# Mineração de Dados

## Pré-Processamento de Dados



## Sumário

- Dados
- Escalar
- 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 número particular de valores, com a possibilidade de valores diferentes para cada observação.

3

## Cardinalidade dos atributos das variáveis

- Qualitativo / Quantitativo
  - Variáveis qualitativas: escalas nominais ou ordinais
  - Variáveis quantitativas: escalas intervalares e proporcionais

4

## 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, existe não 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
  - 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


7

## Transição de Escalas

- Os dados de uma escala de razão podem ser transformados em dados intervalares;
- Os dados intervalares podem ser transformados em ordinais e;
- Os dados ordinais podem ser transformados em nominais.
- Naturalmente, tais transições de escala envolvem, necessariamente, uma **perda de informação**.
- Apenas em certas situações muito especiais, com base em procedimentos axiomático-dedutivos, é possível se fazer a trajetória no sentido inverso.

8

## Cardinalidade: Discreto versus Continuo

- 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



## Cardinalidade: Discreto versus Continuo

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

## 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

13

## 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%).

14

## O que define dados de qualidade ?

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

15

## 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

16

## 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.

17

## Limpeza de Dados

- Preencher valores faltantes
- Identificar "*outliers*" e suavizar ruídos
- 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.

18

## 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

19

## 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
  - 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.

20

## Como manipular dados faltantes ?

- Eliminar a observação:
  - excluir os registros 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.
- 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.

21

## 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 uma única tupla ou um conjunto de tuplas. Demanda conhecimento especialista.
  - Um cliente com idade inferior a 21 possui crédito aprovado.

22

## Dados Ruidosos:

Métodos para tratar os dados ruidosos:

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

23

## 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

24

## Compartimentalização

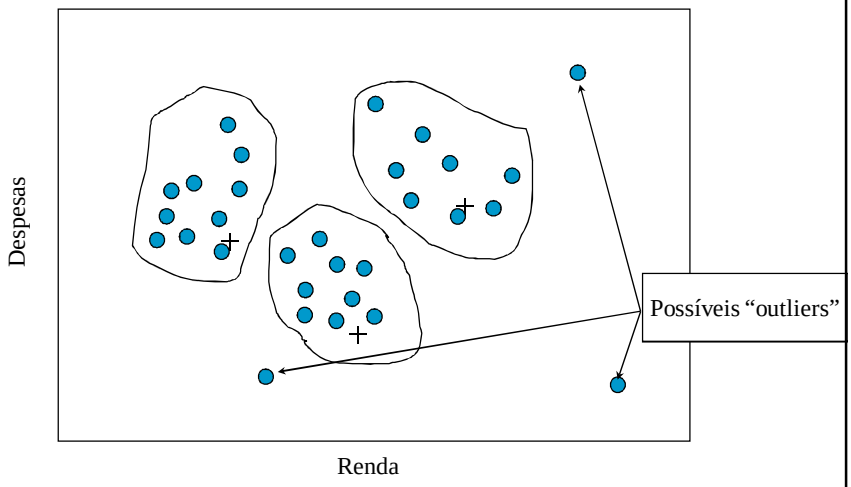
- Particionamento por Distância
  - Divide os dados em N intervalos de mesmo tamanho
  - Se A e B são os valores mínimo e máximo do atributo, a largura dos W intervalos será W=(B-A)/N
  - Esta técnica é a mais direta mas pode ser prejudicada pela presença de “outliers”
- Particionamento por freqüência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



## Compartimentalização

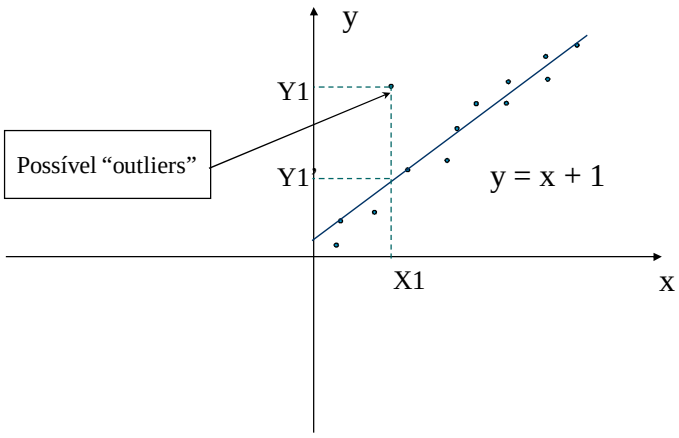
- Exemplo
  - Preço de um produto
    - {4,8,9,15,21,21,24,25,26,28,29,34}
  - Particionamento por freqüência
    - Bin 1: 4, 8, 9, 15
    - Bin 2: 21, 21, 24, 25
    - Bin 3: 26, 28, 29, 34
  - Suavização pela média
    - Bin 1: 9, 9, 9, 9
    - Bin 2: 23, 23, 23, 23
    - Bin 3: 29, 29, 29, 29

# Análise de Cluster




27

# Regressão Linear




28

## 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 Max(| v' |)<1



# Normalização Min-Max

| CPF Cliente | Despesa |
|---|---|
| 999999999999 | 1000 |
| 111111111111 | 2000 |
| 222222222222 | 4000 |

Min=1000
Max=4000

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

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

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

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

| CPF Cliente | Despesa |
|---|---|
| 999999999999 | 0 |
| 111111111111 | 0.333333 |
| 222222222222 | 1 |

# Normalização Z-Score

| CPF Cliente | Despesa |
|---|---|
| 999999999999 | 1000 |
| 111111111111 | 2000 |
| 222222222222 | 4000 |

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

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

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

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

| CPF Cliente | Despesa |
|---|---|
| 999999999999 | -08729 |
| 111111111111 | -0.2182 |
| 222222222222 | 1.0911 |

37

# Escalonamento Decimal

| CPF Cliente | Despesa |
|---|---|
| 999999999999 | 1000 |
| 111111111111 | 2000 |
| 222222222222 | 4000 |

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

| CPF Cliente | Despesa |
|---|---|
| 999999999999 | 0.1 |
| 111111111111 | 0.2 |
| 222222222222 | 0.4 |

38

## 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

39

## Redução

- Seleciona um mínimo de atributos que mantenha as distribuições de probabilidade semelhantes as originais
- Em data mining a supressão de uma coluna (atributo) é muito mais delicada do que a supressão de uma linha (observação)
- Retirar atributos irrelevantes ou permanecer com atributos relevantes
- Pode implicar na descoberta de padrões de baixa qualidade
  - Daí a necessidade de um estágio de seleção de atributos
- Uma abordagem para a seleção é a manual, baseada em conhecimento especialista

40

## Redução: Estratégias

- Bancos de dados muito grandes podem tornar o processo de mineração lento
- Redução de dados
  - Obtém uma representação reduzida do conjunto de dados com menor volume e resultados similares
- Estratégias de redução
  - Redução de dimensionalidade
  - Compressão de dados
  - Redução de casos
  - Discretização



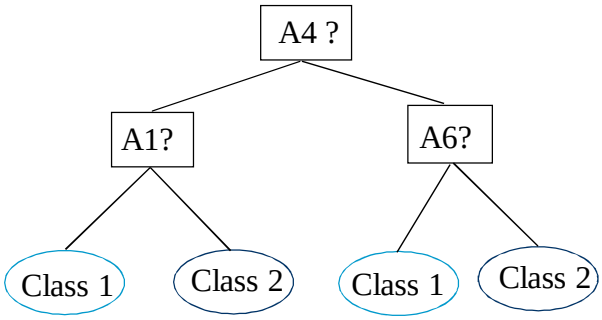
## Redução: de Dimensionalidade

- Métodos Heurísticos
  - Indução por Árvore de Decisão
  - Melhores Atributos Individuais
  - Seleção forward
  - Eliminação backward



## Redução: de Dimensionalidade - Indução por Árvore de Decisão

- Conjunto Inicial de Atributos:
- {A1, A2, A3, A4, A5, A6}




-----> Reduced attribute set: {A1, A4, A6}



## Redução: de Dimensionalidade - Seleção por Heurísticas

- **seleção forward** : a busca é iniciada sem atributos e os mesmos são adicionados um a um. Cada atributo é adicionado isoladamente e o conjunto resultante é avaliado segundo um critério. O atributo que produz o melhor critério é incorporado
- **eliminação backward** : a busca é iniciada com o conjunto completo de atributos e os mesmos são suprimidos um de cada vez. Cada atributo é suprimido isoladamente e o conjunto resultante é avaliado segundo um critério. O atributo que produz o melhor critério é finalmente suprimido

## Redução: Compressão de Dados- Extração de Variáveis

- Objetivo:
  - obter novas variáveis à partir dos atributos iniciais. Em geral as novas variáveis são combinações lineares das variáveis iniciais
  - Limitações: modelo linear (não adequado especialmente para os métodos de data mining baseados em lógica)
- As técnicas de redução de dimensões se propõem a reduzir o número de variáveis com a menor perda possível de informações
- Essas técnicas são úteis também para tratar a redundância de informações (correlação entre variáveis) e ruído

45

## Redução: Compressão de Dados- Extração de Variáveis

- Famílias de Métodos
  - Métodos não supervisionados
  - Métodos supervisionados
- Métodos não supervisionados: Análise de Componentes Principais (variáveis quantitativas) e Análise de Correspondências (variáveis qualitativas)
- Métodos supervisionados: Análise Fatorial

46

## Redução: Compressão de Dados- Extração de Variáveis

- Método supervisionado:
  - Análise Fatorial
- É o termo genérico de uma técnica multivariada, cujo propósito é a redução de dados e sumarização.
- Ela analisa as relações entre variáveis e tenta explicá-las em termos de suas dimensões subjacentes comuns (fatores).

47

## Redução: Compressão de Dados- Extração de Variáveis

- Análise Fatorial
- Os fatores podem ser extraídos como ortogonais ou oblíquos.
  - ortogonais, eles serão independentes entre si e;
  - oblíquos serão correlacionados ou dependentes.
- Fatores ortogonais representam redução de informação, mas podem não ter sentido real.

48

## Redução: Compressão de Dados- Extração de Variáveis

- Análise Fatorial
- A primeira componente é a combinação linear das variáveis iniciais que melhor separa os grupos entre si, isto é, ela toma valores os mais próximos possíveis para os indivíduos de um mesmo grupo e os mais diferentes para indivíduos de grupos distintos.
- A segunda componente é a combinação linear das variáveis iniciais ortogonal a primeira (correlação nula) que melhor separa os grupos entre si. E assim por diante.



## Análise de Componentes Principais (variáveis quantitativas)

A análise de componentes principais é indicada para conjuntos de medidas correlacionadas linearmente, que assim podem ser reduzidas a poucas variáveis sintéticas, denominadas componentes principais (Pielou 1984, Manly 1994).



## Análise de Componentes Principais (variáveis quantitativas)

- É indicada para conjuntos de medidas correlacionadas linearmente, que assim podem ser reduzidas a poucas variáveis sintéticas, denominadas componentes principais
- Ela permite transformar um conjunto de variáveis originais, inter-correlacionadas, num novo conjunto de variáveis não correlacionadas, as componentes principais.
- O objetivo mais imediato é verificar se existe um pequeno número das primeiras componentes principais que seja responsável por explicar uma proporção elevada da variação total associada ao conjunto original.

51

## Análise de Componentes Principais (variáveis quantitativas)

- As vantagens são que ao se descorrelacionar os dados, se está eliminando parte da informação redundante em cada dimensão.
  - os dados podem ser descritos de uma forma mais concisa;
  - certas características escondidas dos dados podem vir à luz depois de transformadas;
  - a distribuição dos dados pode ser representada (aproximadamente) pelas densidades individuais de cada dimensão.

52

## Análise de Correspondências

- É um método de análise fatorial para variáveis categóricas.
- É uma técnica exploratória para estudar a relação entre duas variáveis categóricas (nominais).
- Hierarquizar a informação disponível por ordem decrescente e de acordo com o grau de explicação do fenômeno em estudo;
- Para obter uma representação gráfica da natureza das relações existentes entre as variáveis, ao colocar as categorias semelhantes próximas umas das outras.

53

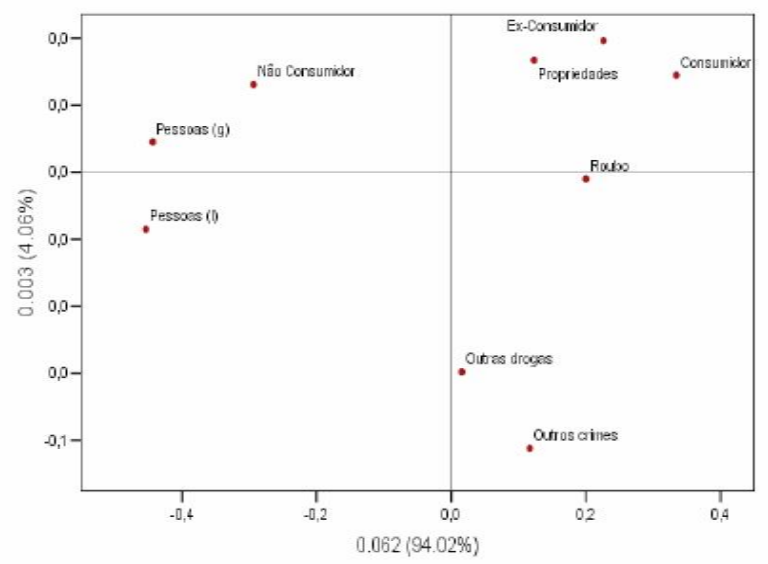
## Análise de Correspondências

Crimes nos EUA nos anos 60

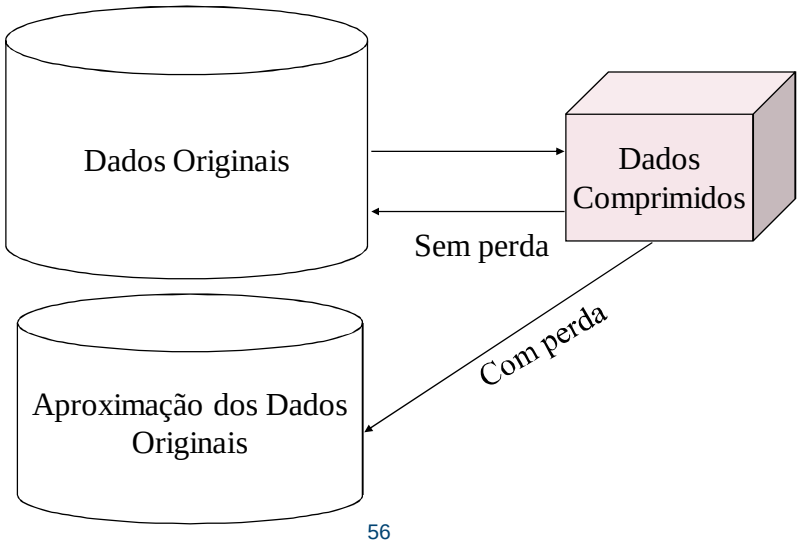
| | *Tipo de Crime* | | | | |
|---|---|---|---|---|---|
| ***Consumo de Heroína*** | **Pessoas (grave)** | **Roubo** | **Pessoas (leve)** | **Propriedades** | **Outros Crimes** |
| **Consumidor** | 30 | 94 | 14 | 237 | 86 |
| **Ex-Consumidor** | 14 | 20 | 5 | 75 | 27 |
| **Outras Drogas** | 93 | 94 | 46 | 253 | 124 |
| **Não Consumidor** | 163 | 79 | 77 | 256 | 93 |

54

## Análise de Correspondências




## Redução: Compressão de Dados



## Redução: Redução de Casos

- Métodos Paramétricos
  - Assume que os dados podem ser representados por um modelo, armazena os parâmetros do modelo e descarta os dados
    - Principais modelos: regressão (simples e múltipla) e modelo log-linear
- Métodos Não-Paramétricos
  - Não usa modelos
  - Histogramas, clusterização, amostragem

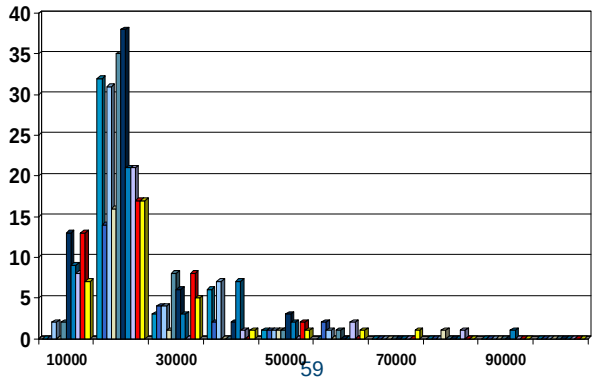


## Redução: Redução de Casos
## Regressão e modelos log-linear

- Regressão linear: os dados são modelados para se ajustarem a uma linha reta
  - Em geral usa o método dos mínimos quadrados para ajustar a linha
- Regressão múltipla: permite que uma variável resposta seja modelada como uma função linear de um vetor de atributos
- Modelo Log-linear : aproxima distribuições de probabilidade discretas multidimensionais

## Redução: Redução de Casos Histogramas

- Divide os dados em blocos e guarda a média de cada bloco
- representação gráfica da distribuição de cada variável em intervalos de freqüência




59

## Clusterização

- Particiona dados em clusters, e armazena apenas a representação do cluster
- Eficiente se os dados podem ser classificados e não estão muito “espalhados”
- Existem diversos algoritmos de clusterização

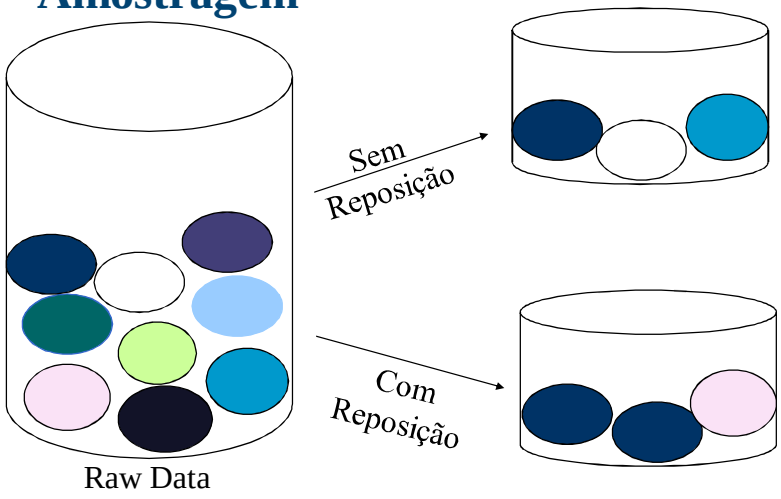
60

## Redução: Redução de Casos Amostragem

- Escolhe um subconjunto de dados representativo
  - Amostragem aleatória simples sem reposição
  - Amostragem aleatória simples com reposição
  - Amostragem de clusters
  - Amostragem estratificada
    - Exemplo: amostragem por faixa etária

61

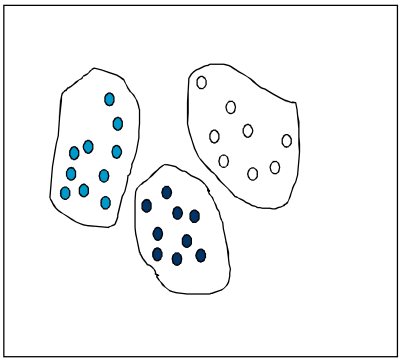
## Redução : Redução de Casos Amostragem



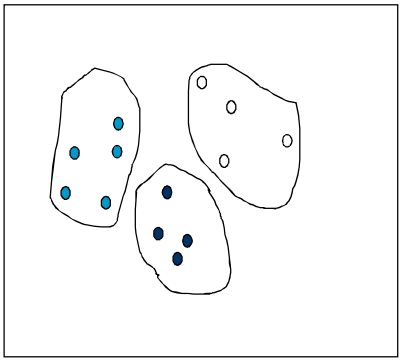

62

## Redução de Casos: Redução de Casos Amostragem

Dados Brutos



Amostra Estratificada



63

## Redução: Redução de Casos Amostragem

- Amostragem incremental
  - O treinamento é realizado em amostras aleatórias cada vez maiores de casos, observar a tendência e parar quando não há mais progresso
  - Um padrão típico de tamanhos de amostras pode ser 10%, 20%, 33%, 50%, 67% e 100%
- Critérios para passar para uma outra amostra
  - O erro diminuiu? A complexidade do tratamento aumentou mais do que a queda da taxa de erro?
  - A complexidade da solução atual é aceitável para a interpretação?

64

## Redução: Redução de Casos Amostragem seguida de voto

- Interesse: quando o método de mineração suporta apenas N casos
- O mesmo método de mineração é aplicado para diferentes amostras de mesmo tamanho resultando em uma solução para cada amostra
- Quando um novo caso aparece, cada solução fornece uma resposta
- A reposta final é obtida por votação (classificação) ou pela média (regressão)



## 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



## Discretização

- Três tipo de atributos
  - Nominais – não ordenável
  - Ordinais – ordenável
  - Contínuos – números reais
- Discretização
  - Divide a faixa de atributos contínuos em intervalos
  - Alguns algoritmos de classificação só aceitam atributos categóricos
  - Reduz o tamanho dos dados por discretização
  - Exemplo: Arredondamento de números
- Hierarquia de Valores
- Hierarquia de Atributos



## Discretização de Variáveis

- Transforma atributos contínuos em atributos categóricos.
- Absolutamente essencial se o método inteligente só manuseia atributos categóricos
- Em alguns casos, mesmo métodos que manuseiam atributos contínuos têm melhor desempenho com atributos categóricos.



## Discretização de Variáveis

- Discretização Supervisionada
  - O método 1R
  - Considera a variável de saída (classe) na discretização
  - Apesar de simples apresenta resultados similares a arvores de decisão.
- Métodos Não Supervisionados consideram somente o atributo a ser discretizado
  - São a única opção no caso de problemas de agrupamento (clustering), onde não se conhecem as classes de saída



## Discretização de Variáveis

- Discretização pelo Método 1R (1-rule)
- Sub-produto de uma técnica de extração automática de regras
- Utiliza as classes de saída para discretizar cada atributo de entrada separadamente
- Ex: Base de dados hipotética de meteorologia x decisão de realizar ou não um certo jogo



## Discretização:
## 1R: aprende uma regra por atributo

- atribuí a classe mais freqüente
- taxa de erro: proporção de instâncias que não pertence a classe majoritária
- escolhe o atributo com menor taxa de erro



## Discretização:
## Pseudo-codigo para 1R

- Para cada atributo
  - para cada valor do atributo, faça uma regra como:
    - conte a freqüência de cada classe
    - encontre a classe mais freqüente
    - atribua a classe mais freqüente a esta regra
    - calcule a taxa de erro da regra
  - escolha a regra com taxa de erro menor



# Discretização: 1R

## Base de Dados Meteorológicos

| Tempo | Temperatura | Umidade | Vento | Jogar? (CLASSE) |
|---|---|---|---|---|
| Sol | 85 | 85 | Não | Não |
| Sol | 80 | 90 | Sim | Não |
| Nublado | 83 | 86 | Não | Sim |
| Chuva | 70 | 96 | Não | Sim |
| Chuva | 68 | 80 | Não | Sim |
| Chuva | 65 | 70 | Sim | Não |
| Nublado | 64 | 65 | Sim | Sim |
| Sol | 72 | 95 | Não | Não |
| Sol | 69 | 70 | Não | Sim |
| Chuva | 75 | 80 | Não | Sim |
| Sol | 75 | 70 | Sim | Sim |
| Nublado | 72 | 90 | Sim | Sim |
| Nublado | 81 | 75 | Não | Sim |
| Chuva | 71 | 91 | Sim | Não |

73

# Discretização: 1R

## Primeiro passo: ordenar pela coluna Temperatura

| Tempo | Temperatura | Umidade | Vento | Jogar? (CLASSE) |
|---|---|---|---|---|
| Nublado | 64 | 65 | Sim | Sim |
| Chuva | 65 | 70 | Sim | Não |
| Chuva | 68 | 80 | Não | Sim |
| Sol | 69 | 70 | Não | Sim |
| Chuva | 70 | 96 | Não | Sim |
| Chuva | 71 | 91 | Sim | Não |
| Sol | 72 | 95 | Não | Não |
| Nublado | 72 | 90 | Sim | Sim |
| Chuva | 75 | 80 | Não | Sim |
| Sol | 75 | 70 | Sim | Sim |
| Sol | 80 | 90 | Sim | Não |
| Nublado | 81 | 75 | Não | Sim |
| Nublado | 83 | 86 | Não | Sim |
| Sol | 85 | 85 | Não | Não |

74

## Discretização: 1R

Segundo passo: discretizar pela Classe de saída

| Tempo | Temperatura | Umidade | Vento | Jogar? (CLASSE) |
|---|---|---|---|---|
| Nublado | 64 | 65 | Sim | Sim |
| Chuva | 65 | 70 | Sim | Não |
| Chuva | 68 | 80 | Não | Sim |
| Sol | 69 | 70 | Não | Sim |
| Chuva | 70 | 96 | Não | Sim |
| Chuva | 71 | 91 | Sim | Não |
| Sol | 72 | 95 | Não | Não |
| Nublado | 72 | 90 | Sim | Sim |
| Chuva | 75 | 80 | Não | Sim |
| Sol | 75 | 70 | Sim | Sim |
| Sol | 80 | 90 | Sim | Não |
| Nublado | 81 | 75 | Não | Sim |
| Nublado | 83 | 86 | Não | Sim |
| Sol | 85 | 85 | Não | Não |

75

## Discretização: 1R

Segundo passo: discretizar pela Classe de saída

| Tempo | Temperatura | Umidade | Vento | Jogar? (CLASSE) |
|---|---|---|---|---|
| Nublado | 64 | 65 | Sim | Sim |
| Chuva | 65 | 70 | Sim | Não |
| Chuva | 68 | 80 | Não | Sim |
| Sol | 69 | 70 | Não | Sim |
| Chuva | 70 | 96 | Não | Sim |
| Chuva | 71 | 91 | Sim | Não |
| Sol | 72 | 95 | Não | Não |
| Nublado | 72 | 90 | Sim | Sim |
| Chuva | 75 | 80 | Não | Sim |
| Sol | 75 | 70 | Sim | Sim |
| Sol | 80 | 90 | Sim | Não |
| Nublado | 81 | 75 | Não | Sim |
| Nublado | 83 | 86 | Não | Sim |
| Sol | 85 | 85 | Não | Não |

76

# Discretização: 1R

## Terceiro passo: ajustar divisões

| Tempo | Temperatura | Umidade | Vento | Jogar? (CLASSE) |
|---|---|---|---|---|
| Nublado | 64 | 65 | Sim | Sim |
| Chuva | 65 | 70 | Sim | Não |
| Chuva | 68 | 80 | Não | Sim |
| Sol | 69 | 70 | Não | Sim |
| Chuva | 70 | 96 | Não | Sim |
| Chuva | 71 | 91 | Sim | Não |
| Sol | 72 | 95 | Não | Não |
| Nublado | 72 | 90 | Sim | Sim |
| Chuva | 75 | 80 | Não | Sim |
| Sol | 75 | 70 | Sim | Sim |
| Sol | 80 | 90 | Sim | Não |
| Nublado | 81 | 75 | Não | Sim |
| Nublado | 83 | 86 | Não | Sim |
| Sol | 85 | 85 | Não | Não |

77

# Discretização: 1R

## Terceiro passo: ajustar divisões

| Tempo | Temperatura | Umidade | Vento | Jogar? (CLASSE) |
|---|---|---|---|---|
| Nublado | (1) 64 | 65 | Sim | Sim |
| Chuva | 65 (2) | 70 | Sim | Não |
| Chuva | 68 | 80 | Não | Sim |
| Sol | (3) 69 | 70 | Não | Sim |
| Chuva | 70 | 96 | Não | Sim |
| Chuva | 71 | 91 | Sim | Não |
| Sol | 72 (4) | MUITAS DIVISÕES! | | Não |
| Nublado | 72 | | | Sim |
| Chuva | (5) 75 | 80 | Não | Sim |
| Sol | 75 | 70 | Sim | Sim |
| Sol | 80 (6) | 90 | Sim | Não |
| Nublado | (7) 81 | 75 | Não | Sim |
| Nublado | 83 | 86 | Não | Sim |
| Sol | 85 (8) | 85 | Não | Não |

78

# Discretização: 1R

Quarto passo: mínimo de valores da maior classe (ex: 3)

| Tempo | Temperatura | Umidade | Vento | Jogar? (CLASSE) |
|---|---|---|---|---|
| Nublado | 64 | 65 | Sim | Sim |
| Chuva | 65 | 70 | Sim | Não |
| Chuva | 68 | 80 | Não | Sim |
| Sol | 69 | 70 | Não | Sim |
| Chuva | 70 | 96 | Não | Sim |
| Chuva | 71 | 91 | Sim | Não |
| Sol | 72 | 95 | Não | Não |
| Nublado | 72 | 90 | Sim | Sim |
| Chuva | 75 | 80 | Não | Sim |
| Sol | 75 | 70 | Sim | Sim |
| Sol | 80 | 90 | Sim | Não |
| Nublado | 81 | 75 | Não | Sim |
| Nublado | 83 | 86 | Não | Sim |
| Sol | 85 | 85 | Não | Não |

79

# Discretização: 1R

Quarto passo: mínimo de valores da maior classe (ex: 3)

| Tempo | Temperatura | Umidade | Vento | Jogar? (CLASSE) |
|---|---|---|---|---|
| Nublado | 64 | 65 | Sim | Sim |
| Chuva | 65 | 70 | Sim | Não |
| Chuva | 68 | 80 | Não | Sim |
| Sol | 69 | 70 | Não | Sim |
| Chuva | 70 | 96 | Não | Sim |
| Chuva | 71 | 91 | Sim | Não |
| Sol | 72 | 95 | Não | Não |
| Nublado | 72 | 90 | Sim | Sim |
| Chuva | 75 | 80 | Não | Sim |
| Sol | 75 | 70 | Sim | Sim |
| Sol | 80 | 90 | Sim | Não |
| Nublado | 81 | 75 | Não | Sim |
| Nublado | 83 | 86 | Não | Sim |
| Sol | 85 | 85 | Não | Não |

80

## Discretização: 1R

Quarto passo: mínimo de valores (ex: 3)

| Tempo | Temperatura | | Umidade | Vento | Jogar? (CLASSE) |
|---|---|---|---|---|---|
| Nublado | 64 | | 65 | Sim | Sim |
| Chuva | 65 | | 70 | Sim | Não |
| Chuva | 68 | 1 | 80 | Não | Sim |
| Sol | 69 | | 70 | Não | Sim |
| Chuva | 70 | | 96 | Não | Sim |
| Chuva | 71 | | 91 | Sim | Não |
| Sol | 72 | | 95 | Não | Não |
| Nublado | 72 | 2 | 90 | Sim | Sim |
| Chuva | 75 | | 80 | Não | Sim |
| Sol | 75 | | 70 | Sim | Sim |
| Sol | 80 | | 90 | Sim | Não |
| Nublado | 81 | 3 | 75 | Não | Sim |
| Nublado | 83 | | 86 | Não | Sim |
| Sol | 85 | | 85 | Não | Não |

81

## Discretização: 1R

Quarto passo: mínimo de valores (ex: 3)

| Tempo | Temperatura | | Umidade | Vento | Jogar? (CLASSE) |
|---|---|---|---|---|---|
| Nublado | 64 | | 65 | Sim | Sim |
| Chuva | 65 | | 70 | Sim | Não |
| Chuva | 68 | 1 | 80 | Não | Sim |
| Sol | 69 | | 70 | Não | Sim |
| Chuva | 70 | | 96 | Não | Sim |
| Chuva | 71 | | 91 | Sim | Não |
| Sol | 72 | | 95 | Não | Não |
| Nublado | 72 | 2 | 90 | Sim | Sim |
| Chuva | 75 | | 80 | Não | Sim |
| Sol | 75 | | 70 | Sim | Sim |
| Sol | 80 | | 90 | Sim | Não |
| Nublado | 81 | 3 | 75 | Não | Sim |
| Nublado | 83 | | 86 | Não | Sim |
| Sol | 85 | | 85 | Não | Não |

82

# Discretização: 1R

## Quarto passo: mínimo de valores (ex: 3)

| Tempo | Temperatura | Umidade | Vento | Jogar? (CLASSE) |
|---|---|---|---|---|
| Nublado | 64 | 65 | Sim | Sim |
| Chuva | 65 | 70 | Sim | Não |
| Chuva | 68 (1) | 80 | Não | Sim |
| Sol | 69 | 70 | Não | Sim |
| Chuva | 70 | 96 | Não | Sim |
| Chuva | 71 | 91 | Sim | Não |
| Sol | 72 | 95 | Não | Não |
| Nublado | 72 | 90 | Sim | Sim |
| Chuva | 75 | 80 | Não | Sim |
| Sol | 75 | 70 | Sim | Sim |
| Sol | 80 | 90 | Sim | Não |
| Nublado | 81 (2) | 75 | Não | Sim |
| Nublado | 83 | 86 | Não | Sim |
| Sol | 85 | 85 | Não | Não |



# Discretização: 1R

## Quarto passo: mínimo de valores (ex: 3)

| Tempo | Temperatura | Umidade | Vento | Jogar? (CLASSE) |
|---|---|---|---|---|
| Nublado | 64 | 65 | Sim | Sim |
| Chuva | 65 | 70 | Sim | Não |
| Chuva | 68 (1) | 80 | Não | Sim |
| Sol | 69 | 70 | Não | Sim |
| Chuva | 70 | 96 | Não | Sim |
| Chuva | 71 | | | |
| Sol | 72 | | | |
| Nublado | 72 | | | |
| Chuva | 75 | | | |
| Sol | 75 | 70 | Sim | Sim |
| Sol | 80 | 90 | Sim | Não |
| Nublado | 81 (2) | 75 | Não | Sim |
| Nublado | 83 | 86 | Não | Sim |
| Sol | 85 | 85 | Não | Não |

Categoria 1: Temperatura $\leq 77.5$
Categoria 2: Temperatura $> 77.5$



## Discretização: 1R

Base de Dados Meteorológicos - Umidade

| Tempo | Temperatura | Umidade | Vento | Jogar? (CLASSE) |
|---|---|---|---|---|
| Nublado | 64 | 65 | Sim | Sim |
| Chuva | 65 | 70 | Sim | Não |
| Sol | 69 | 70 | Não | Sim |
| Sol | 75 | 70 | Sim | Sim |
| Nublado | 81 | 75 | Não | Sim |
| Chuva | 68 | 80 | Não | Sim |
| Chuva | 75 | 80 | Não | Sim |
| Sol | 85 | 85 | Não | Não |
| Nublado | 83 | 86 | Não | Sim |
| Sol | 80 | 90 | Sim | Não |
| Nublado | 72 | 90 | Sim | Sim |
| Chuva | 71 | 91 | Sim | Não |
| Sol | 72 | 95 | Não | Não |
| Chuva | 70 | 96 | Não | Sim |

85

## Métodos de Discretização Não Supervisionada

- Três abordagens básicas:
  - Número pré-determinado de intervalos
    - uniformes (equal-interval binning)
  - Número uniforme de amostras por intervalo
    - (equal-frequency binning)
  - Agrupamento (clustering): intervalos arbitrários

86

## Métodos de Discretização Não Supervisionada

- Número pré-determinado de intervalos uniformes
  - (equal-interval binning)
- No exemplo (temperatura):

  64 65 68 69 70 71 72 72 75 75 80 81 83 85
- Bins com largura 6: $x \leq 60$

  $60 < x \leq 66$
  $66 < x \leq 72$
  $72 < x \leq 78$
  $78 < x \leq 84$
  $84 < x \leq 90$

87

## Métodos de Discretização Não Supervisionada

- Número pré-determinado de intervalos uniformes
  - (equal-interval binning)
- No exemplo (temperatura):

  64 65 68 69 70 71 72 72 75 75 80 81 83 85
- Bins com largura 6: $x \leq 60$: n.a.

  $60 < x \leq 66$: 64, 65
  $66 < x \leq 72$: 68, 69, 70, 71, 72, 72
  $72 < x \leq 78$: 75, 75
  $78 < x \leq 84$: 80, 81, 83
  $84 < x \leq 90$: 85

88

## Métodos de Discretização Não Supervisionada

Equal-interval binning: Problemas

- Como qualquer método não supervisionado, arrisca destruir distinções úteis, devido a divisões muito grandes ou fronteiras inadequadas
- Distribuição das amostras muito irregular, com algumas bins com muitas amostras e outras com poucas amostras



## Métodos de Discretização Não Supervisionada

- Número uniforme de amostras por intervalo
  - (equal-frequency binning)
- Também chamado de equalização do histograma
- Cada bin tem o mesmo número aproximado de amostras
- Histograma é plano
- Heurística para o número de bins: $\sqrt{N}$
- N = número de amostras



## Métodos de Discretização Não Supervisionada

- Número uniforme de amostras por intervalo
  - (equal-frequency binning)
- No exemplo (temperatura):
- 64 65 68 69 |70 71 72 72 |75 75 80| 81 83 85
- 14 amostras: 4 Bins
  - $x \leq 69{,}5$: 64, 65, 68, 69
  - $69{,}5 < x \leq 73{,}5$: 70, 71, 72, 72
  - $73{,}5 < x \leq 80{,}5$: 75, 75, 80
  - $x > 80{,}5$: 81, 83, 85

91

## Hierarquia de Atributos

- Especialista do domínio apresenta hierarquia
- Exemplo
  - Logradouro < Bairro < Cidade < Estado
  - Especialista estabelece nível de corte

92

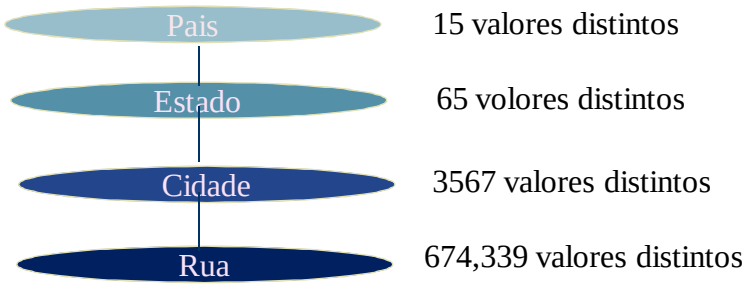
## Hierarquia de Valores

- Também necessita do especialista
- Exemplo
  - {tênis, sapato, sandália} = sapato
  - {bermuda, calça, camisa, paletó} = roupa

93

## Hierarquias de conceitos para dados categóricos

- Hierarquia conceitual pode ser gerada automaticamente com base no número de valores distintos por atributo.
- O atributo com o maior número de valores distintos é colocado no nível mais baixo da hierarquia




94

# Métodos de Discretização Não Supervisionada

- Agrupamento (Clustering)
- Pode-se aplicar um algoritmo de agrupamento
- No caso unidimensional
- Para cada grupo (cluster), atribuir um valor discreto