SCC0173 – Mineração de Dados Biológicos

Preparação de Dados: Parte A

Prof. Ricardo J. G. B. Campello

SCC / ICMC / USP

Créditos

- O material a seguir consiste de adaptações e extensões:
  - dos originais gentilmente cedidos pelo professor André C. P. L. F. de Carvalho
  - dos originais de Tan et al., *Introduction to Data Mining*, Addison-Wesley, 2006

Aula de Hoje

- Instâncias e Atributos
- Tipos de Atributos
- Preparação de Dados
  - Qualidade de Dados
  - Pré-Seleção de Dados
  - Ruído
  - Valores inconsistentes, duplicados e ausentes
  - Normalizações

Introdução

- **Dados**:
  - coleção de instâncias com seus atributos
- **Instâncias**
  - exemplos, objetos, registros, pontos, casos, entidades, ...
    - Ex.: clientes de um banco, pacientes de um hospital
  - cada instância é formada (descrita) por um conjunto de **atributos**

Introdução

- **Atributo**
  - variável, campo, característica
    - Ex.: altura, sintoma, renda
  - cada atributo representa uma propriedade ou característica específica das instâncias
  - coleção de valores específicos dos atributos descreve uma instância particular
  - seus valores podem ser números ou símbolos

Exemplo

- Seja uma aplicação de DM em medicina
  - Descoberta de conhecimento a partir de um conjunto de dados de pacientes
  - Conjunto possui o cadastro de vários pacientes de um hospital
    - Diagnosticados de acordo com uma dada patologia em:
      - Saudáveis
      - Doentes

Exemplo

Atributos

Instâncias

Nome	Febre	Enjôo	Mancha	Diagnóstico
João	sim	sim	pequena	doente
Pedro	não	não	pequena	saudável
Maria	sim	sim	grande	saudável
José	sim	não	pequena	doente
Ana	sim	não	grande	saudável
Leila	não	não	grande	doente

valor de um atributo

Principais Tipos de Atributos

- **Nominal**
  - cor, identificação, profissão, ...
- **Ordinal**
  - gosto (ruim, médio, bom), dias da semana, ...
- **Numérico**
  - peso, tamanho, idade, temperatura, ...

Tipos de Atributos

	Tipo de Atributo	Descrição	Exemplos
Categórico (Qualitativo)	**Nominal**	Valores são simplesmente nomes (símbolos) diferentes, i.e., atributos nominais provêm apenas informação suficiente para distinguir uma instância de outra: (=, ≠)	Sexo, Estado Civil, CEP, …
Categórico (Qualitativo)	**Ordinal**	Os valores de atributos ordinais provêm informação suficiente para distinguir e ordenar instâncias, i.e.: (=, ≠) e (<, >)	Grau de Educação, Números de Endereço, …
Numérico (Quantitativo)	**Intervalo**	Atributos para os quais a diferença entre valores faz sentido, i.e., existe uma unidade de medida com referência (zero) arbitrário.	Datas, Temperatura em Celsius ou Fahrenheit, ...
Numérico (Quantitativo)	**Razão**	Atributos para os quais não apenas a diferença entre valores faz sentido, mas também a razão entre valores (zero é absoluto).	Contagens, Massa, Largura, Corrente Elétrica, Quantidades Monetárias, Temp. em Kelvin...

Exemplo

- Identificar tipo de cada atributo no cadastro de pacientes abaixo:

Nome	Temp.	Enjôo	Mancha	Dor	Salário	Diagnóstico
João	37,7	sim	pequena	sim	1000	doente
Pedro	37	não	pequena	não	1100	saudável
Maria	38,2	sim	grande	não	600	saudável
José	39	não	pequena	sim	2000	doente
Ana	37,3	não	grande	sim	1800	saudável
Leila	36,9	não	grande	sim	900	doente

Exemplo

Nome	Temp	Enjôo	Mancha	Dor	Salário	Diagnóstico
João	37,7	sim	pequena	sim	1000	doente
Pedro	37	não	pequena	não	1100	saudável
Maria	38.2	sim	grande	não	600	saudável
José	39	não	pequena	sim	2000	doente
Ana	37,3	não	grande	sim	1800	saudável
Leila	37,7	não	grande	sim	900	doente

Exercício

- Definir o tipo dos seguintes atributos:
  - Renda mensal
  - Número de palavras de um texto
  - Endereço de E-mail
  - Número de matrícula
  - Data de nascimento
  - Código postal
  - Posição em uma corrida

Tipos de Atributos

- Atributos também pode ser divididos em:
  - **Discretos** (categóricos ou numéricos)
    - assumem um número contável de valores
      - no. finito ou infinito
  - **Contínuos** (numéricos)
    - assumem uma quantidade incontável de valores

Atributos Contínuos

- Assumem valores que são números reais
  - Temperatura
  - Peso
  - Distância
  - ...

Atributos Discretos

- No. finito ou infinito e enumerável de valores
  - estações do ano, cores elementares, ...
  - no. de filhos, no. estrelas no universo, no. de anos, ...
- Caso especial: Atributos **Binários**
  - 0 ou 1
  - V ou F
  - ...

Atributos Assimétricos

- Um caso ainda mais particular de atributo discreto são os **atributos binários assimétricos**
- Embora assuma dois valores como qualquer atributo binário, apenas a presença de um deles é relevante
  - indica que a instância possui uma dada característica
  - p. ex., aluno matriculado ou não em cada disciplina
    - se no. de disciplinas disponíveis for grande, alunos são todos similares com relação às disciplinas que não cursam...

Conjunto de Dados Alvo

- Conforme já vimos antes, dados alvo são selecionados a partir de BDs "brutos"
  - sob a ótica da aplicação de interesse
- Procedimentos na geração dos dados alvo:
  - Descarte de instâncias
  - Descarte de atributos
  - Integração de bases distribuídas
  - ...

Descarte de Instâncias

- Visa selecionar apenas os registros de interesse ao problema em questão
- Exemplo:
  - Descartar registros de pacientes do sexo masculino em um processo de KDD sobre gravidez e questões clínicas relacionadas

Descarte de Atributos

- Descartar atributos claramente irrelevantes:

Nome	Febre	Enjôo	Mancha	Dor	Salário	Diagnóstico
João	sim	sim	pequena	sim	1000	doente
Pedro	não	não	pequena	não	1100	saudável
Maria	sim	sim	grande	não	600	saudável
José	sim	não	pequena	sim	2000	doente
Ana	sim	não	grande	sim	1800	saudável
Leila	não	não	grande	sim	900	doente

Descarte de Atributos

- Descartar atributos claramente irrelevantes:

Febre	Enjôo	Mancha	Dor	Diagnóstico
sim	sim	pequena	sim	doente
não	não	pequena	não	saudável
sim	sim	grande	não	saudável
sim	não	pequena	sim	doente
sim	não	grande	sim	saudável
não	não	grande	sim	doente

Integração de BDs

- Dados podem ser oriundos de diferentes fontes
- Nesse caso, dados precisam ser consistentes
  - Identificar e remover instâncias duplicadas
    - gerenciando possíveis inconsistências de valores entre elas
  - Gerenciar atributos em comum e exclusivos
    - diferentes bases de dados podem ser descritas por conjuntos de atributos que podem diferir entre si em algum grau

Qualidade de Dados

- Dados a serem utilizados geralmente não são gerados com o propósito específico de mineração
  - em geral apresentam problemas de diversos tipos
- Algoritmos de DM precisam ou têm melhor desempenho na presença de dados "limpos"
  - problemas nos dados precisam ser minimizados

Qualidade de Dados

- Dados quase nunca serão ideais
- Problemas podem ocorrer nas medições e coleta de dados
- Causas:
  - Erros humanos
  - Falhas ou limitações do dispositivo de medição
  - Problemas no procedimento de coleta de dados

Qualidade de Dados

- Algumas Conseqüências:
  - **Valores errados**
    - p. ex. inconsistentes ou fora de faixa
      - pessoa gestante do sexo masculino
      - temperatura ambiente 200 graus Celsius
  - **Valores faltantes**
    - p. ex. não mensurados/coletados ou perdidos
      - indivíduo que se recusou a responder seu salário

Pré-Processamento e Limpeza

- Correção ou atenuação de erros nos dados
- Principais problemas:
  - Dados com ruído
    - no sentido amplo do termo
  - Dados incompletos ou ausentes
  - Dados duplicados
    - que não tenham sido eliminados na etapa de seleção
  - ...

Ruído

- Componente não-determinístico (aleatório) de algum tipo de erro
  - Erro randômico introduzido nos dados
- Formas
  - Distorção dos valores de atributos
  - Adição de instâncias espúrias

Ruído

- Normalmente é um problema que merece atenção em DM, por diferentes razões:
  - Nem sempre é possível ter certeza de que os dados estão contaminados com ruído
  - Mesmo em aplicações nas quais se tem esta certeza, muitas vezes não é possível eliminar ou mesmo reduzir o ruído

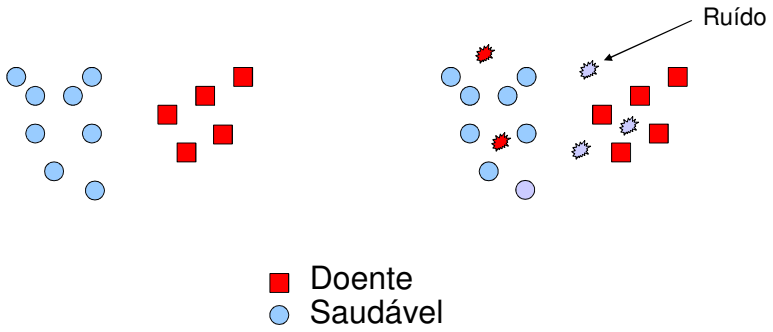
Ruído

- ... merece atenção, por diferentes razões :
  - Quando se pode eliminar ou reduzir o ruído, as técnicas são tipicamente específicas para cada domínio de aplicação em particular
    - p. ex. dados de imagens, de sensores variados, ...
  - Modelos gerados a partir de dados com ruído estão muito mais sujeitos a super-ajuste (overfitting)
    - exemplo: no quadro...

Ruído (Exemplo)

Dados sem ruído

Dados com ruído



Doente
Saudável

Valores Ausentes

- Não é raro uma instância não ter o valor de um ou mais atributos
- Algumas possíveis causas:
  - Desconhecimento do valor do atributo ou recusa em fornecê-lo na ocasião do preenchimento
  - Distração na ocasião do preenchimento
  - Inexistência de valor para o atributo em algumas instâncias
  - Problema com dispositivo / processo de coleta
  - ...

Tratamento de Valores Ausentes

Alternativas:

1. Descartar instâncias com atributos que apresentem valores ausentes
   - Simples e eficiente se as instâncias remanescentes ainda forem representativas
   - Proibitivo se parte significativa das instâncias possuírem ausentes

Tratamento de Valores Ausentes

Alternativas:

2. Descartar atributos com valores ausentes
   - Simples e eficiente se esses atributos não forem fundamentais para a solução do problema
     - Irrelevantes, ou
     - Redundantes a outros que não possuem valores ausentes
   - Proibitivo caso contrário

Tratamento de Valores Ausentes

Alternativas:

3. Modificar algoritmo para lidar com ausentes
   - Vários algoritmos de DM podem ser adaptados para lidar com valores ausentes
   - Por exemplo:
     - algoritmos puramente baseados em comparações de distâncias entre instâncias da base de dados

Tratamento de Valores Ausentes

Exemplo (Distância Euclidiana Normalizada entre I1 e I3):

Inst. /Atrib.	A1	A2	A3	A4
I1	2	-1	**???**	0
I2	7	0	-4	8
I3	**???**	3	5	2
I4	**???**	10	**???**	5

- no quadro...

- **Exercício**: calcule todas as demais distâncias !

Tratamento de Valores Ausentes

Alternativas:

4. Estimar (imputar) valores ausentes
   - Alternativa se outras não forem aplicáveis / eficazes
   - Existem várias técnicas
     - Área ativa de pesquisa!
   - Dentre as mais elementares tem-se, por exemplo:
     - Média (atributos numéricos) ou moda (atributos nominais) dos valores do atributo para instâncias da mesma classe

Tratamento de Valores Ausentes

- Exemplo (Temp. 1ª inst. e Dor 6ª inst.): no quadro...

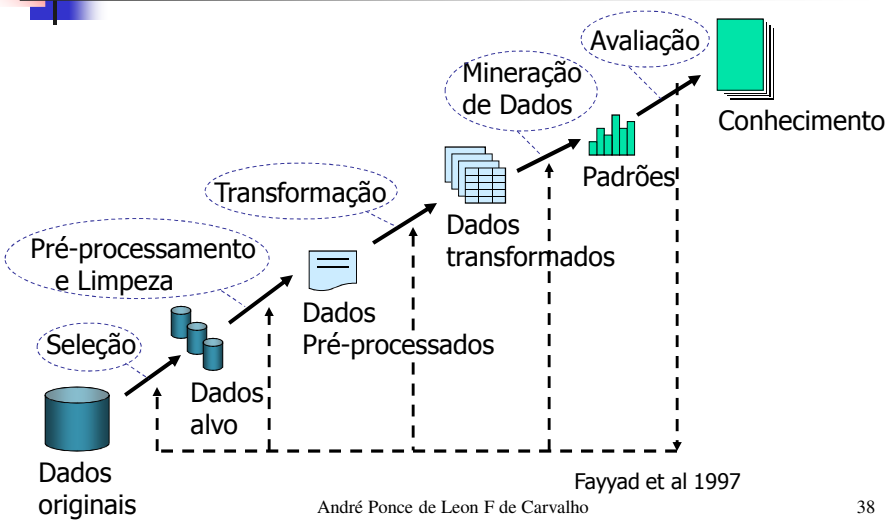
Temp.	Exame	Mancha	Dor	Diagnóstico
???	105	amorfa	sim	doente
37	155	ausente	não	saudável
37	175	ausente	não	saudável
38	135	amorfa	sim	doente
37	130	???	sim	saudável
40	120	circular	???	doente
37	90	amorfa	não	doente
37	???	circular	não	saudável

- **Exercício:** Exame 8ª instância e Mancha 5ª instância ?

Agregação

- Agregação é outra técnica de pré-processamento que também pode ser útil, com os seguintes propósitos:
  - Redução de dados
    - Reduz memória e tempo de processamento
    - Permite uso de algoritmos mais sofisticados
  - Dados mais estáveis
    - Dados agregados tendem a ter menos variabilidade
- Exemplo
  - Resumir preços horários (e.g. ações) ou transações unitárias (e.g. vendas) em uma média diária /mensal

Relembrando KDD...



Fayyad et al 1997

Transformação

- Uma vez que os dados estão limpos e pré-processados, segue a etapa de transformação:
  - Redução de Dimensionalidade
    - seleção de atributos / extração de características
    - veremos posteriormente no curso...
  - Discretizações e Conversões
    - próxima aula...
  - Normalizações
    - a seguir...

Transformação de Atributos

- Algumas vezes, os valores de um atributo numérico precisam ser ajustados
  - Algumas das possíveis razões são:
    - Grande variação de valores / escalas
      - evitar que um atributo predomine sobre outro...
      - a menos que isso seja importante
    - Propriedades estatísticas indesejadas
      - dados com distribuição não Gaussiana

Transformação de Atributos

- Por ex., análise de anomalias em rede de computadores pela avaliação da quantidade **x** de bytes movimentados em uma sessão
  - **x** varia de 1 a 1 bilhão
  - seções típicas de transferência de arquivos grandes movimentam entre 10^8 e 10^9 bytes
    - essas seções podem ser mais similares entre si do que seções que movimentem 10 e 1000 bytes, respectivamente
    - mas os dados brutos não mostram isso: $10^9 - 10^8 >> 10^3 - 10$
  - possível solução: $\mathbf{log_{10}\ x} \Rightarrow 9 - 8 < 3 - 1$

Transformação de Atributos

- Aplicada a todos os valores de um atributo
  - Valores correspondentes a todas as instâncias
  - Exemplo simples:
    - supor que apenas a magnitude do atributo é importante
    - transformação: valor absoluto
      - - 4, 5 e –2 se tornam 4, 5 e 2
  - Classe mais comum
    - Normalização

Normalização

- Transformação que, quando aplicada de maneira independente a dois ou mais atributos, faz com que eles exibam alguma propriedade em comum
- No contexto de mineração de dados, **normalizações lineares** são mais usuais:
  - **Re-escalar**
  - **Padronizar**

Re-Escalar

- Re-escalar os valores de um atributo:
  1. Adicionar ou subtrair uma constante
  2. Multiplicar ou dividir por uma constante
- Utilizado para mudar unidade de medida dos dados
- Permite converter todos os valores de um atributo para o intervalo [0, 1] ou [-1,+1]

$$x' = \frac{(x - \min_x)}{(\max_x - \min_x)}$$

Re-Escalar

- Propriedade comum ao re-escalar atributos:
  - Valores mínimos e máximo (escalas) iguais
- Muito usada em regressão (e.g. redes neurais)
  - dentre outras razões para evitar problemas numéricos
- É muito sensível a valores incomuns com grande magnitude (outliers)
  - por isso evitado, dependendo da natureza da aplicação

Padronizar

- Padronizar os valores de um atributo:
  1. Adicionar ou subtrair uma medida de localização
  2. Multiplicar ou dividir por uma medida de escala
- Se os valores têm uma distribuição Gaussiana
  - Subtrair a média (μ)
  - Dividir pelo desvio padrão (σ)
  - Produz distribuição normal padrão: N(0,1)
  - Denomina-se normalização **score-z**

Padronizar

- Normalização score-z:

$$x' = \frac{x - \mu_x}{\sigma_x}$$

- Propriedade comum ao padronizar atributos:
  - médias e variâncias iguais
- Muito utilizada em tarefas de DM que demandam cálculo de distâncias

Exemplo

- Seja um conjunto de dados com atributos salário e idade
  - Diferenças em salário serão bem maiores que diferenças em idade
    - Isso influencia o resultado de técnicas de DM que usam informação sobre diferenças (p. ex. distância Euclidiana)
  - Se essa discrepância não deve ser refletida pelo algoritmo de DM, atributos devem ser padronizados
    - questão de projeto: ciclo de KDD

Exercício

- Converter os seguintes valores numéricos utilizando re-escala e padronização
  - [0, 1] e N(0,1)

Valores	Re-escala	Padronização
3		
9		
5		
11		
5		
7		