

Mineração de Dados

Pré-processamento



- 1 Pré-processamento de Dados
- 2 Limpeza dos Dados
- 3 Integração dos Dados
- 4 Redução de Dados
- 5 Transformação dos Dados

Pré-processamento de Dados

Pré-processamento de Dados

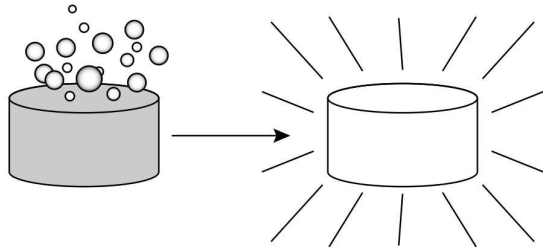
- ▶ Qualidade dos dados
- ▶ Por que os dados devem ser pré-processados?
- ▶ Algumas formas de medir a qualidade dos dados
 - ▶ Acurácia: correto ou errado
 - ▶ Completude: não guardado, indisponível, ...
 - ▶ Consistência: faltam modificações, pendências, ...
 - ▶ Atualidade: podem haver discrepâncias temporais nos dados
 - ▶ Credibilidade: confiança na corretude dos dados
 - ▶ Interpretabilidade: facilidade em entender os dados

Principais Tarefas do Pré-processamento de Dados

- ▶ Limpeza dos dados
 - ▶ Preencher dados faltantes, suavizar dados com ruído, identificar e/ou remover *outliers*, e resolver inconsistências
- ▶ Integração dos dados
 - ▶ Integrar múltiplos bancos de dados, arquivos, ...
- ▶ Redução dos dados
 - ▶ Redução de dimensionalidade
 - ▶ Redução do número de instâncias
 - ▶ Compressão dos dados
- ▶ Transformação dos dados
 - ▶ Discretização
 - ▶ Normalização
 - ▶ Hierarquização: logradouro < cidade < estado < país

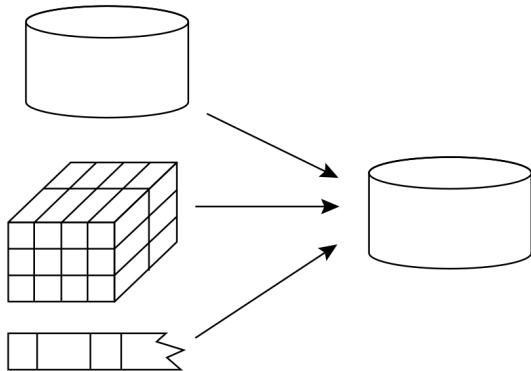
Pré-processamento: Limpeza

Data cleaning



Pré-processamento: Integração

Data integration



Data reduction

Attributes: A1, A2, A3, ..., A126

Transactions: T1, T2, T3, T4, ..., T2000

→

Attributes: A1, A3, ..., A115

Transactions: T1, T4, ..., T1456

Data transformation

$-2, 32, 100, 59, 48 \longrightarrow -0.02, 0.32, 1.00, 0.59, 0.48$

Limpeza dos Dados

Limpeza dos Dados

- ▶ Os dados em situações reais são comumente sujos
 - ▶ incompleto: falta de valores, falta de atributos de interesse, contém apenas dados agregados
 - ▶ ruído: contém ruídos, erros ou *outliers*
 - ▶ inconsistência: discrepância entre valores de atributos distintos (podendo ser “duplicações”)
 - ▶ intencional: valores faltantes disfarçados

Dados Incompletos

- ▶ Nem sempre todos os dados estão disponíveis
 - ▶ algumas instâncias podem não ter algum valor de certos atributos
- ▶ A falta de dados pode vir de
 - ▶ mau funcionamento de um equipamento
 - ▶ inconsistência com outros registros e, assim, foi deletado
 - ▶ um valor não foi guardado por ter sido mal interpretado
 - ▶ certos atributos podem não terem sido considerados importantes em algum momento da coleta dos dados
 - ▶ não haver histórico dos registros ou alteração dos dados
- ▶ Dados faltantes podem ser substituídos por valores inferidos

Como Lidar com os Dados Faltantes?

- ▶ Ignorar a instância
 - ▶ estratégia normalmente adotada quando o “valor esperado” está faltando
 - ▶ não é efetivo quando a porcentagem de dados faltantes é grande
- ▶ Completar os valores faltantes manualmente
 - ▶ tendencioso
 - ▶ pode gerar valores infactíveis
- ▶ Completar o valor automaticamente com
 - ▶ uma constante global: por exemplo, uma classe “desconhecida”
 - ▶ o valor médio daquele atributo
 - ▶ o valor médio daquele atributo para instâncias da mesma classe (alternativa mais esperta)
 - ▶ o valor “mais provável”: por exemplo, inferindo-o com alguma técnica de aprendizado de máquina

Dados com Ruído

- ▶ Ruído: erro aleatório ou variância em uma variável medida
- ▶ Valores incorretos podem vir de
 - ▶ falha dos instrumentos que coletam os dados
 - ▶ problemas na entrada dos dados
 - ▶ problemas de transmissão
 - ▶ limitação tecnológica
 - ▶ inconsistência na convenção dos nomes
- ▶ Outras situações que requerem limpeza nos dados:
 - ▶ dados duplicados
 - ▶ dados incompletos
 - ▶ dados inconsistentes

Como Tratar os Dados com Ruído?

- ▶ Alisamento
 - ▶ Distribuição dos dados ordenados, tendo como referência os seus vizinhos
 - ▶ Ordenação: 1, 1, 2, 3, 3, 3, 4, 5, 5, 7
 - ▶ Particionamento em “caixas”: $\{1, 1, 2\}$, $\{3, 3, 3\}$, $\{4, 5, 5, 7\}$
 - ▶ Alisamento pela mediana: $\{1, 1, 1\}$, $\{3, 3, 3\}$, $\{5, 5, 5, 5\}$
 - ▶ Pode-se usar alternativamente: média, fronteiras, ...
- ▶ Regressão
 - ▶ suaviza os dados ajustando-os por funções de regressão
- ▶ Agrupamento
 - ▶ detecta e remove *outliers*
- ▶ Combinando inspeção por computador e humana
 - ▶ detecta-se valores suspeitos automaticamente e verifica-se por um humano
 - ▶ tratamento de *outliers*

Exemplo: Vizinhança

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15

Bin 2: 21, 21, 24

Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9

Bin 2: 22, 22, 22

Bin 3: 29, 29, 29

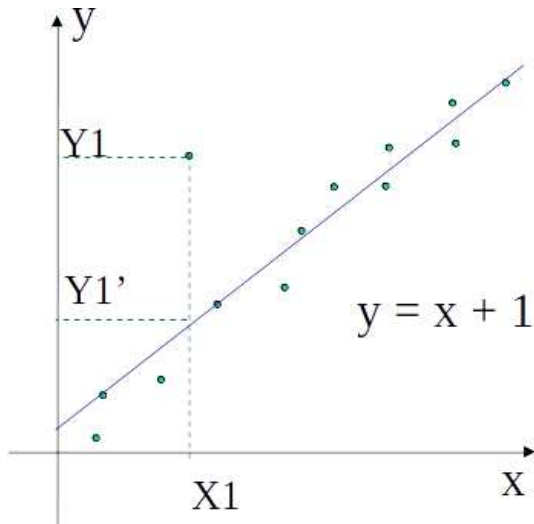
Smoothing by bin boundaries:

Bin 1: 4, 4, 15

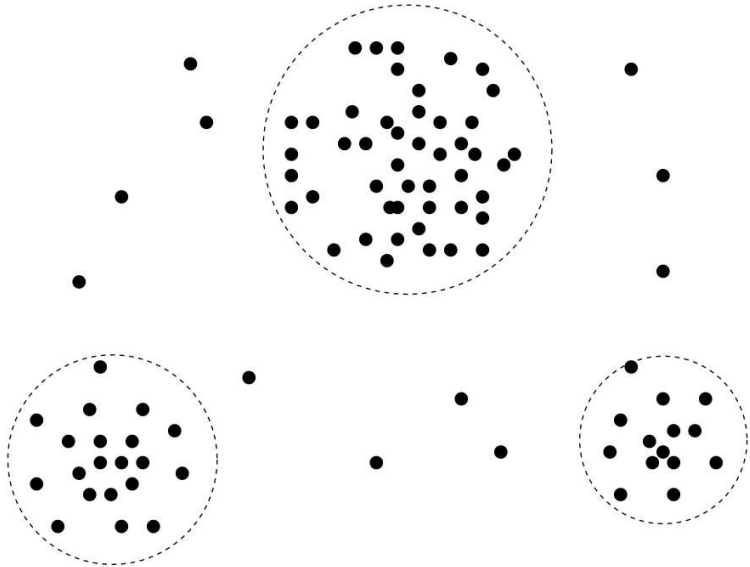
Bin 2: 21, 21, 24

Bin 3: 25, 25, 34

Exemplo: Regressão



Exemplo: Agrupamento



Limpeza dos Dados como um Processo

- ▶ Comum em migração e integração de dados
- ▶ Detecção de discrepância nos dados
 - ▶ Use metadados: domínio, limites de valores, dependências, distribuições,...
 - ▶ Verifique sobrecarga de campo
 - ▶ Verifique regras de unicidade, valores consecutivos, valores nulos
 - ▶ Adote ferramentas: limpando os dados usando ferramentas do domínio dos dados (corretor ortográfico, verificador de CEP, etc) ou sistemas de auditoria (violação de valores)

Integração dos Dados

Integração dos Dados

- ▶ Combina dados de múltiplas fontes
- ▶ Esquema de integração
 - ▶ integra metadados de diferentes fontes
- ▶ Problema de identificar entidades
 - ▶ Pode ser difícil caracterizar entidades de múltiplas fontes
 - ▶ Por exemplo: Bill Gates = Willian Gates
- ▶ Detectando e resolvendo valores conflitantes nos dados
 - ▶ valores diferentes para uma mesma entidade vindos de múltiplas fontes
 - ▶ razões possíveis: representação, escala, métrica, etc

Tratando Redundância na Integração dos Dados

- ▶ Redundância pode ocorrer quando integrando múltiplos banco de dados
 - ▶ identificação de objetos: o mesmo atributo ou objeto pode ter nomes diferentes
 - ▶ dado derivável: um atributo pode ser uma derivação de um atributo em outra tabela
- ▶ Atributos redundantes podem ser detectados usando análise de correlação e covariância
- ▶ Integrar cuidadosamente dados de múltiplas fontes pode ajudar a reduzir/evitar redundâncias e inconsistências

Análise de Correlação (Dados Nominais)

- ▶ Teste χ^2 (qui-quadrado)
- ▶ $\chi^2 = \sum \frac{(\text{observado} - \text{esperado})^2}{\text{esperado}}$
 - ▶ Mais especificamente: $\chi^2 = \sum_{i=1}^{nl} \sum_{j=1}^{nc} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$
 - ▶ o_{ij} e e_{ij} são as frequências, respectivamente, observadas e esperadas, e nl e nc o número de linhas e colunas, respectivamente
 - ▶ $e_{ij} = \frac{\text{quantidade}(A = a_i) \times \text{quantidade}(B = b_j)}{n}$
- ▶ Hipótese nula, H_0 : Não há associação entre os grupos, ou seja, as variáveis são independentes.
 - ▶ O teste usa graus de liberdade, onde $GL = (nl - 1) \times (nc - 1)$
 - ▶ Se $\chi^2 \geq \chi^2_{\text{tabelado}}$ rejeita-se H_0

Tabela do χ^2

g/P	0,99	0,95	0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	,0002	0,004	0,016	0,064	0,148	0,455	1,074	1,642	2,706	3,841	5,412	6,635	10,827
2	0,020	0,103	0,211	0,446	0,713	1,386	2,408	3,219	4,605	5,991	7,824	9,210	13,815
3	0,115	0,352	0,584	1,005	1,424	2,366	3,665	4,642	6,251	7,815	9,837	11,345	16,266
4	0,297	0,711	1,064	1,649	2,195	3,357	4,878	5,989	7,779	9,488	11,668	13,277	18,467
5	0,554	1,145	1,610	2,343	3,000	4,351	6,064	7,289	9,236	11,070	13,388	15,080	20,515
6	0,872	1,635	2,204	3,070	3,828	5,348	7,231	8,558	10,645	12,592	15,033	16,812	22,457
7	1,239	2,167	2,833	3,822	4,671	6,346	8,383	9,803	12,017	14,067	16,622	18,475	24,322
8	1,646	2,733	3,490	4,594	5,527	7,344	9,524	11,030	13,362	15,507	18,168	20,090	26,125
9	2,088	3,325	4,168	5,380	6,393	8,343	10,656	12,242	14,684	16,919	19,679	21,666	27,877
10	2,558	3,940	4,865	6,179	7,267	9,342	11,781	13,442	15,987	18,307	21,161	23,209	29,588
11	3,053	4,575	5,578	6,989	8,148	10,341	12,899	14,631	17,275	19,675	22,618	24,725	31,264
12	3,571	5,226	6,304	7,807	9,034	11,340	14,011	15,812	18,549	21,026	24,054	26,217	32,909
13	4,107	5,892	7,042	8,634	9,926	12,340	15,119	16,985	19,812	22,362	25,472	27,688	34,528
14	4,660	6,571	7,790	9,467	10,821	13,339	16,222	18,151	21,064	23,685	26,873	29,141	36,123
15	5,229	7,261	8,547	10,307	11,721	14,339	17,322	19,311	22,307	24,996	28,259	30,578	37,697
16	5,812	7,692	9,312	11,152	12,624	15,338	18,418	20,465	23,542	26,296	29,633	32,000	39,252
17	6,408	8,672	10,085	12,002	13,531	16,338	19,511	21,615	24,769	27,587	30,995	33,409	40,790
18	7,015	9,390	10,865	12,857	14,440	17,338	20,601	22,760	25,989	28,869	32,346	34,805	42,312

Exemplo de Cálculo de χ^2

	<i>male</i>	<i>female</i>	<i>Total</i>
<i>fiction</i>	250 (90)	200 (360)	450
<i>non_fiction</i>	50 (210)	1000 (840)	1050
Total	300	1200	1500

$$\chi^2 = \frac{(250 - 90)^2}{90} + \frac{(50 - 210)^2}{210} + \frac{(200 - 360)^2}{360} + \frac{(1000 - 840)^2}{840}$$

$$= 284.44 + 121.90 + 71.11 + 30.48 = 507.93.$$

- ▶ $GL = (2 - 1) \times (2 - 1) = 1$
- ▶ Com nível de significância de 0.001, o valor tabelado é 10.828
- ▶ Portanto, rejeita-se H_0 : os atributos são correlacionados

Covariância: Valores Numéricos

$$Cov(A, B) = E[(A - \bar{A})(B - \bar{B})] = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{n}$$

- ▶ O sinal na covariância indica o tipo de relação que as duas variáveis tem
- ▶ Um sinal positivo indica que elas movem juntas e um negativo que elas movem em direções opostas
- ▶ Independência implica em covariância nula, mas o contrário não é válido

Covariância

- $$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{n}$$
 It can be simplified in computation as

$$Cov(A, B) = E(A \cdot B) - \bar{A}\bar{B}$$

- Suppose two stocks A and B have the following values in one week: (2, 5), (3, 8), (5, 10), (4, 11), (6, 14).
- Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?
 - $E(A) = (2 + 3 + 5 + 4 + 6) / 5 = 20 / 5 = 4$
 - $E(B) = (5 + 8 + 10 + 11 + 14) / 5 = 48 / 5 = 9.6$
 - $Cov(A, B) = (2 \times 5 + 3 \times 8 + 5 \times 10 + 4 \times 11 + 6 \times 14) / 5 - 4 \times 9.6 = 4$
- Thus, A and B rise together since $Cov(A, B) > 0$.

Covariância

<i>Time point</i>	<i>AllElectronics</i>	<i>HighTech</i>
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

$$E(\text{AllElectronics}) = \frac{6 + 5 + 4 + 3 + 2}{5} = \frac{20}{5} = \$4$$

$$E(\text{HighTech}) = \frac{20 + 10 + 14 + 5 + 5}{5} = \frac{54}{5} = \$10.80$$

$$\begin{aligned} \text{Cov}(\text{AllElectronics}, \text{HighTech}) &= \frac{6 \times 20 + 5 \times 10 + 4 \times 14 + 3 \times 5 + 2 \times 5}{5} - 4 \times 10.80 \\ &= 50.2 - 43.2 = 7. \end{aligned}$$

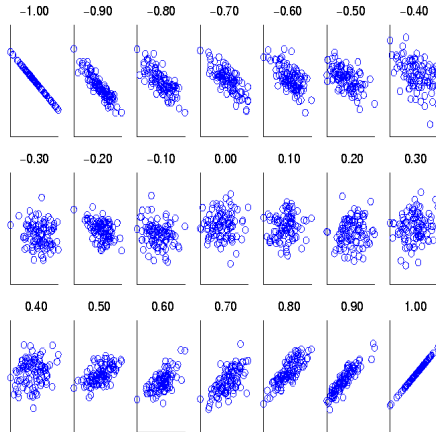
Coefficiente de Correlação: Dados Numéricos

$$\rho_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$$

$$r_{A,B} = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{\sqrt{\sum_{i=1}^n (a_i - \bar{A})^2} \sqrt{\sum_{i=1}^n (b_i - \bar{B})^2}} = \frac{\sum_{i=1}^n a_i b_i - n \bar{A} \bar{B}}{(n-1) \sigma_A \sigma_B}$$

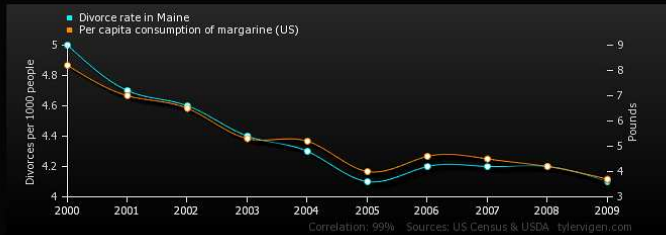
- ▶ n é o número de instâncias, \bar{A} e \bar{B} são as médias de A e B , e σ_A e σ_B são os desvios padrões
- ▶ Se $r_{A,B} > 0$, A e B são positivamente correlacionados
- ▶ Se $r_{A,B} < 0$, A e B são negativamente correlacionados
- ▶ Se $r_{A,B} = 0$, A e B são independentes
- ▶ Quanto maior $|r_{A,B}|$ mais forte é a correlação

Coeficiente de Correlação

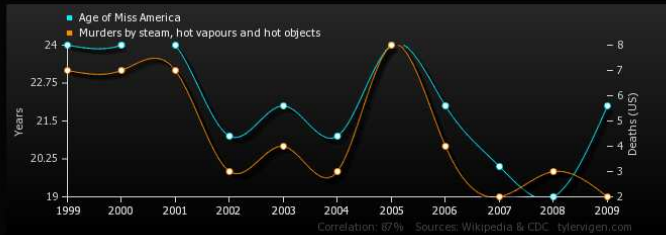


- Correlação não implica em causalidade
 - Número de hospitais e número de carros roubados numa cidade estão correlacionados (a causa é o tamanho da população)

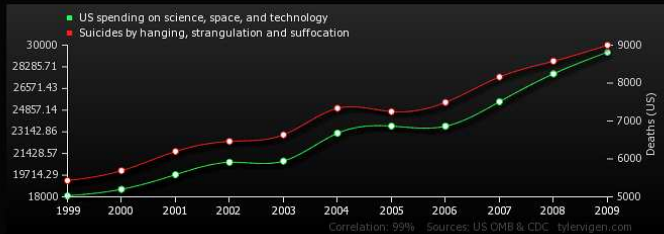
Correlação não implica em causalidade



Correlação não implica em causalidade



Correlação não implica em causalidade



Redução de Dados

Estratégias de Redução

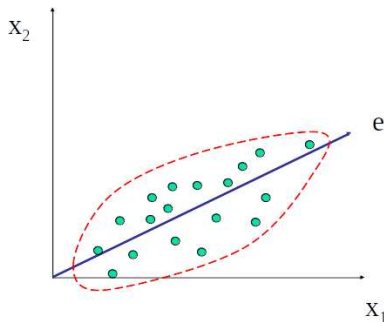
- ▶ Redução dos dados: obter uma representação reduzida dos dados que seja muito menor do que o volume original e que produza os mesmos resultados analíticos (ou similares)
- ▶ Por que reduzir os dados?
 - ▶ Grandes quantidades de dados são difíceis de serem tratados
- ▶ Estratégias para redução dos dados
 - ▶ Redução de dimensionalidade
 - ▶ Redução no número de instâncias
 - ▶ Compressão dos dados

Redução de Dimensionalidade

- ▶ Maldição da dimensionalidade
 - ▶ Quando a dimensão aumenta, os dados se tornam esparsos
 - ▶ Densidade e distância entre pontos, que são elementos críticos para agrupamentos e análises de *outliers*, tornam-se menos significativos
 - ▶ As combinações possíveis de subespaços crescerão
- ▶ Redução de dimensionalidade
 - ▶ Evita a maldição da dimensionalidade
 - ▶ Ajuda a eliminar atributos irrelevantes e reduz ruído
 - ▶ Reduz o tempo e o espaço de memória necessários para a mineração dos dados
 - ▶ Facilita a visualização
- ▶ Técnicas de redução de dimensionalidade
 - ▶ Análise de Componentes Principais
 - ▶ Técnicas supervisionadas: por exemplo, seleção de características

Análise de Componentes Principais (PCA)

- ▶ Encontra a projeção que captura a maior quantidade de variação dos dados
- ▶ Os dados são então projetados num espaço de menor dimensão
- ▶ O novo espaço é definido pelos auto-vetores da matriz de covariância



Análise de Componentes Principais (PCA)

- ▶ Dados N vetores n -dimensionais, encontre $k \leq n$ vetores ortogonais (componentes principais) que melhor representem os dados
 - ▶ normalize os dados: os atributos devem estar dentro de um mesmo limite
 - ▶ compute os componentes principais
 - ▶ ordene os componentes inversamente em relação à “significância”
 - ▶ pode-se reduzir a dimensão eliminando os componentes “mais fracos”
- ▶ Aplicável apenas a dados numéricos

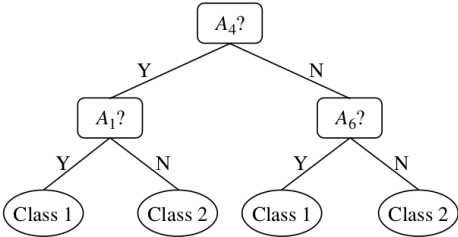
Seleção de Atributos

- ▶ Atributos redundantes
 - ▶ duplicações
 - ▶ uma informação espalhada em vários atributos
- ▶ Atributos irrelevantes
 - ▶ não contém um dado relevante para a mineração

Seleção de Atributos: Busca Heurística

- ▶ Há 2^d possíveis combinações de d atributos
- ▶ Métodos comuns para a seleção de atributos:
 - ▶ melhor atributo segundo testes de significância
 - ▶ adição atributo a atributo
 - ▶ eliminação atributo a atributo
 - ▶ combinação de seleção e eliminação de características
 - ▶ via árvores de decisão
- ▶ O critério de parada pode ser definido como um limitador sobre alguma medida de qualidade

Seleção de Atributos: Busca Heurística

Forward selection	Backward elimination	Decision tree induction
<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p> <p>Initial reduced set: $\{\}$ $\Rightarrow \{A_1\}$ $\Rightarrow \{A_1, A_4\}$ \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$</p>	<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p> <p>$\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_4, A_5, A_6\}$ \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$</p>	<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p>  <pre> graph TD A4["A4?"] -- Y --> A1["A1?"] A4 -- N --> A6["A6?"] A1 -- Y --> C1_1((Class 1)) A1 -- N --> C2_1((Class 2)) A6 -- Y --> C1_2((Class 1)) A6 -- N --> C2_2((Class 2)) </pre> <p>\Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$</p>

Geração de Atributos

- ▶ Cria novas características que capturam as informações importantes do conjunto de dados mais eficientemente do que os atributos originais
- ▶ Alguns métodos:
 - ▶ Extração de atributos: específico de domínio de conhecimento
 - ▶ Mapeando dados num outro espaço: transformada de Fourier, transformação *wavelet*,...
 - ▶ Construção de atributos: combinando características

Redução de Instâncias

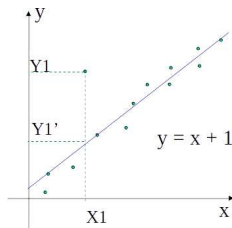
- ▶ Diminui o volume de dados tentando manter a mesma representatividade
- ▶ Métodos Paramétricos
 - ▶ Assume que os dados se ajustam a um dado modelo, estima-se os parâmetros do modelo, guarda-se esses parâmetros e descarta-se os dados
 - ▶ Pode-se manter os *outliers*
 - ▶ Modelos de regressão
- ▶ Métodos Não Paramétricos
 - ▶ Não utiliza um modelo
 - ▶ Exemplos: histograma, agrupamento, amostragem,...

Modelos de Regressão e Log-Linear

- ▶ Regressão Linear
 - ▶ Modela-se os dados por uma reta
 - ▶ Normalmente utiliza o Método dos Mínimos Quadrados
- ▶ Regressão Múltipla
 - ▶ Modela a resposta esperada como uma função linear e multidimensional das variáveis
- ▶ Modelo Log-linear
 - ▶ Aproxima a distribuição de probabilidade

Análise de Regressão

- ▶ Técnicas para modelar e analisar dados numéricos consistindo de
 - ▶ variável dependente ou variável de resposta
 - ▶ variáveis independentes ou explicatórias
- ▶ Pretende-se obter o modelo que “melhor” se ajuste aos dados
- ▶ Critério comum de avaliação: soma das diferenças entre os valores esperados e preditos ao quadrado
 - ▶ Mínimos Quadrados



Análise de Regressão

▶ Regressão Linear

- ▶ $y = wx + b$

- ▶ Os coeficientes w e b devem ser determinados

▶ Regressão Múltipla

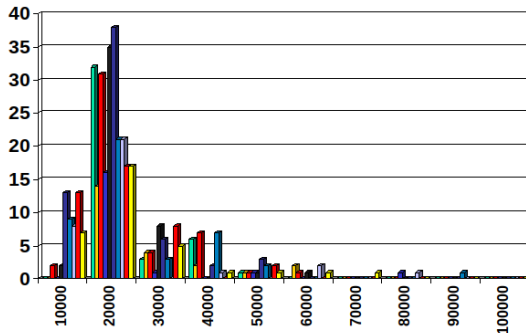
- ▶ $y = b_0 + \sum_{i=1}^n b_i x_i$

- ▶ Os coeficientes b_i devem ser determinados

▶ Modelo Log-linear

- ▶ Aproxima distribuições de probabilidade multidimensionais e discretas
- ▶ Probabilidade de cada instância

Análise de Histograma

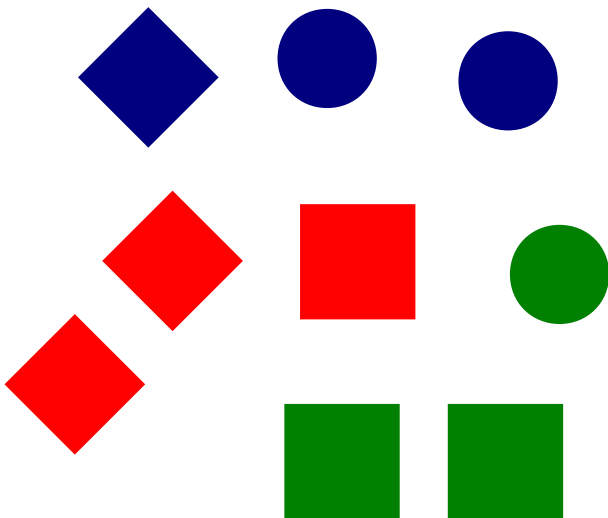


- ▶ Divide os dados em pacotes e guarda a média de cada pacote
- ▶ Regras de particionamento:
 - ▶ Largura: os pacotes tem larguras iguais
 - ▶ Frequência

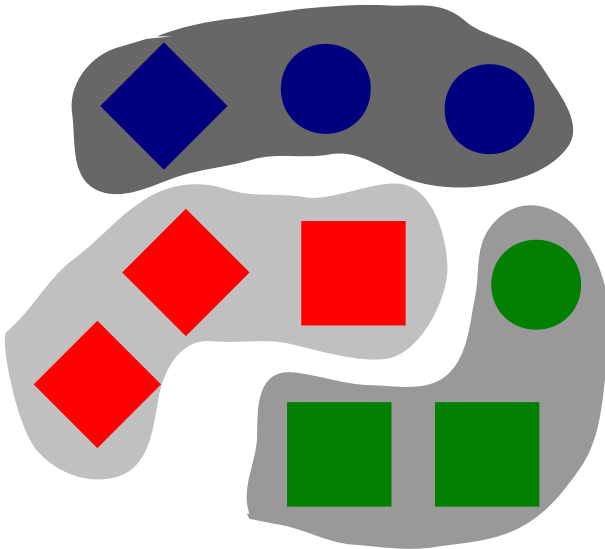
Agrupamento

- ▶ Particionar os dados em grupos baseando-se em similaridade
- ▶ Apenas uma representação dos grupos é necessária
 - ▶ Centroides e diâmetro, por exemplo
- ▶ Pode não ser efetivo se os dados estiverem espalhados
- ▶ Há agrupamentos hierárquicos
 - ▶ Podem ser armazenados em estruturas de árvore de índices
- ▶ Há vários algoritmos e definições de agrupamento

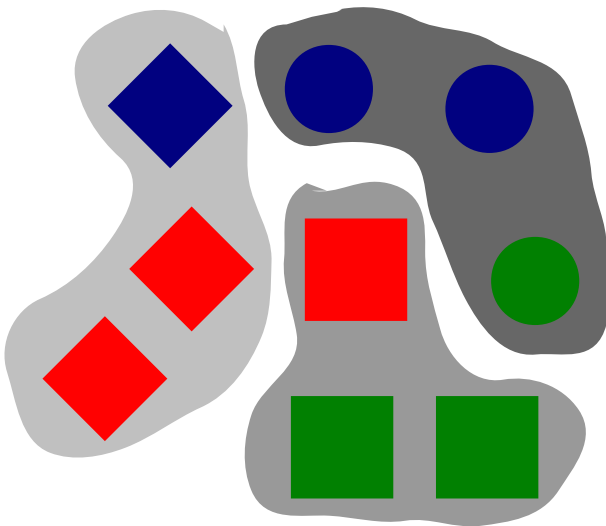
Agrupamento



Agrupamento



Agrupamento



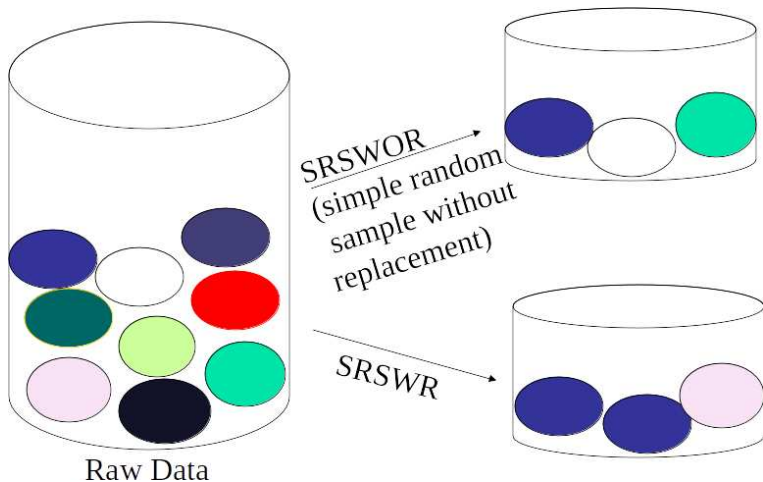
Amostragem

- ▶ Obter uma amostra s pequena para representar todo o conjunto N de dados
- ▶ Permite que as técnicas de aprendizagem executem em menos tempo
- ▶ Os dados amostrados devem ser representativos
 - ▶ Amostragem aleatória simples pode ser enviesada
 - ▶ Métodos adaptativos podem ser adotados
- ▶ A amostragem não reduz os acessos iniciais ao banco de dados
 - ▶ Há redução nas análises posteriores

Tipos de Amostragem

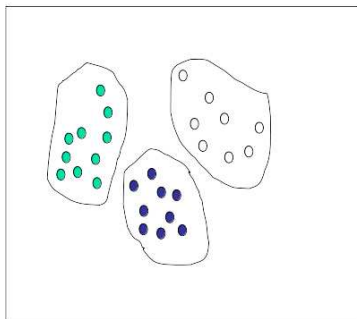
- ▶ Amostragem aleatória simples
 - ▶ Todas as instâncias tem probabilidade igual de serem selecionadas
- ▶ Amostragem sem reposição
 - ▶ Cada instância pode ser selecionada apenas uma vez
- ▶ Amostragem com reposição
 - ▶ Uma instância pode ser selecionada mais de uma vez
- ▶ Amostragem estratificada
 - ▶ Os dados são particionados e as amostras são obtidas em igual proporção de cada partição
 - ▶ Importante quando os dados estão desbalanceados

Amostragem com e sem Repetição

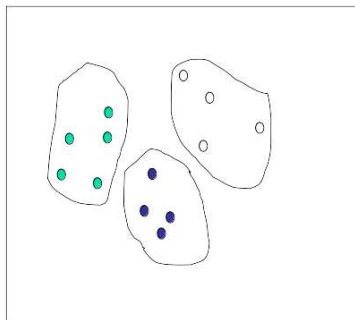


Amostragem

Raw Data



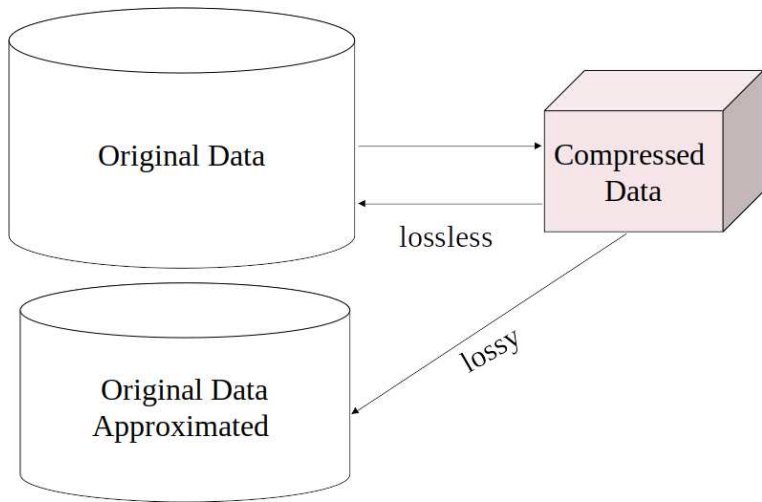
Cluster/Stratified Sample



Compressão dos Dados

- ▶ Compressão de texto
 - ▶ Há uma extensa teoria e algoritmos
 - ▶ Normalmente não há perdas, mas apenas uma manipulação limitada é possível sem que haja expansão
- ▶ Compressão de áudio e vídeo
 - ▶ Usualmente há perdas
 - ▶ Pode ser feita progressivamente
 - ▶ Fragmentos de sinal podem ser reconstruídos sem precisar de manipular o todo
- ▶ Redução de dimensionalidade e de volume dos dados podem ser consideradas compressão dos dados

Compressão de Dados



Transformação dos Dados

Transformação dos Dados

- ▶ Busca tratar os dados para
 - ▶ melhorar o processo de geração de modelos
 - ▶ tornar os padrões gerados mais fáceis de serem interpretados
- ▶ Está relacionado a outros componentes do pré-processamento dos dados

Estratégias de Transformação

- ▶ Suavização
 - ▶ Busca reduzir o ruído dos dados
 - ▶ Relacionado à etapa de limpeza dos dados
 - ▶ Técnicas envolvem alisamento, regressão e agrupamento
- ▶ Construção de atributos
- ▶ Agregação ou sumarização
 - ▶ O dado agregado num intervalo de tempo pode ser importante do que o valor isolado na instância original
- ▶ Normalização
 - ▶ Os atributos são escalados para um intervalo determinado
 - ▶ Visto no conteúdo de “Conhecendo os seus Dados”

Estratégias de Transformação

- ▶ Discretização
 - ▶ Alterar um valor (normalmente numérico) para um conjunto de possibilidades discretas
 - ▶ Pode ser feita via alisamento, histograma, agrupamento, árvores de decisão, ...
 - ▶ Ajuda na interpretação do modelo
 - ▶ Pode reduzir ruídos e remover *outliers*
- ▶ Hierarquização
 - ▶ Transformação de dados nominais
 - ▶ Reduz a quantidade de valores distintos possíveis
 - ▶ rua < cidade < estado < país