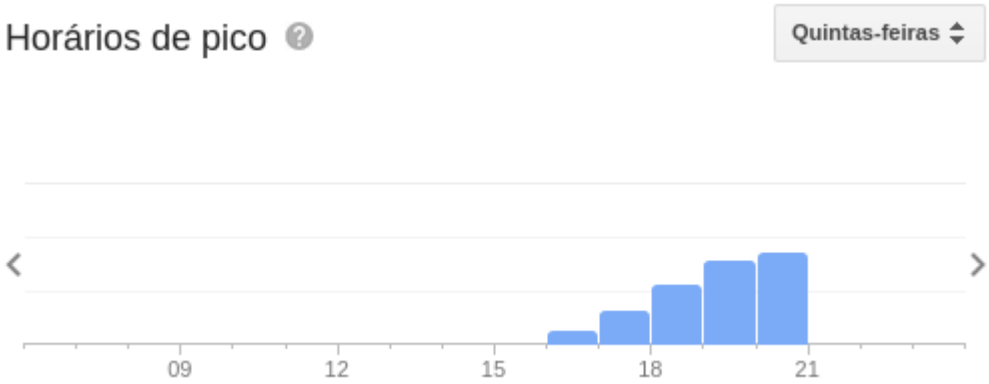
Dados ordenados

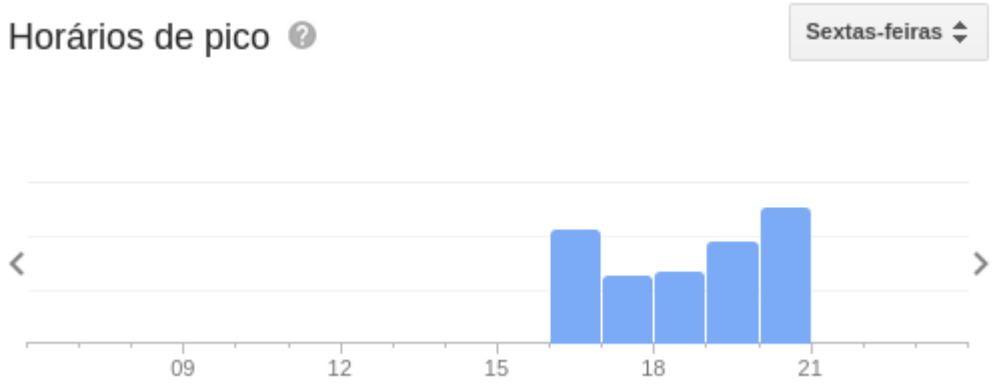
- Sequências de transações, por exemplo.



Dados ordenados

- Série temporal
- Bar do Didi - Curitiba





Qualidade dos dados

- Dados com baixa qualidade afetam negativamente muitos esforços de processamento de dados

- **Ex:** um modelo de classificação para detectar quem é um risco para conceder empréstimos com dados de baixa qualidade
  - Alguns bons candidatos possuem crédito negado
  - Mais crédito é concedido para indivíduos padrão

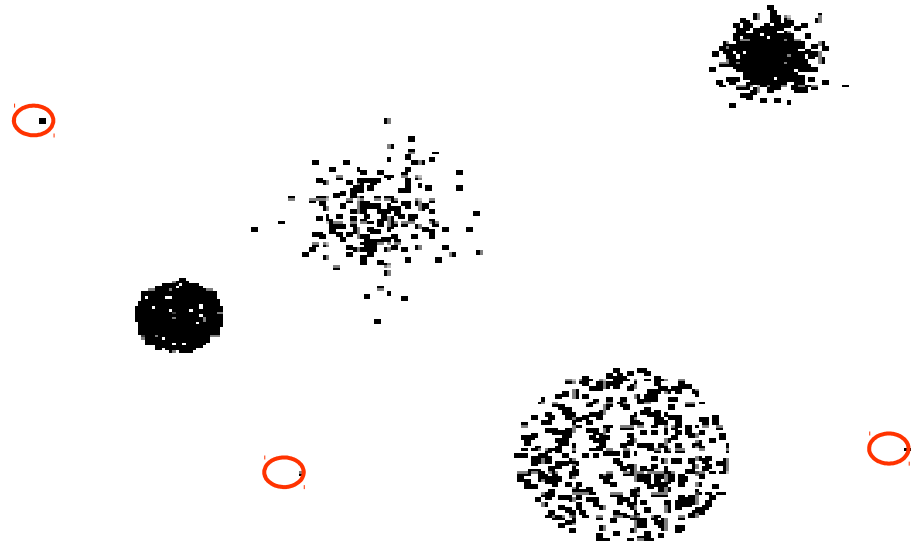
Exemplos de problema na qualidade dos dados:

Ruído e *outliers*
Dados faltantes
Dados duplicados
Dados errados

Outliers

Outliers são objetos de dados com características que são consideravelmente diferentes da maioria dos outros objetos

- **Caso 1:** Outliers são ruído que interferem nas análises
- **Caso 2:** Outliers são o objetivo das análises
  - Fraude de cartão de crédito
  - Detecção de intrusos
- Causas?



Dados faltantes

- Razões para dados faltantes:
  - Informação não é coletada: e.g., pessoa se recusa a fornecer a idade
  - Atributos podem não ser aplicados a todos os casos: e.g., renda anual não é aplicável para crianças
- Trabalhando com dados faltantes
  - Eliminar objetos de dados ou atributos
  - Estimar dados faltantes
    - Ex: séries temporais de temperatura
    - Ex: resultados do censo
  - Ignorar dados faltantes durante as análises

Dados duplicados

- Datasets podem incluir objetos duplicados (ou parcialmente duplicados)
  - Principalmente ao fundir dados de fontes diferentes

- Ex:
  - A mesma pessoa com o mesmo ID em diferentes redes.

Reformatação e Limpeza

Estas tarefas tendem a ser demoradas e tediosas

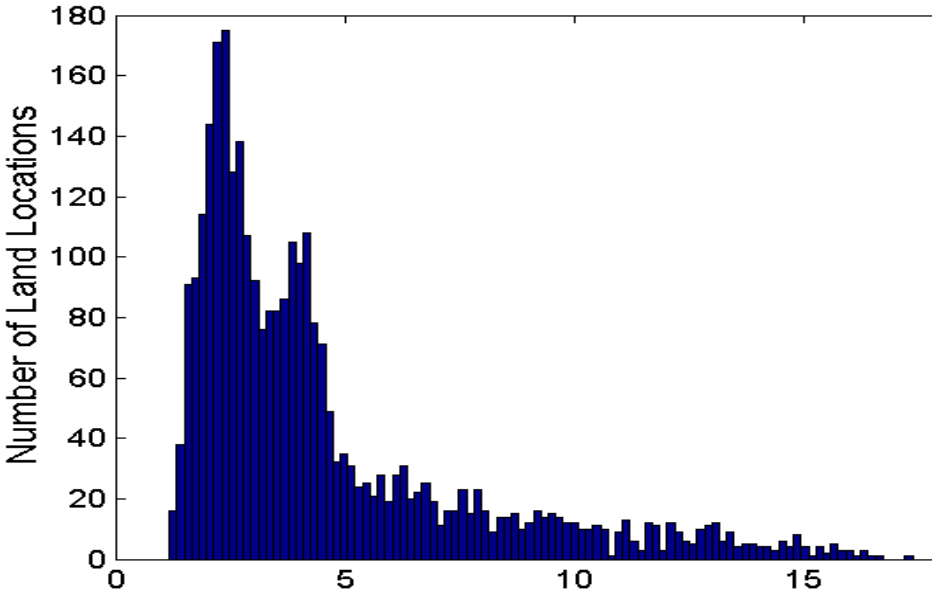
CrowdFlower
2015 Data Scientist Report

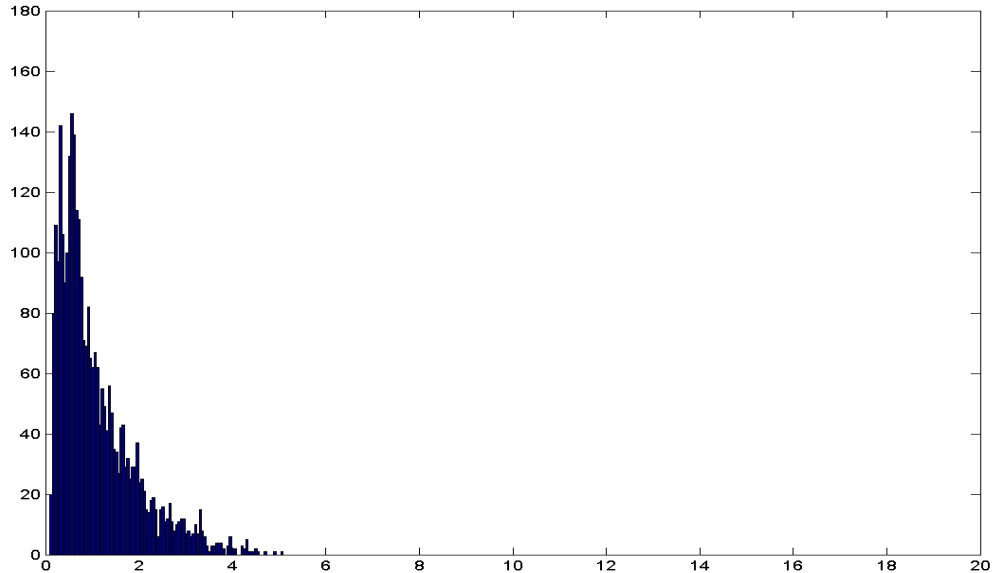
Agregação

- Combina 2 ou mais atributos em um único atributo
- **Objetivos**
  - Redução de dados
  - Reduzir o número de atributos
  - Mudar a escala
    - Cidades agregadas em regiões ou estados, etc.
    - Dias em semanas, meses ou anos
  - Dados mais “estáveis”
    - Dados agregados tendem a ter menos variabilidade

Agregação

Variação da precipitação na Austrália (grid cells)



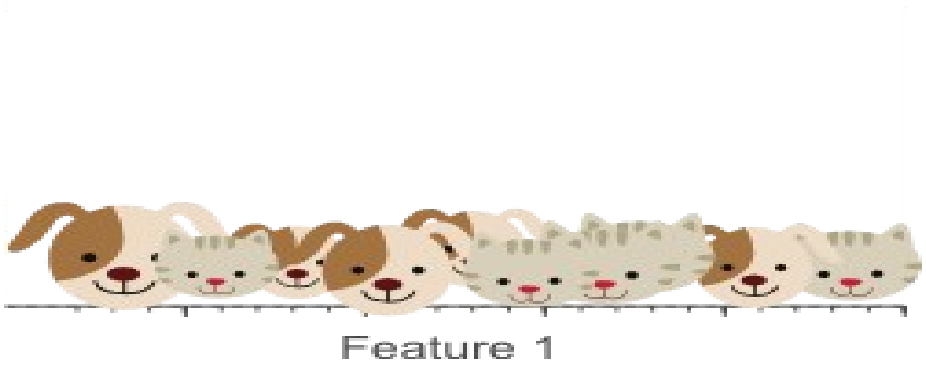


Desvio padrão médio da precipitação por mês

Desvio padrão médio da precipitação por ano

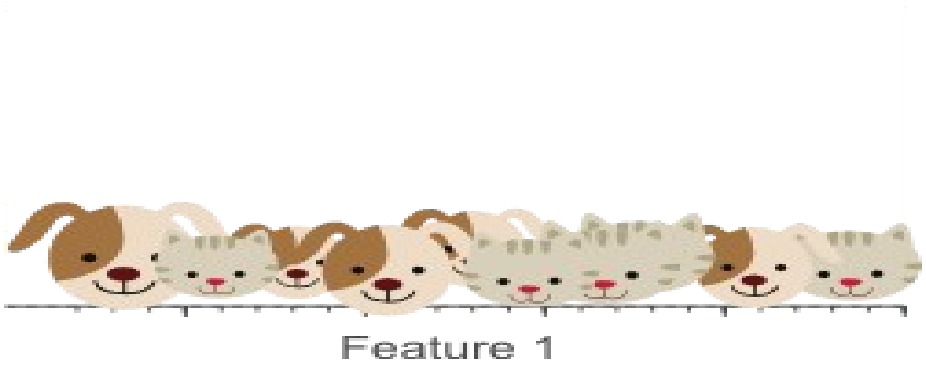
A precipitação média por ano varia menos

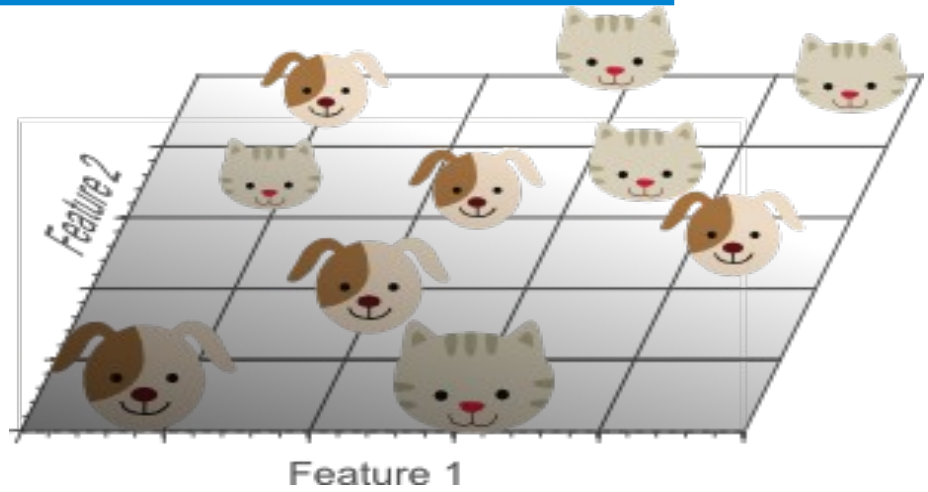
Importância das dimensões



Imagens de Vincent Spruyt

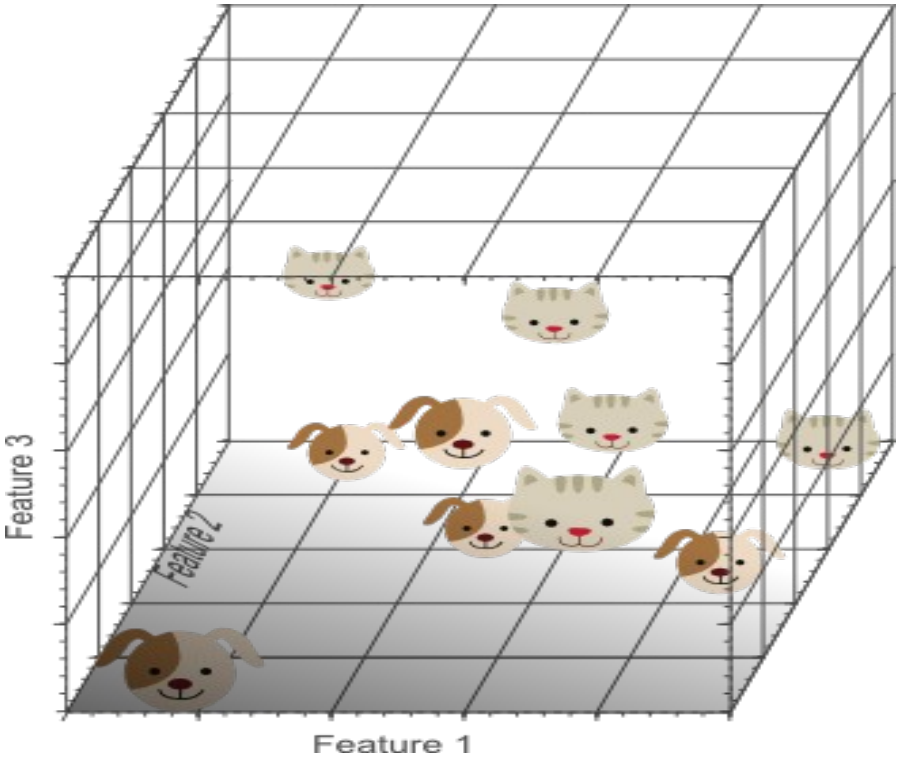
Importância das dimensões





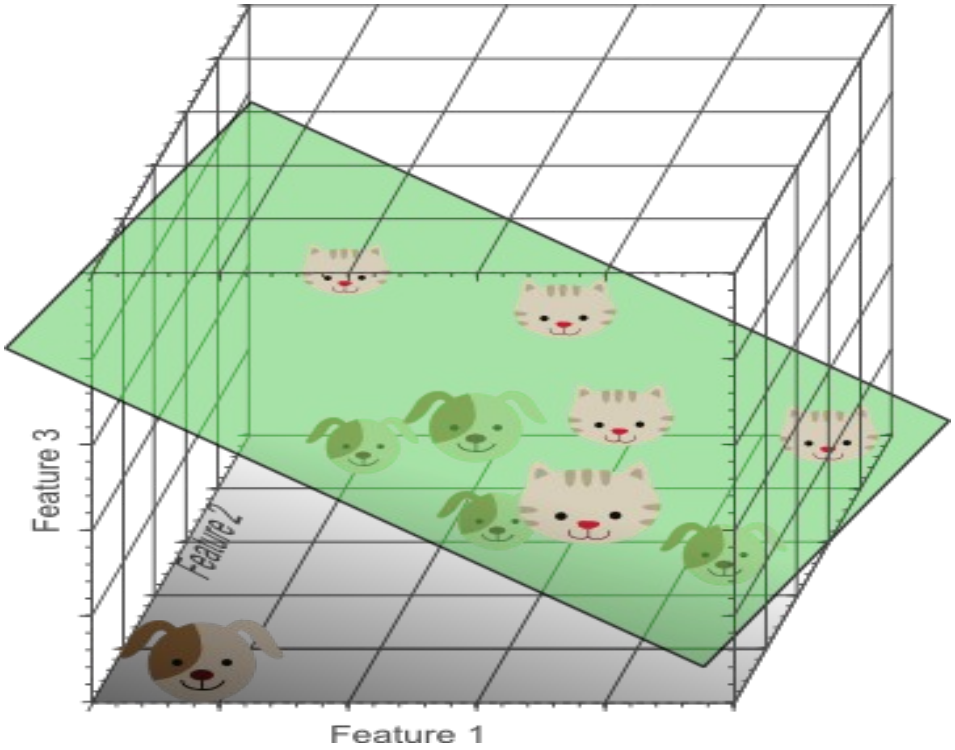
Imagens de Vincent Spruyt

Importância das dimensões



Imagens de Vincent Spruyt

Importância das dimensões



Imagens de Vincent Spruyt

Criação de features

Objetivo: capturar informações importantes no dataset mais eficientemente do que com os atributos originais

- Três metodologias gerais:
  - Extração de features
  - Ex: bordas de imagens
  - Construção de features
  - Ex: dividir a massa pelo volume para obter a densidade
  - Mapear dados em um novo espaço
    - Ex: Análises de Fourier e wavelet

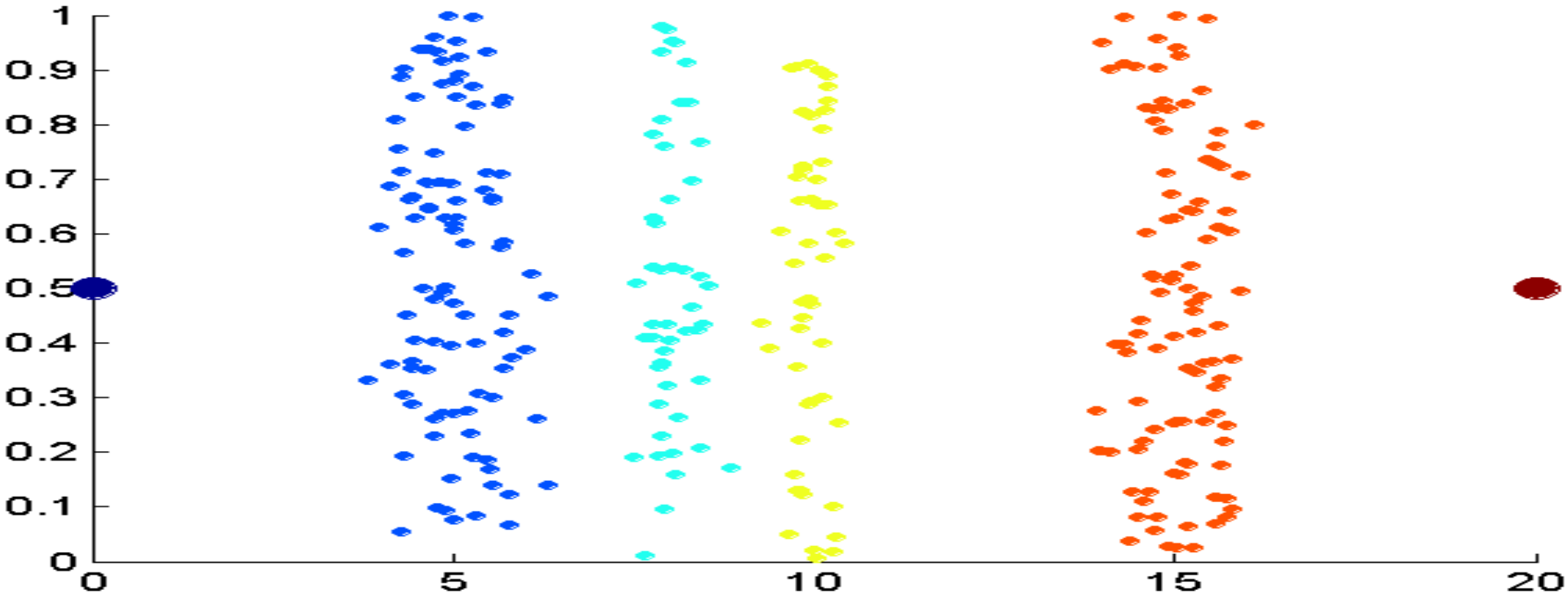
Discretização

Processo de converter um atributo contínuo em um atributo categórico (ordinal ou nominal)

Potencialmente, um número infinito de valores numéricos são mapeados em um número menor de categorias

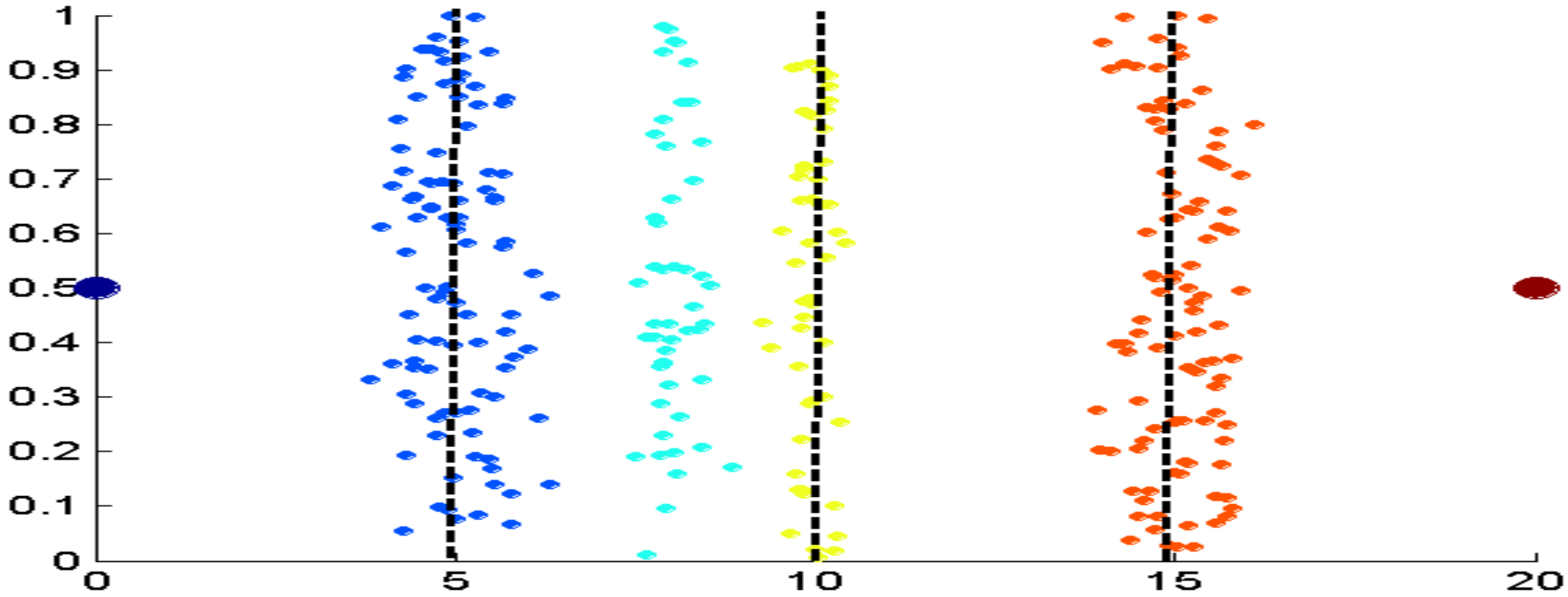
- Vários algoritmos de classificação funcionam melhor se variáveis possuem menos valores

Discretização



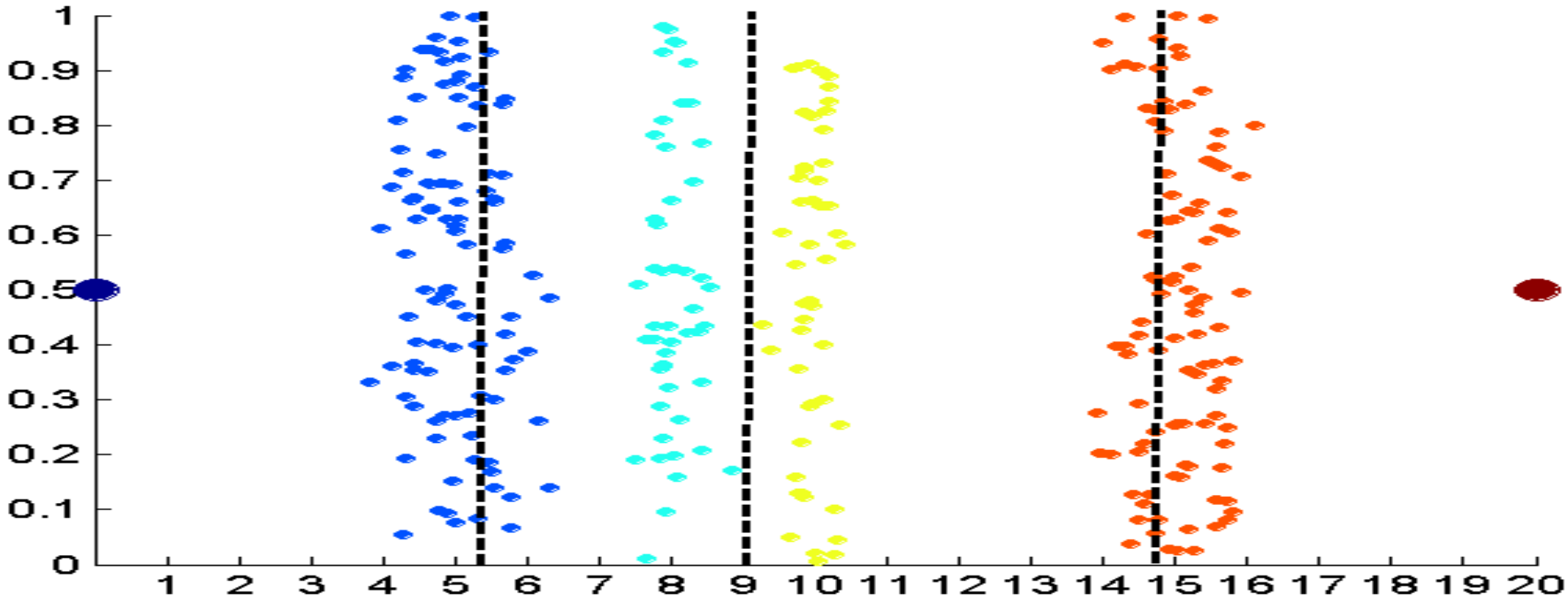
Dados consistem de quatro grupos de pontos e dois outliers

Discretização



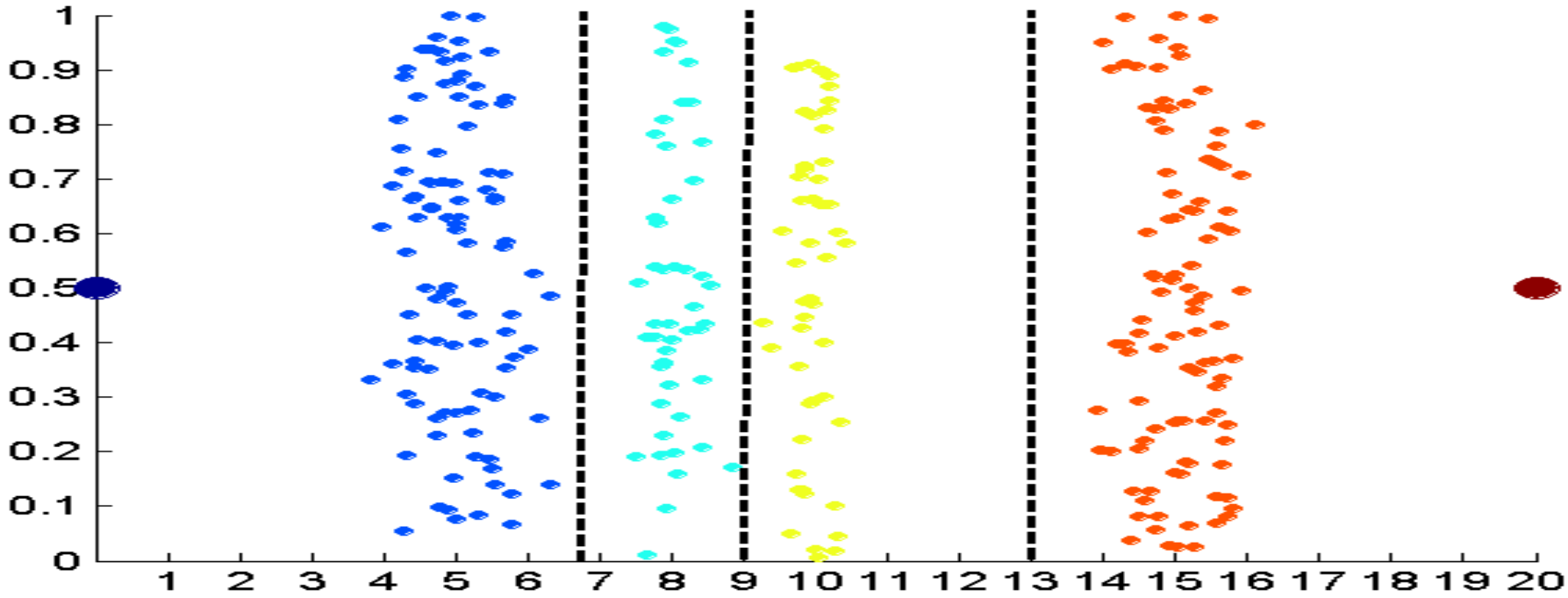
Intervalos de largura iguais

Discretização



Igual frequência

Discretização



Agrupamento

Binarização

- Mapeia um atributo contínuo ou categórico em uma ou mais variáveis binárias
- Usado bastante para análises de associação
- Geralmente converte um atributo contínuo em categórico e depois em um conjunto binário

Table 2.6. Conversion of a categorical attribute to five asymmetric binary attributes.

Categorical Value	Integer Value	x_1	x_2	x_3	x_4	x_5
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
OK	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

Agradecimentos

Parte deste material é derivado do livro:
Introduction to Data Mining - Tan, Steinbach, Kumar