

MATB16

Laboratório de Inteligência Artificial

PRÉ-PROCESSAMENTO

Tatiane Nogueira Rios
Ricardo Araújo Rios

LabIA
Instituto de Computação - UFBA



Objetivo

- Apresentar conceitos necessários para execução de pré-processamento e preparação de dados para aplicação de métodos de AM.



Agenda

- Introdução
- Limpeza
- Exploração
- Transformação



Introdução

- Dados de entrada
 - Coleção de instâncias com seus atributos
- Instâncias
 - padrões, exemplos, objetos, registros, pontos, amostras, casos, entidades



Introdução

- Atributos
 - Características e/ou propriedades que compõem uma instância
 - valores específicos dos atributos descrevem uma instância particular



Introdução

- Exemplo

Dados do paciente 1	
Idade	67 anos
Sexo	Masculino
Tipo de dor no peito	Assintomática
Pressão arterial de repouso (diastólica)	160mmHg
Colesterol no sangue (sérico)	286mg/ dl
Nível de glicose no sangue	>120mg/ dl



Introdução

- Exemplo

Instância

Dados do paciente 1	
Idade	67 anos
Sexo	Masculino
Tipo de dor no peito	Assintomática
Pressão arterial de repouso (diastólica)	160mmHg
Colesterol no sangue (sérico)	286mg/ dl
Nível de glicose no sangue	>120mg/ dl



Introdução

- Exemplo

Dados do paciente 1		Atributo
Idade		67 anos
Sexo		Masculino
Tipo de dor no peito		Assintomática
Pressão arterial de repouso (diastólica)		160mmHg
Colesterol no sangue (sérico)		286mg/ dl
Nível de glicose no sangue		>120mg/ dl



Introdução

- Exemplo

Dados do paciente 1		Atributos
I dade	67 anos	
Sexo	Masculino	
Tipo de dor no peito	Assintomática	
Pressão arterial de repouso (diastólica)	160mmHg	
Colesterol no sangue (sérico)	286mg/ dl	
Nível de glicose no sangue	>120mg/ dl	



Introdução

- Matriz Atributo x Valor

Exemplo	Idade	Sexo	Dor	Pressão	Colesterol	Glicose
Paciente1						
Paciente 2						
...						

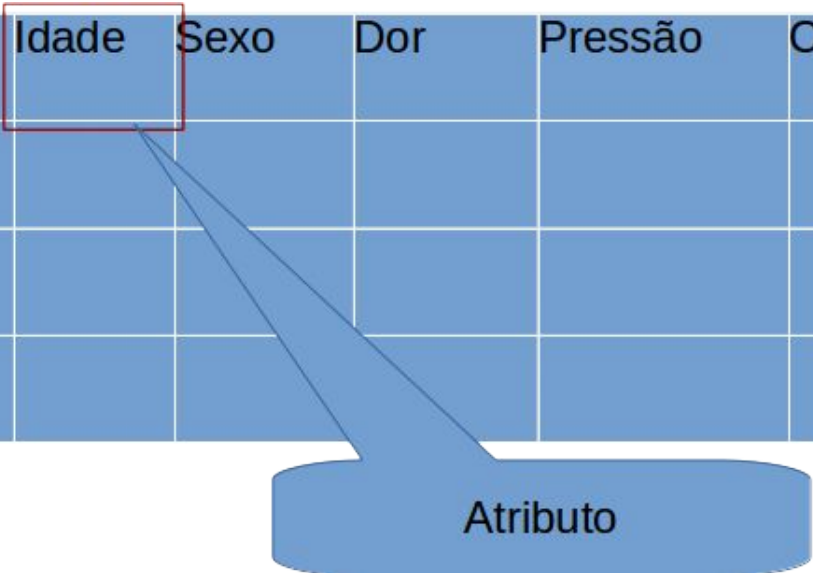
Instância



Introdução

- Matriz Atributo x Valor

Exemplo	Idade	Sexo	Dor	Pressão	Colesterol	Glicose
Paciente1						
Paciente 2						
...						



Atributo



Introdução

- Matriz Atributo x Valor

Exemplo	Idade	Sexo	Dor	Pressão	Colesterol	Glicose
Paciente1						
Paciente 2						
...						

Valor de atributo



Introdução


- Matriz Atributo x Valor - supervisionado

Exemplo	Idade	Sexo	Dor	Pressão	Colesterol	Glicose	Saída
Paciente1	67	M	Assint.	160	286	>120	Doente
Paciente 2	50	F	Sint.	150	200	<120	Não Doente
...
Paciente n	40	M	Assint.	120	150	>120	?



Introdução

- Matriz Atributo x Valor - não-supervisionado

Exemplo	Idade	Sexo	Dor	Pressão	Colesterol	Glicose	
Paciente1	67	M	Assint.	160	286	>120	
Paciente 2	50	F	Sint.	150	200	<120	
...	
Paciente n	40	M	Assint.	120	150	>120	

Rótulo
desconhecido!



Limpeza de dados

- Nem sempre é possível aplicar técnicas de AM diretamente sobre os dados
- Pré-processamento:
 - Eliminação manual;
 - Integração;
 - Amostragem;
 - Balanceamento;
 - Limpeza;
 - Transformação e redução da dimensionalidade;



Limpeza de dados

- Seleção adequada de atributos
- Atributo não existe para todas as instâncias
 - Base de Dados: Meios de Transporte
 - Atributo: Quantidade de pneus
 - Instância: Navios



Limpeza de dados

- Seleção adequada de atributos
 - A existência de um atributo é dependente de outro:
 - “nome da(o) esposa(a)” e “casado (sim/não)”



Limpeza de dados

- Seleção adequada de atributos
 - Atributo irrelevante para o que deseja-se aprender
 - “nome do paciente”



Limpeza de dados

- Tipo dos atributos
 - Numérico (contínuo)
 - Pressão arterial, colesterol, ...
 - Nominal (categórico)
 - Nome, sexo, ...



Limpeza de dados

- Escala dos atributos
 - Nominal
 - ==, !=
 - Cor, identificação, profissão, ...
 - Ordinal
 - <, >, <=, >=
 - Dias da semana, Intensidade (alta, média, baixa), ...



Limpeza de dados

- Escala dos atributos
 - Intervalar
 - Temperatura em Celsius, ...
 - Racional
 - Peso, tamanho, idade, ...



Limpeza de dados

- Dados são obtidos de diferentes fontes
 - Ex: Ao analisar um supermercado, os dados são provenientes de diferente setores: vendas, serviço, contas, ...
 - Problemas nas medições e coletas dos dados: diferentes departamentos armazenam informações com diferentes valores de atributos, períodos, chaves primária, ...



Limpeza de dados

- Para obter um bom conjunto de dados é preciso:
 - Coletar dados de diferentes domínios;
 - Integrar;
 - Limpar;



Limpeza de dados

- Consequências de um conjunto de dados ruim:
 - Valores errados
 - Valores ausentes
 - Instâncias duplicadas



Limpeza de dados

- Ruído
 - Erro aleatório introduzido nos dados
 - Distorção dos valores de atributos



Limpeza de dados

- Valores inconsistentes
 - Altura do paciente: -1.97m ou 4m
 - Adulto com 5kg
 - Temperatura de um ambiente = 300 °C
 - CEP: 00000-000
 - Inconsistências produzidas ao acaso podem ser consideradas ruído.



Limpeza de dados

- Valores ausentes
 - Atributo não era monitorado quando os primeiros dados foram coletados
 - Distração/Recusa ao fornecer uma informação
 - Inexistência para certas instâncias
 - Erro ou limitação do equipamento de medição



Limpeza de dados

- Valores ausentes
 - Tratamento
 - Descarte
 - Vantagem: evita introdução de erros
 - Desvantagem: pode comprometer a qualidade da modelagem



Limpeza de dados

- Valores ausentes
 - Tratamento
 - Estimar valores ausentes
 - Nominiais: moda
 - Contínuos: média
 - Temporais: splines
 - Bons resultados com poucos valores ausentes
 - Erro de estimação pode ser acumulativo → problemas com muitos dados ausentes



Limpeza de dados

- Instancias duplicadas
 - Instâncias idênticas ou que não diferem significativamente para o domínio do problema
 - **Ilegítimas:** cadastro duplicado de um cliente devido a pequenas diferenças na representação do nome
 - **Legítimas:** dois pacientes com as mesmas características



Limpeza de dados

- Outliers
 - Instâncias “anômalas”, i.e., possuem características (valor de um ou mais atributos) diferentes da maioria dos demais
 - Definição de “diferente” usualmente é estatística



Limpeza de dados

- Conheça seus dados!
 - Uso de ferramentas são úteis e importantes, mas uma simples olhada em uma planilha pode fazer toda a diferença.



Limpeza de dados

- Exemplo de análises simples e importantes
 - Se um atributo numérico apresenta somente 6 valores bem separados, então provavelmente este deveria ser um atributo categórico.
 - Se todos os valores de um atributo são idênticos, então este atributo poderia ser descartado.



Limpeza de dados

- Exemplo de análises simples e importantes
 - Se todos os valores de um atributo são idênticos e somente um é diferente, é preciso decidir se este valor representa um ruído, ou se o atributo apresenta dois valores nominais.



Limpeza de dados

- Alguns valores podem estar fora do intervalo de valores esperados para o atributo
 - Ex: Se o intervalo for $[200, 5000]$, o valor 22654,8 está fora deste intervalo, mas pode ter sido apenas um erro de digitação ocasionado pela repetição do primeiro valor.



Limpeza de dados

- Alguns valores podem estar fora do intervalo de valores esperados para o atributo
 - Ex: No mesmo intervalo, o valor 38597 também está fora deste intervalo, mas pode ter sido apenas um erro de digitação em que faltou digitar o ponto decimal.



Limpeza de dados

- Alguns valores podem estar fora do intervalo de valores esperados para o atributo
 - Se estes dados foram coletados automaticamente, deve ser considerado o mal funcionamento do equipamento.



Limpeza de dados

- Pode ser observada a entrada de valores em quantidades anormais
 - Ex: Ao preencher um formulário na web, usuários podem apenas selecionar a primeira opção para “país”, gerando um valor default que não representa a realidade



Limpeza de dados

- Pode ser observada a entrada de valores em quantidades anormais
 - Ex: Valor default para data de nascimento, ao preencher formulário e o atributo idade não corresponder ao esperado (ex: mais idosos do que idosos respondendo à uma pesquisa direcionada à estudantes)



Limpeza de dados

Data cleaning is a time-consuming and labor-intensive procedure, but one that is absolutely necessary for successful data mining. With a large dataset, people often give up—how can they possibly check it all?

Instead, you should sample a few instances and examine them carefully. You'll be surprised at what you find. Time looking at your data is always well spent (Witten et. Al, 2011).



Transformação

- Várias técnicas de AM são limitadas ao tipo dos atributos: apenas valores numéricos ou apenas valores simbólicos
 - RNAs e SVMs são exemplos de técnicas que lidam apenas com dados numéricos
 - Solução: conversão de valores



Transformação

- Conversão Simbólico-Numérico
 - Atributo nominal com dois valores que representam presença ou ausência de uma característica
 - Substituir por um dígito binário
 - Atributo nominal com dois valores que representam relação de ordem
 - Substituir por um dígito binário



Transformação

- Conversão Simbólico-Numérico
 - Atributo nominal com mais de dois valores
 - Abordagem 1: Sequência de bits, em que cada valor possível de atributo possui apenas 1 bit com o valor 1 e os demais com valor 0
 - Problema: Dependendo dos valores nominais, a sequência binária pode ficar muito longa.
 - Abordagem 2: Pseudoatributos



Transformação

- Conversão Simbólico-Numérico
 - Atributo ordinal com mais de dois valores
 - A nova codificação deve preservar a relação de ordem
 - Abordagem 1: ordenar os valores categóricos ordinais e codificar cada valor com sua posição na ordem
 - Ex: Primeiro=1, Segundo=2, Terceiro=3, ...



Transformação

- Conversão Numérico-Simbólica
 - Uma parcela dos algoritmos de classificação e de associação foram desenvolvidos para trabalhar com valores qualitativos
 - Atributo quantitativo do tipo discreto ou binário, com apenas dois valores
 - Conversão trivial: associar um nome a cada valor



Transformação

- Conversão Numérico-Simbólica
 - Atributos quantitativos numéricos
 - Discretização: transformação de valores numéricos em intervalos
 - Existem vários métodos de discretização, o mais simples é a média.



Transformação

- Transformação de atributos numéricos
 - Transformar um valor numérico em outro valor numérico
 - Isso ocorre quando os limites inferior e superior de valores dos atributos são muito diferentes
 - Ou quando vários atributos estão em escalas diferentes
 - A transformação é necessária para evitar que um atributo predomine sobre outro



Transformação

- Normalização (0-1)

$$\hat{X} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

