A Preparação dos Dados

Escolhas Básicas

Objetos, casos, instâncias

Objetos do mundo real: carros, arvores, etc

Ponto de vista da mineração: um objeto é descrito por uma coleção de características sobre as quais podem ser realizadas medidas

Conceito

A entidade a ser aprendida. Alguns algoritmos de aprendizagem fornecem uma descrição de um conceito

Medidas

O que é possível medir sobre as características: meu carro é azul escuro, 2 portas, 6 cilindros, 5 passageiros

Variáveis, descritores

Uma variável representa uma medida que toma um numero particular de valores, com a possibilidade de valores diferentes para cada observação.

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 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 de valores numéricos, existe não apenas uma ordem entre os valores, mas também existe diferença entre esses valores. O zero é relativo.
Ex: Temperatura em Graus Celsius

Escala Proporcional

Nessa escala de valores numéricos, além da diferença, tem sentido calcular a proporção entre valores (o zero é absoluto).
Ex: Peso, Altura, etc.

Cardinalidade dos atributos das variáveis

Qualitativo / quantitativo

Variáveis qualitativas: escalas nominais ou ordinais

Variáveis quantitativas: escalas intervalares e proporcionais

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)

Cardinalidade: Discreto versus Continuo

Variáveis Discretas

Qualquer variável que possui um conjunto finito de valores distintos.
Ex: Departamentos do CIn

Variáveis contínuas

Podem, em principio, assumir qualquer valor dentro de um intervalo.

Exemplo: Peso, altura

Valores ausentes e valores inaplicáveis

Valores ausentes

Um valor ausente é aquele ausente no conjunto de dados mas existente no contexto em que a medida foi realizada

Numa base de dados eles são indicados por valores negativos ou nulos em atributos numéricos.

Em atributos não numéricos por brancos ou traços.

As vezes são indicados por uma mesma constante

Valores ausentes e valores inaplicáveis

Valores inaplicáveis

Um valor inaplicável é um valor ausente e inexistente no contexto em que a medida foi realizada.

Ex: Sexo = Masculino e Número de Partos = null

Sexo = Feminino e Número de Partos = 0

Valores ausentes e valores inaplicáveis

Valores ausentes e vazios

A diferenciação entre valores ausentes e valores inaplicáveis é importante mais ainda não se dispõe de técnicas automáticas para fazer isso. Deve-se faze-lo manualmente

Quase todas as ferramentas de modelização dispõem de técnicas para tratar dados ausentes: ignora - los, atribuir um valor fixo aos valores ausentes ou estimar os valores ausentes à partir de outras variáveis

Em algumas situações os dados ausentes são altamente informativos e ao serem tratados perde-se essa informação

Mudança de Escala

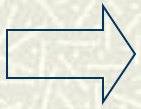
Interesse Muitos modelos só se aplicam à variáveis de mesma escala

Intervalar ⇨ **Ordinal**

Ex: Idade O = [0, 150]
0-20: jovem; 20-60: adulto; >60: idoso
O'={jovem, adulto, idoso}

Trata-se de subdividir O em subintervalos contíguos e associar a cada um deles uma modalidade

Mudança de Escala

Intervalar  Ordinal

Perda de informação

- Distinção entre objetos de uma mesma categoria
- Amplitude da diferença entre objetos de categorias diferentes

Mudança de Escala

Ordinal ⇨ Nominal

Basta desconsiderar a ordem entre as modalidades

Ordinal ou Nominal ⇨ Binária

Cada modalidade é transformada em uma variável binária

- Codificação disjuntiva
- Codificação aditiva

Mudança de Escala

Ordinal ou Nominal ⇨ Binária

Cor: 1(verde), 2(azul), 3(marrom)
Idade: 1(0-20), 2(20-60), 3(> 60)

	Cor	Idade
w	1	2
w'	2	1

	Verde		Azul	Marrom	0-20	0-60	>60
w	1	0	0	1	1	0	
w'	0	1	0	1	0	0	

Representação de Dados para a Mineração

Representação dos Dados

Tabelas de Dados (flat file): as colunas representam as variáveis e as linhas representam as observações

	y_1	y_2	...	y_p
i_1				
⋮				
i_n				

Necessidade do pré-processamento dos Dados

Os dados no mundo real estão "sujos":

- Incompletos
  - ausência de atributos de interesse
  - apenas dados agregados
  - ausência de valores
- Ruidosos
  - erros aleatórios
  - valores aberrantes (outliers)
- Inconsistentes
  - discrepâncias nas codificações ou nos nomes

Sem dados de boa qualidade o resultado da mineração é pobre

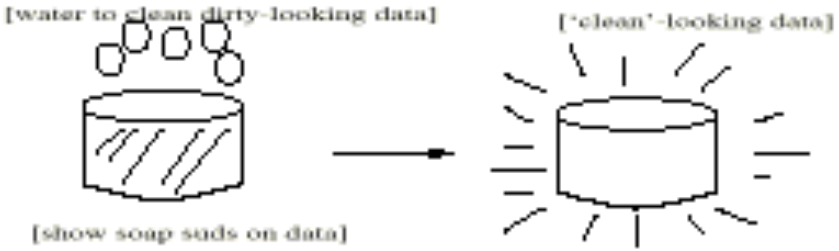
Pré-processamento dos dados

Principais etapas na preparação de dados

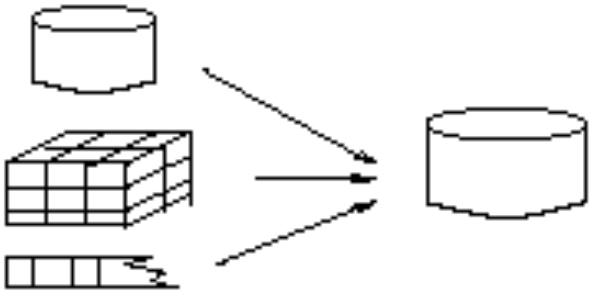
- Limpeza dos dados
  - preencher dados ausentes, “alisar” ruído, identificar e/ou remover
  - valores aberrantes, resolver inconsistências
- Integração e transformação de Dados
  - integração de múltiplas bases de dados, cubos e arquivos
  - Normalização e agregação
- Redução de Dados
  - redução no volume de dados com resultados similares
- Discretização e Construção de Hierarquias Conceituais
  - importante para dados numéricos

Pré-processamento dos dados

Data Cleaning

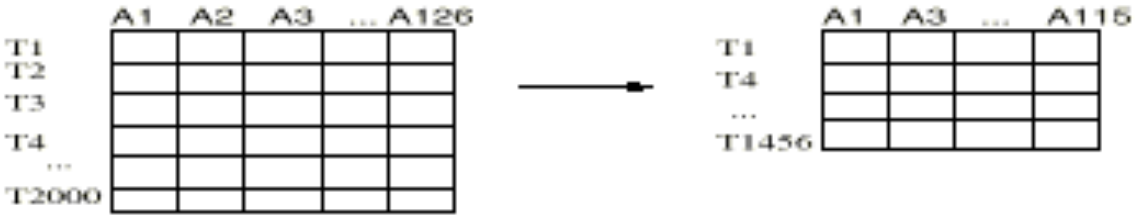


Data Integration



Data Transformation -2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



Limpeza dos dados

Em que consiste a “limpeza” dos dados?

- preencher dados ausentes
- “alisar” o ruido
- identificar valores aberrantes
- Identificar inconsistências
- etc

Valores ausentes

- Dados não estão sempre disponíveis
  - Ex., muitas tuplas não tem nenhum valor gravado para vários atributos (renda do cliente em dados relativos a vendas
- A ausência de dados pode ser consequência
  - mau funcionamento do equipamento
  - inconsistência com outros dados gravados e conseqüente supressão
  - Não entrada de dados devido a enganos
  - determinados dados podem não ser considerados importantes no momento do registro
  - etc.
- Pode ser necessário inferir os dados ausentes

Valores ausentes

Quais os tratamentos usuais para valores ausentes?

- Ignorar a descrição do indivíduo ou mesmo eliminar o descritor;
- Preencher os valores ausentes manualmente;
- Usar uma constante global para representar os valores ausentes (não recomendado, pois o sistema pode identificar esse valor como um conceito);
- Usar a média (ou a moda);
- Usar a média (ou a moda) por classe
- Usar o valor mais provável segundo um modelo (regressão, regra de Bayes, árvores de decisão)

Dados com ruído e/ou valores aberrantes

Ruído: erro aleatório ou variabilidade presente em descritores

Algumas técnicas para a remoção de ruído

- Alisamento
- Regressão

Algumas técnicas para a identificação de valores aberrantes

- Clustering
- Inspeção

Dados com ruído e /ou valores aberrantes

Alisamento: consiste em distribuir dados ordenados em caixas 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

*caixa*1 *caixa*2 *caixa*3

Alisamento pela mediana

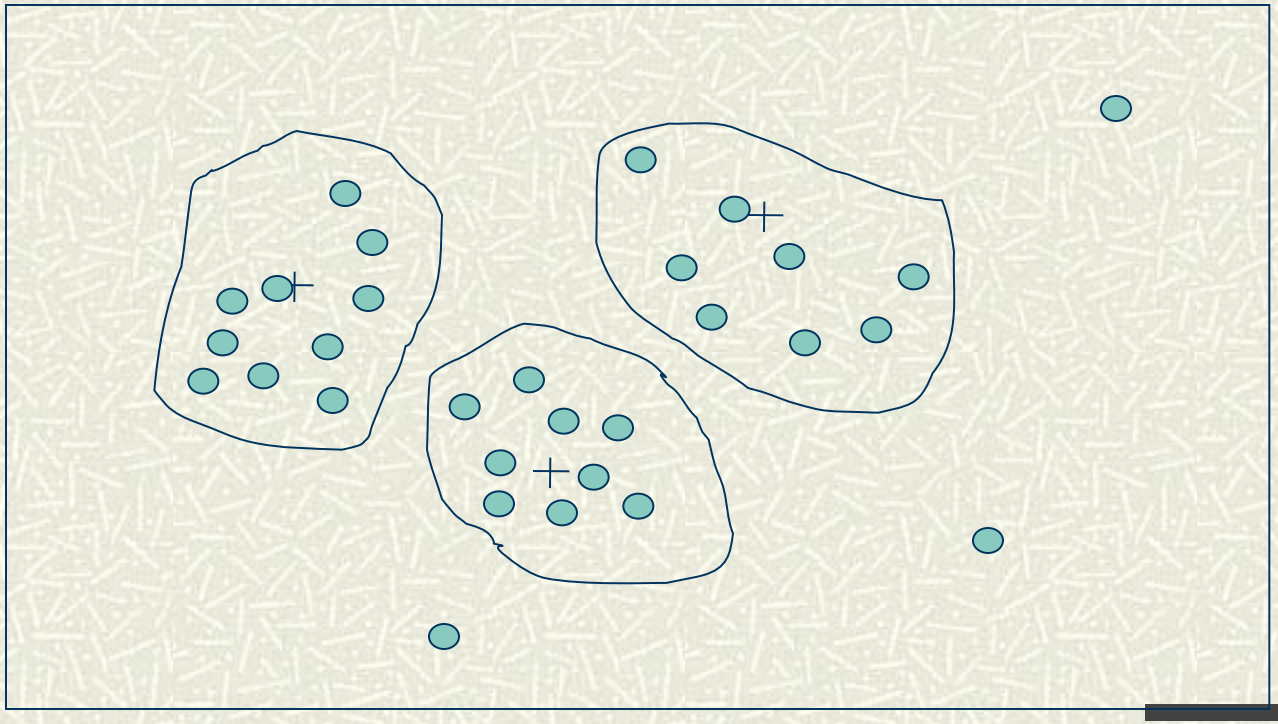
1,1,1 3,3,3 5,5,5,5

*caixa*1 *caixa*2 *caixa*3

Outras alternativas: média, fronteiras

Dados com ruído e /ou valores aberrantes

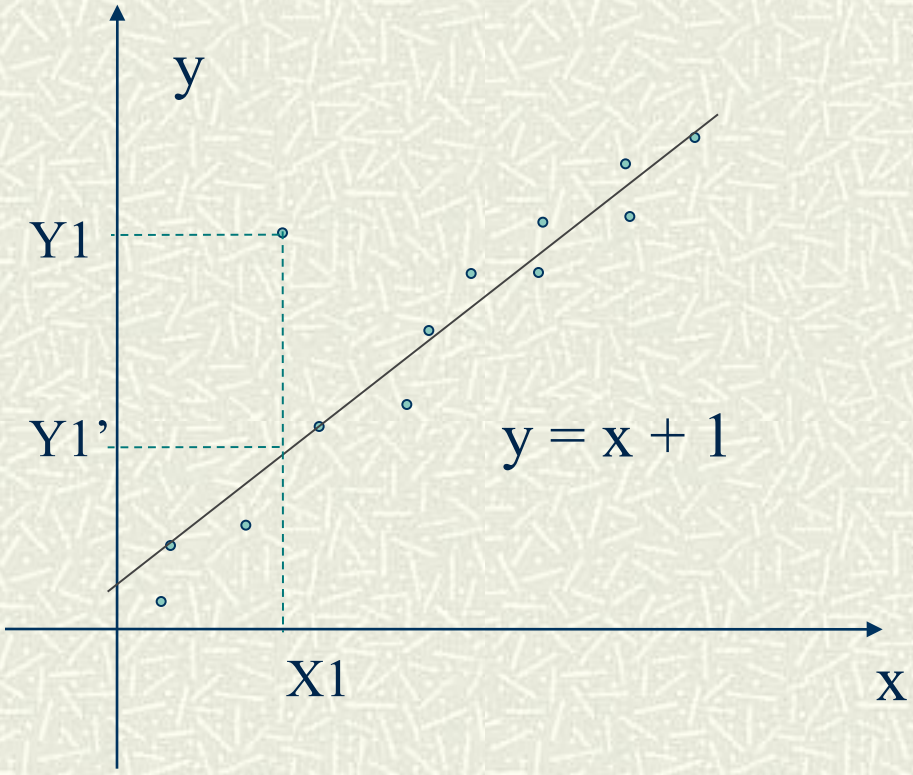
- Clustering: deteção e remoçõ de valores aberrantes
  - os valores são organizados em grupos; os valores isolados podem ser considerados aberrantes;



Dados com ruído e /ou valores aberrantes

- Regressão:

  - os dados podem ser alisados pelo ajustamento a uma função (regressão linear, por exemplo);



Dados Inconsistentes

- Erros no momento de introdução dos dados
- Erros oriundos da integração de várias bases de dados
  - mesmo atributo com diferentes codificações;
  - duplicação de objetos
- etc

Integração e Transformação de Dados

- Integração de dados

  - Fusão de dados à partir de diferentes fontes em uma única fonte coerente. As fontes podem ser bases de dados, cubos ou arquivos texto

- Transformação de Dados

  - é necessário para obter os mesmos em uma forma apropriada para a mineração

Integração de Dados

Esquema em bases de dados relacionais

- identificação das mesmas entidades do mundo real a partir de múltiplas fontes de dados
- Integração dos metadados de diferentes fontes

Redundância

Dados redundantes ocorrem quando da integração de bases de dados

- Diferentes nomes para um mesmo atributo;
- Um atributo pode ser derivado diretamente de outro;

Análise de correlação: instrumento para a detecção de redundâncias

Duplicação de objetos;

Integração de Dados

Detecção e resolução de conflitos

Os valores de um mesmo atributo pode diferir segundo as diversas fontes

Isso pode acontecer devido a diferenças na representação,
Escala ou codificação

Peso (em libras ou em quilos)
Altura (valor numérico ou categórico (médio, pequeno...)
Preço (pode indicar serviços diferentes)

Transformação de dados

Objetivo:

obter os dados em uma forma mais apropriada para a mineração

Alisamento

Agregação: sumários dos dados (soma, etc) quando da construção de cubos para OLAP

Generalização

Dados primitivos são substituídos por conceitos de ordem superior via uma hierarquia de conceitos.

Ex. valores do atributo numérico idade são mapeados em jovem, meia-idade, etc.

Construção de novos atributos

Transformação de dados

Normalização

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

As variáveis podem ser normalizadas segundo a amplitude ou segundo a distribuição

Algumas ferramentas de modelização são beneficiadas com a Normalização (redes neurais, KNN, clustering)

Transformação de dados

Normalização segundo a amplitude

Justificativa: unidades diferentes ou dispersões muito heterogêneas

$$a \quad y = \frac{x - m}{s}$$

$$b \quad y = \frac{x - \min}{\max - \min}$$

$$c \quad y = \frac{x}{10^k}, \text{ para o menor k tal que max } \left|\frac{x}{10^k}\right| \quad 1$$

Transformação de dados

Normalização distribucional

A normalização distribucional é interessante em várias situações: remoção de distorções de valores aberrantes, obtenção de simetria etc.

As transformações mais comuns são:

$$\sqrt{x} \qquad \log x \qquad -\frac{1}{x}$$

A mais suave é a raiz e a mais forte é a inversa negativa

Redução de Dados

Razões para a redução de dados:

- ultrapassagem da capacidade de processamento dos programas de aprendizagem
- tempo muito longo para obter uma solução

Redução de dados:

- Obtem uma respresentação reduzida da série de dados de que é muito menor no volume mas contudo produz os mesmos (ou quase os mesmos) resultados analíticos

Outras vantagens da redução de dados:

- redução do tempo de aprendizagem
- interpretação mais fácil dos conceitos aprendidos

Redução de Dados

Estratégias para a redução de dados

- Agregação via cubo
- Redução de dimensão
- Compressão de dados
- Redução de casos
- Discretização e construção de hierarquias conceituais

Redução de Dados

Redução de dimensão

Em data mining a supressão de uma coluna (atributo) é muito mais Delicada do que a supressão de uma linha (observação)

Retirar atributos relevantes ou permanecer com atributos irrelevantes 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

Redução de Dados

Algumas abordagens automáticas de seleção de variáveis

Seleção do menor conjunto de atributos

Selecionar o menor conjunto de atributos suficiente para dividir o espaço das instancias de tal maneira que a distribuição das classes no novo espaço é tão próxima quanto possível daquela do espaço original

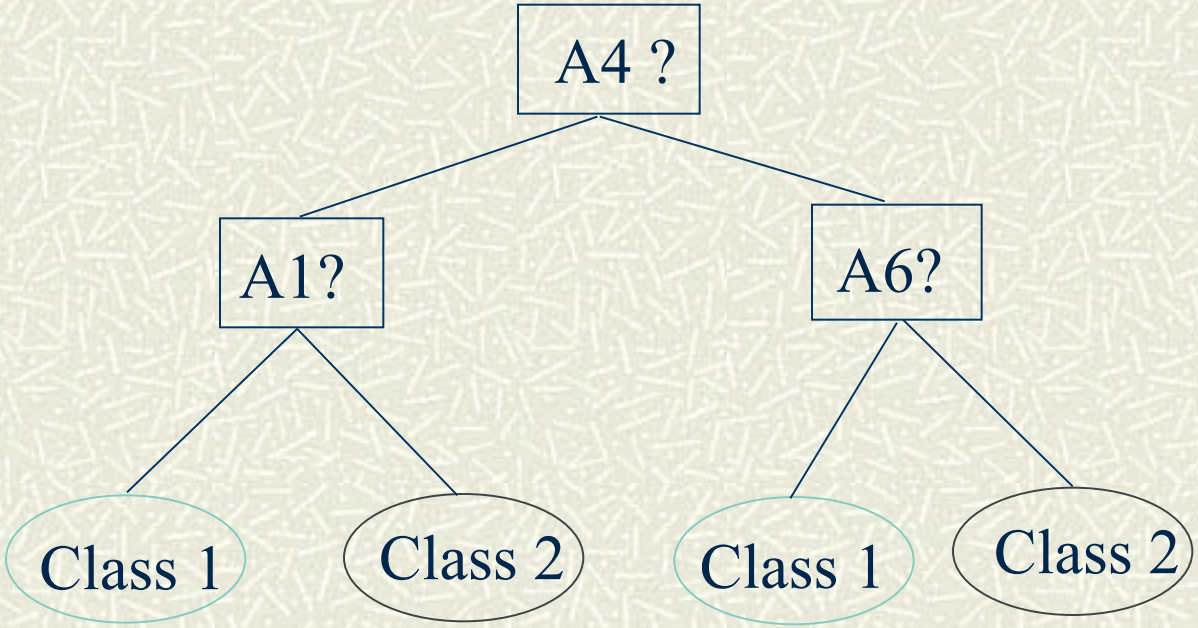
Problemas: busca exaustiva e overfitting

Algoritmo de construção de árvores de decisão

Aplicar esse algoritmo nos dados completos e então selecionar apenas as variáveis presentes na árvore de decisão

Redução de Dados

Conjunto inicial de atributos:
{A1, A2, A3, A4, A5, A6}



-------> Conjunto reduzido de atributos: {A1, A4, A6}

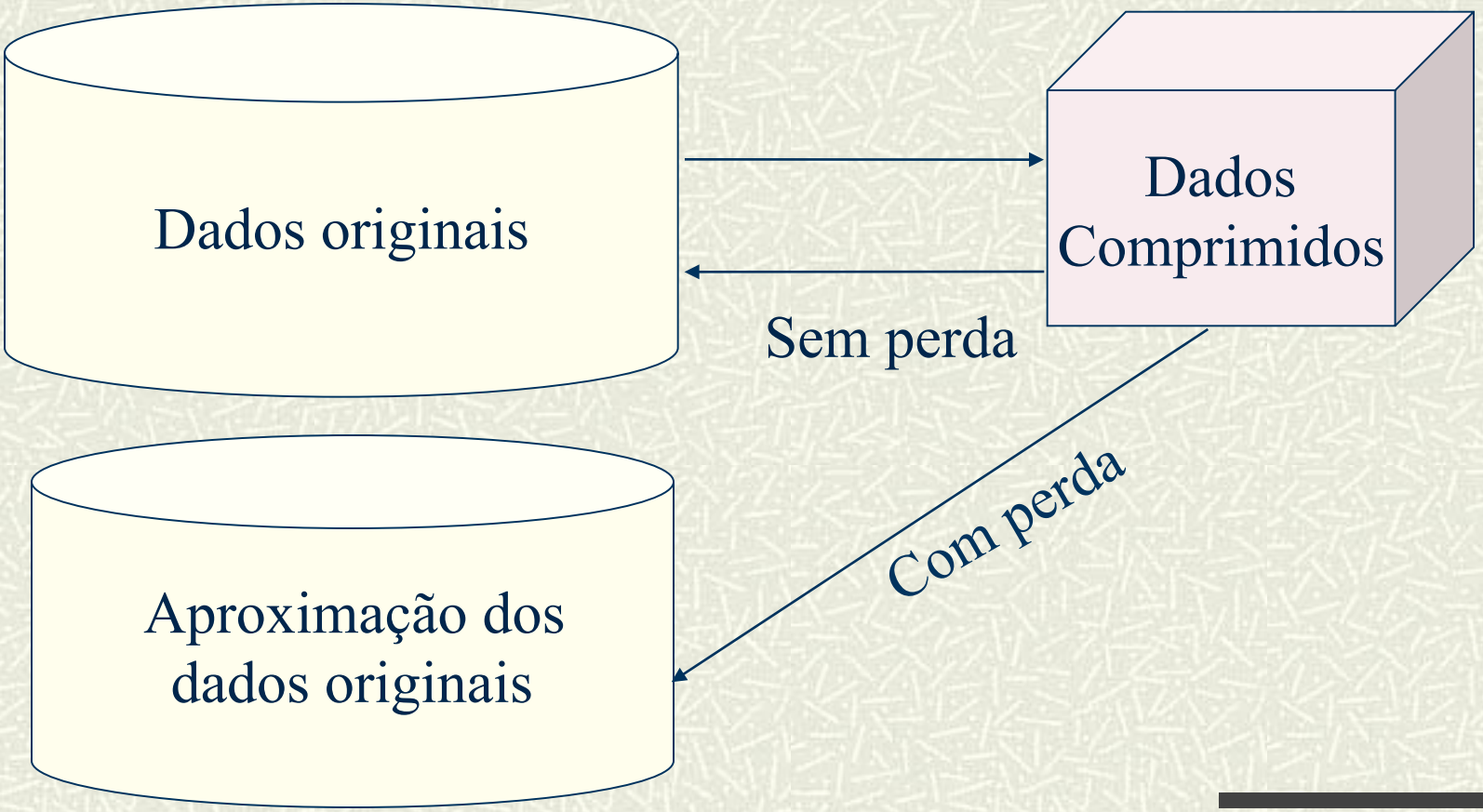
Redução de Dados

Seleção por busca no espaço de atributos

- Existem 2^d possíveis sub-conjuntos de a partir de d atributos
- Várias heurísticas para a seleção de variáveis
  - 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
  - combinação da seleção forward com a eliminação backward

Compressão de Dados

Essas técnicas comprimem os dados originais



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 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

Compressão de Dados

Extração de Variáveis

Famílias de Métodos

- Métodos não supervisionados
- Métodos supervisionados

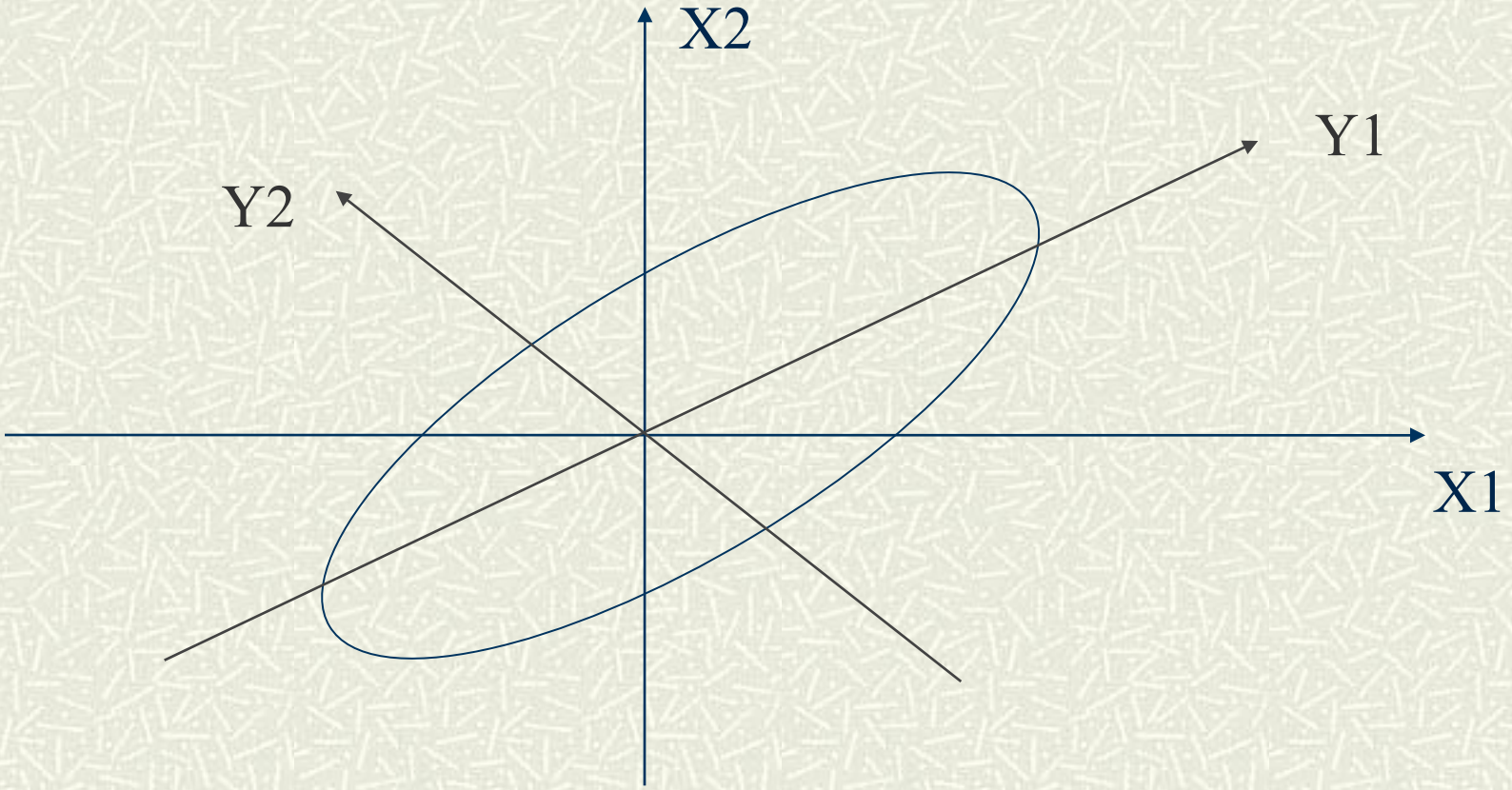
Métodos não supervisionados:

- Análise de Componentes Principais (variáveis quantitativas)
- Análise de Correspondências (variáveis qualitativas)

A primeira componente é a combinação linear das variáveis iniciais de maior variância (maximiza a separação entre os indivíduos). A segunda componente é ortogonal a primeira (correlação nula), é também combinação linear das variáveis iniciais e apresenta a segunda maior variância. E assim por diante.

Compressão de Dados

Extração de Variáveis



Compressão de Dados

Extração de Variáveis

Métodos supervisionados

Análise Fatorial Discriminante

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.

Redução de Casos

Redução do volume de dados via representação *econômica* dos mesmos

- Métodos paramétricos
  - Supõe que os dados ajustam um modelo, estimam os parâmetros do modelo, armazena apenas os parâmetros e descarrega os dados (exceto os aberrantes)
  - Principais modelos: regressão (simples e multipla) e modelo log-linear
- Métodos não paramétricos
  - Não assume modelos
  - Famílias principais: histogramas, clustering, amostragem

Redução de Casos

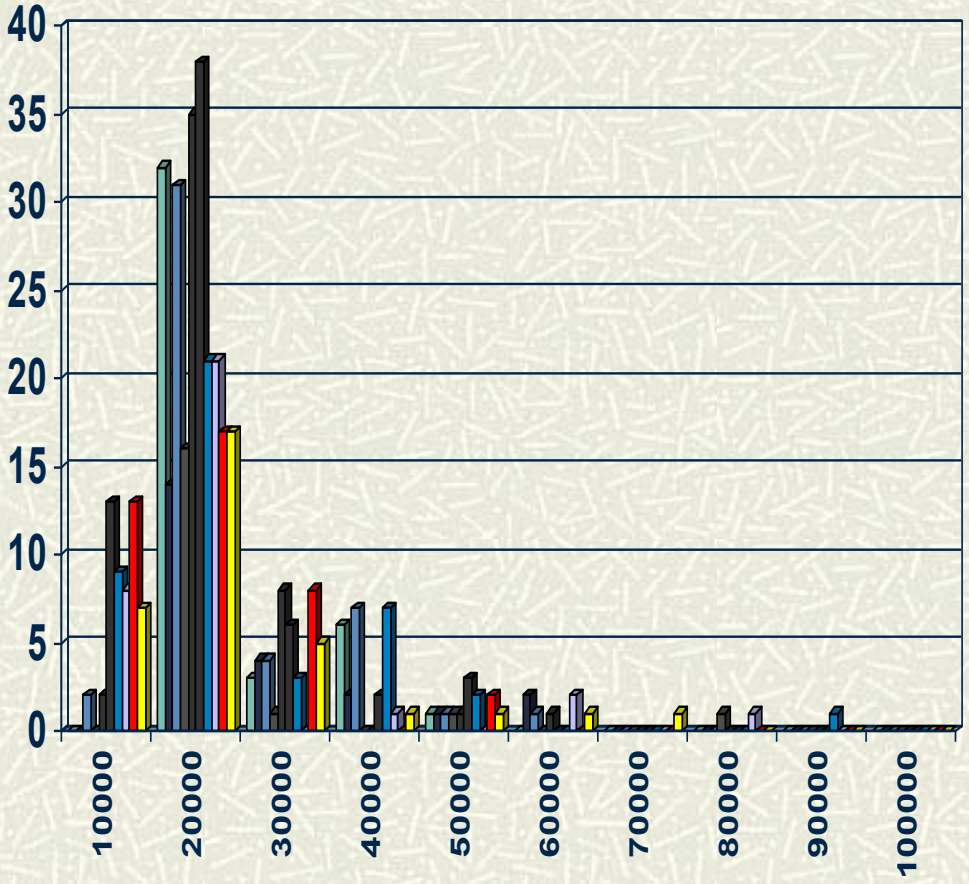
Regressão e modelos log-linear

- Regressão linear: os dados são modelados para se ajustarem a uma linha reta
  - Em geral usa o método dos quadrados mínimos 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 de Casos

Histogramas

- Particiona os dados em caixas e armazena a frequencia média dos valores
- Em uma dimensão pode ser construido pela otimização de um critério via programação dinâmica



Redução de Casos

Clustering

- Os dados são particionados em clusters e armazena-se apenas a representação do mesmo
- Pode ser muito eficaz se os dados são agrupados mas não se estão apenas sujos
- Existem muitas opções de métodos de e algoritmos de agrupamento

Redução de Casos

Amostragem

- Permite que os algoritmos de mineração tratem enormes bases de dados pela redução dos casos

- Tipos de Amostragem:
  - Amostragem aleatória simples com reposição
  - Amostragem aleatória simples sem reposição
  - Amostragem estratificada
  - Amostragem por conglomerado

Redução de Casos

Amostragem

Duas formas básicas de amostragem são interessantes no contexto da mineração de dados:

- Amostragens incrementais
- Amostragens seguida de voto

Redução de Casos

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?

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)

Discretização e Construção de Hierarquias

Interesse: redução do numero de valores.

Muito interessante em árvores de decisão

- Discretização
  - reduz o número de valores de um dado atributo contínuo pela divisão da amplitude do atributo em intervalos. Os rótulos dos intevalos substituem os valores.
- Hierarquias Conceituais
  - reduz os dados pela substituição de rótulos de nível inferior (como os valores numéricos do atributo idade) por rótulos de nível superior (tais como jovem, meia-idade, etc)

Discretização e Construção de Hierarquias

Ferramentas

- Alisamento
- Histograma
- Clustering
- Discretização baseada em entropia
- Segmentação via particionamento “natural”

Discretização e Construção de Hierarquias

Abordagens para a discretização de intervalos:

- discretização não supervisionada
- discretização supervisionada

Discretização não supervisionada

a discretização é realizada sem levar em conta os grupos a que pertencem as instâncias no conjunto de treinamento

Discretização supervisionada

a discretização é realizada levando em conta os grupos a que pertencem as instâncias no conjunto de treinamento

Discretização e Construção de Hierarquias

Técnicas de Discretização não supervisionada

- Partição em intervalos iguais
  riscos: escolher fronteiras que colocam juntas muitas instancias de diferentes classes; intervalos sem nenhuma instancia outras com muitas
- Partição por efetivos iguais
  riscos: escolher fronteiras que colocam juntas muitas instancias de diferentes classes
- Partição em intervalos arbitrários
- Partição por minimização da variância

Discretização e Construção de Hierarquias

Técnicas de Discretização supervisionada

- Discretização divisiva (top-down)

  Exemplo: procura recursiva da partição binária que minimiza o ganho de entropia

- Discretização aglomerativa (bottom-up)

  Exemplo: isolar cada instancia em um intervalo e em seguida fusionar intervalos segundo um critério estatístico

Hierarquias de conceitos para dados categóricos

- Especificação explícita de uma ordem parcial dos atributos ao nível do esquema pelos usuários e/ou especialistas
- Especificação de uma porção de hierarquia via agrupamento de dados
- Especificação do conjunto de atributos, mas não da ordem parcial
- Especificação de de um conjunto de atributos parcialmente

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.