Principais Características dos Dados

Mineração de Dados

Universidade Federal do ABC

Bases de dados

- Tipos de bases de dados: tabela atributo-valor, grafos etc
- Tipos de atributos: uma revisão breve
- Qualidade dos dados: Nenhuma base de dados é perfeita (em especial as grandes). Entre os diferentes problemas que podemos encontrar, os mais comuns são:
  - Valores ausentes
  - Erro humano
  - Sensores defeituosos

Nomenclatura

- Objetos, casos, instâncias, examplos, tuplas:
  - A unidade em que iremos realizar predições ou descrever.
- Atributos, variáveis, descritores, características:
  - Propriedades do objeto que será utilizado na análise

Nomenclatura — Exemplos

- **Aplicação:** Detecção de fraude
  - **Objeto:** Transação
  - **Atributos:** Dia da semana, hora, valor, tipo de estabelecimento
- **Aplicação:** Estimação de renda
  - **Objeto:** Indivíduo
  - **Atributos:** Cargo, empresa, currículo, grau de conectividade (Q.I.)
- **Aplicação:** Análise de crédito
  - **Objeto:** Indivíduo
  - **Atributos:** Renda, Valor em bens próprios, Valor em bens de familiares

Tipos de bases de dados

- Conjunto de registros:
  - Cesta de compras: conjunto de produtos comprados em uma transação
  - *Tabela atributo-valor*: formato mais comum. Cada célula da tabela refere-se ao valor de um objeto de acordo com um atributo.
- Conjunto de grafos:
  - Relacionamentos entre objetos representados por meio de um grafo: análise de páginas web
  - Objetos com estrutura representada como um grafo: compostos químicos e estrutura de proteina

Tipos de bases de dados

- Conjunto ordenado de observações:
  - Dados sequenciais: agregação de tempo em cada registro do conjunto, por exemplo, cesta de compras com *timestamp*
  - Dados de sequência: um registro é uma sequência de itens (não relacionado a tempo). Exemplo mais comum seria DNA (sequência de A, T, G e C)
  - Séries temporais: um registro é uma série de medidas tiradas no tempo. Ex: ações
  - Dados espaciais: um registro é relacionado a uma localização. Ex: dados sobre cidades
- Focaremos em bases de dados no formato de conjunto de registros, em especial, no formato tabela atributo-valor.
  - Se houver tempo, discutiremos sobre dados sequenciais ou séries temporais.

Tipos de atributos

- Escalas
  - ordinal
  - nominal
  - intervalar
  - proporcional
- Relação entre operações com os tipos de escala:

Tipo de atributo	Operações
Nominal	$=, \neq$
Ordinal	$<, \leq, \geq, >$
Intervalar	$+, -$
Proporcional	$*, /$

Tipos de atributos

- Exemplo clássico de atributo intervalar é temperatura em grau Celsius ou Fahrenheit. O zero não é absoluto, não pode-se dizer que 20ºC é o dobro de temperatura que 10ºC. A escala Kelvin não sofre este problema.

Tipos de atributos

- Considere a seguinte tabela atributo-valor, formato mais usual de dados usados em mineração de dados:

RA	Sexo	Conceito	QI	Nota
AB123	M	B	100	8.0
BC234	M	B	150	7.5
CD345	M	C	78	6.0
DE456	F	B	110	7.7
EF567	F	B	91	7.0
GH678	F	A	166	10.0

Tipos de atributos

- Atributos nominais e ordinais são normalmente chamados de atributos qualitativos, enquanto que os demais contínuos ou quantitativos.
- Atributos assimétricos
  - Encontrados em diversas aplicações, por exemplo, mineração de textos.
  - Apenas valores diferentes de 0 são relevantes.
  - Exemplo: as matrículas realizadas por um aluno pode ser representada por um vetor binário.
  - Este tipo de atributo pode ser discreto ou contínuo.

Qualidade dos dados

- Uma vez que os dados utilizados em MD não foram normalmente coletados para a MD, tem-se que considerar as mais diversas fontes de erro.
- *Garbage In→Garbage Out.*
- Uma fonte de erro nos dados que foge do controle do analista de dados é o erro de medição.
  - Valores de uma dada medida não são os valores *verdadeiros*.
  - Considera-se algoritmos *robustos*, que são capazes de relevar tais erros em suas estimativas.

Outliers

- *Outliers* ocorrem com frequência em aplicações.
  - objetos que possuem características diferentes dos demais (ex: análise de cidades do Brasil em números absolutos, SP tem ~12mi habitantes enquanto que a segunda maior, RJ, tem ~6mi)
  - valores de um atributo que são anormais em relação aos valores típicos daquele atributo (ex: análise de jogos de futebol, uma partida com um placar de 15x0)
- Note que *outliers* não são, necessariamente, resultados de erros e sim valores verdadeiros. Aplicações em que se busca identificar estes registros atípicos são chamados de **detecção de anomalia/*outliters***

Definição

- Valores ausentes são casos em que uma célula está "vazia" em uma tabela atributo-valor.
- Existem três categorias do modo que valores ausentes são distribuídos:
  - *Missing Completely At Random* (MCAR): a probabilidade de um valor ausente ocorrer não depende de nenhum valor da base de dados
  - *Missing At Random* (MAR): a probabilidade de um valor ausente ocorrer depende dos valores conhecidos na base de dados
  - *Missing Not At Random* (MNAR): a probabilidade de um valor ausente ocorrer depende do próprio valor ausente

Exemplo

- Considere o seguinte exemplo:
  - Será realizado um levantamento baseado-se em respostas de um questionário com os alunos de um departamento da UFABC.
  - Após terem sido preenchidos, descobriu-se um problema no servidor de banco de dados que removia entradas aleatoriamente.
  - Cada valor ausente é resultado de um processo completamente aleatório (MCAR).

Exemplo

- Considere o seguinte exemplo:
  - Será realizado um levantamento baseado-se em respostas de um questionário com os alunos de um departamento da UFABC.
  - Uma das perguntas no questionário refere-se à grau de depressão.
  - Normalmente, homens não gostam de falar de depressão e tem uma probabilidade maior de não responder esta questão.
  - Nesse caso, o valor ausente ocorre com probabilidade condicional a se o respondente é homem (MAR).

Exemplo

- Considere o seguinte exemplo:
  - Será realizado um levantamento baseado-se em respostas de um questionário com os alunos de um departamento da UFABC.
  - Considere que uma pergunta é "qual a renda de sua família em R$?".
  - Normalmente, pessoas com renda familiar maior vão preferir não responder a pergunta.
  - Em outras palavras, a probabilidade de ausência é maior justamente pelo fato de o valor que seria reportado ser alto (MNAR).

Tratamento

- Existem diversas formas de se tratar valores ausentes. As mais comuns são:
  - Remoção de objetos ou atributos com valores ausentes
  - Ignorar valores ausentes na análise
  - Estimar valores ausentes
    - Neste caso, estimar valores ausentes no caso MNAR pode não ser confiável. Mas, se temos um outro atributo que possui alta correlação com o atributo com valor ausente, ainda pode haver esperança

Representatividade e Viés de Seleção

- Suponha que pretendemos induzir um modelo para estimar a altura de uma pessoa baseada em outras características.
  - Nossos dados foram obtidos de todas as pessoas matriculadas no curso de computação da universidade.
  - O fato de a maior parte da amostra ser referente a alunos do sexo masculino (característica ainda comum nos cursos de computação) terá algum impacto?

Representatividade e Viés de Seleção

- Temos um problema de representatividade se queremos usar nosso modelo para estimar a altura de qualquer pessoa independente do sexo.
- Aplicações baseadas em dados de redes sociais podem sofrer consideravelmente com isso.
- Sabe-se que as pessoas tendem a se *conectar* com pessoas de pensamento parecido.
  - Efeito *bolha*

Population drift

- Suponha que estamos trabalhando com modelos de detecção de spam.
  - Podemos ter um modelo excelente com capacidade de detecção de 95%.
  - Os *spammers* vão se adaptar e modificar o padrão de seus e-mails
  - Nosso modelo pode se tornar inútil
- Não necessariamente envolve um adversário
- Por exemplo, análise de crédito para empréstimo
- Velocidade de captura e fluxo de informações é muito maior do que antigamente.

Referências

- P. Tan, M. Steinbach e V. Kumar, Introduction to Data Mining. **Capítulo 2**
- D. Hand, H. Manilla e P. Smith. Principles of Data Mining. **Capítulo 2**
- João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. ACM Comput. Surv. 46, 4, Article 44 (March 2014), 37 pages. DOI: http://dx.doi.org/10.1145/2523813