



Pré-processamento de Dados

Terminologia
Limpeza de Dados
Conversão de Dados

Huei Diana Lee

Inteligência Artificial
CECE/UNIOESTE-FOZ

Terminologia



Terminologia

Conceitos:

Tipos de “coisas” que se pode aprender

Exemplos:

- Objetos
- Casos
- Ocorrências específicas do conceito

Atributos:

- Características
- Campos
- Variáveis

Valores de atributos

Podem ser definidos por:

- Tipo

Grau de quantização nos dados

- Escala

Significância relativa dos valores



Conhecer o tipo/escala dos atributos auxilia a identificar o modo adequado de preparar os dados e posteriormente modelá-los

Tipos de atributos

Quantitativo (numérico)

- Representa quantidades
- Valores podem ser ordenados e usados em operações aritméticas
- Podem ser **contínuos** ou **discretos**
- Possuem unidade associada

Qualitativo (simbólico ou categórico)

- Representa qualidades
- Valores podem ser associados a categorias
- Alguns podem ser ordenados, mas operações aritméticas não são aplicáveis
- *Ex. {pequeno, médio, grande}*

Tipos de atributos

Atributos Quantitativos

Contínuos

- Podem assumir um número infinito de valores
- Geralmente resultados de medidas
- Frequentemente representados por números reais
- *Ex. peso, distância*

Discretos

- Número finito ou infinito contável de valores
- Caso especial: atributos binários (booleanos)
- Ex. {12, 23, 45}, {0, 1}

Tipos de atributos

Ex. Conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Qualitativo

Quantitativo discreto

Quantitativo contínuo

Tipos de atributos

Ex. Conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Qualitativo

Qualitativo discreto

Quantitativo contínuo

Observar atributo **Peso**

Tipos de atributos

Ex. Conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Alguns atributos qualitativos são representados por números, mas não faz sentido a utilização de operadores aritméticos sobre seus valores

Escala de atributos

Define operações que podem ser realizadas sobre os valores dos atributos:

- Nominais
- Ordinais
- Intervalares
- Racionais

Escalas de atributos

- Define operações que podem ser realizadas sobre os valores dos atributos

- Nominais

- Ordinais

- Intervalares

- Racionais

Qualitativos

Escalas de atributos

- Define operações que podem ser realizadas sobre os valores dos atributos
 - Nominiais
 - Ordinais
 - Intervalares
 - Racionais

Quantitativos

Escalas de atributos

Escala nominal

- Valores são nomes diferentes e carregam a menor quantidade de informação possível
- Não existe relação de ordem entre os valores
- **Operações aplicáveis:** =, ≠
- *Ex.: número de conta em banco, cores, sexo*

Escala ordinal

- Valores refletem ordem das categorias representadas
- Operações aplicáveis: =, ≠, <, >, ≥, ≤
- Ex.: hierarquia militar, avaliações qualitativas de temperatura

Escalas de atributos

Escala intervalar

- Números que variam em um intervalo
- É possível definir ordem e diferença em magnitude entre dois valores
- Origem da escala definida de maneira arbitrária
- **Operações aplicáveis:** $=$, \neq , $<$, $>$, \leq , \geq , $+$, $-$
- *Ex.: temperatura em $^{\circ}\text{C}$ ou $^{\circ}\text{F}$, datas*

Escala racional

- Carregam mais informações
- Têm significado absoluto (existe 0 absoluto)
- Razão tem significado
- **Operações aplicáveis:** $=$, \neq , $<$, $>$, \leq , \geq , $+$, $-$, $*$, $/$
- *Ex.: tamanho, distância, salário, saldo em conta*

Escalas de atributos

Ex. conjunto de dados `hospital`

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Grandes	38,0	2	SP	Doente
3217	Maria	18	F	67	Pequenas	39,5	4	MG	Doente
4039	Luiz	49	M	92	Grandes	38,0	2	RS	Saudável
1920	José	18	M	43	Grandes	38,5	20	MG	Doente
4340	Cláudia	21	F	52	Médias	37,6	1	PE	Saudável
2301	Ana	22	F	72	Pequenas	38,0	3	RJ	Doente
1322	Marta	19	F	87	Grandes	39,0	6	AM	Doente
3027	Paulo	34	M	67	Médias	38,4	2	GO	Saudável

Nominal

Ordinal

Intervalar

Racional

Tipos de atributos

- Por que especificar os tipos de atributos?
 - Para que comparações e aprendizado de conceitos sejam feitos corretamente
 - **Tempo** > “ensolarado” não faz sentido, enquanto que **Umidade** > **70** faz sentido
- Usos adicionais para tipos de atributos:
 - Verificar validade de valores
 - Tratar valores faltantes
 - Entre outros

Limpeza de Dados



Limpeza de Dados

Aquisição de dados

Valores faltantes

Formato unificado de datas

Conversão de nominais para
numéricos

Detecção de duplicados

Limpeza de Dados

Aquisição

- Dados podem estar em SGBD
- Dados em arquivos texto (flat file)
 - Formato delimitado: tab, vírgula e outros
 - Por exemplo: C4.5 (.data) e Weka (.arff) usam dados delimitados por vírgulas

Limpeza de Dados

Reformatação

- Converter os dados para o formato padrão (ex. arff, data ou csv)
- Tratar valores faltantes (VF) (*Missing values*)
- Tratar *outliers*
- Converter valores nominais ordenados para valores numéricos

Limpeza de Dados

Significados para o Termo VF

- Faltantes de modo randômico:
 - Em geral é o melhor caso
 - Usualmente não são verdadeiros
- Faltantes de modo não randômico
- Presupostos como valores normais e portanto não mensurados
- Faltante por casualidade:
 - Por causa de valores de outros atributos ou por causa do valor do atributo meta

Limpeza de Dados

Por que VF
existem?

Defeito de equipamentos

Mensurações incorretas

Dados de censos ou dados anônimos

Falta de preenchimento manual de
dados

- Bastante frequente em questionários para cenários médicos
- Muito baixa frequência de valores faltantes pode ser suspeito

Limpeza de Dados

Por que VF
são
Importantes?

- **Perda de eficiência:**
Menos padrões são extraídos ou conclusões são estatisticamente menos fortes
- **Complicações na manipulação e análise de dados:**
Em geral, os métodos não estão preparados para tratar valores faltantes
- ***Bias* resultante da diferença entre valores faltantes e completos:**
Métodos de Mineração de Dados geram modelos diferentes

Limpeza de Dados

Estratégias para Tratar VF

Descartar exemplos com valores faltantes:

- Estratégia mais simples
- Permite o uso de métodos sem modificá-los
- Funciona se há poucos exemplos com valores faltantes, caso contrário pode-se introduzir *bias*

Limpeza de Dados

Estratégias para Tratar VF

Converter os valores faltantes em novos valores:

- Usar um valor especial para isso
- Adicionar um atributo que identifica se o valor é faltante ou não
- Aumenta bastante a dificuldade de se realizar o processo de Mineração de Dados

Limpeza de Dados

Estratégias para Tratar VF

Métodos de imputação:

- Atribui um valor para o faltante baseado no restante do conjunto de dados
- Permite o uso de métodos sem modificá-los

**Limpeza de
Dados**

**Imputação de
Dados**

Adequado para
valores faltantes de
modo randômico

Não adequado para
valores faltantes de
modo não randômico

Limpeza de Dados

Do Not Impute (DNI)

- Simplesmente use a estratégia de VF do algoritmo
- Adequado somente se tal estratégia existe
- Exemplo para aprendizado de regras:
Atributos com VF seriam considerados irrelevantes

Limpeza de Dados

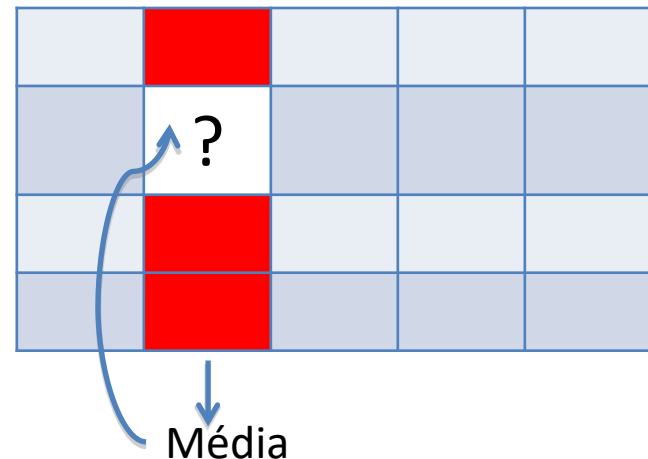
Random Imputation

- Predizer VF e adicionar componente de erro escolhido de modo randômico
- Repetir diversas vezes para melhorar a estimativa do erro

Limpeza de Dados

Most Common (MC) value

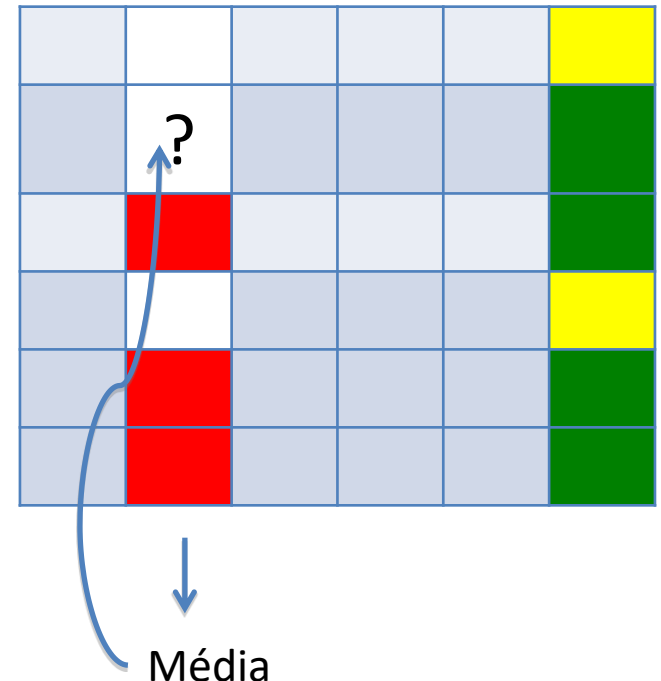
- Se os VF são:
 - Contínuos, substituir pela média
 - Discretos, substituir pela moda
- Simples e rápido de ser computado
- Assume que cada variável apresenta distribuição normal



Limpeza de Dados

Concept Most Common (CMC) value

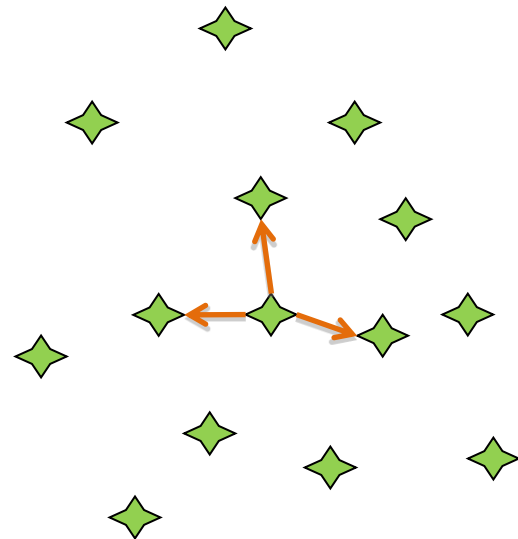
- Refinamento da estratégia MC
- O VF é substituído pela média/moda computada a partir dos exemplos *pertencentes à mesma classe*
- Assume que a distribuição de um atributo para todos os exemplos da mesma classe é normal




Limpeza de Dados

Imputação com *k-Nearest Neighbour* (KNNI)

- Selecionar os k vizinhos mais próximos
- Substituir os VF com a média/moda desses k exemplos



Conversão de Dados



Conversão de Nominal para Numérico

Alguns algoritmos tratam internamente valores nominais

Outros métodos requerem apenas valores numéricos como entrada (RNA, kNN, Regressão)

Conversão Binário para Numérico

- Campos Binários como gênero M e F
- Converter para valores 0 e 1
 - Gender = M \rightarrow Gender_0_1 = 0
 - Gender = F \rightarrow Gender_0_1 = 1

Conversão Ordenado para Numérico

- Atributos ordenados, como Nota, podem ser convertidos para números preservando a ordem natural
 - A \rightarrow 4.0
 - A- \rightarrow 3.7
 - B+ \rightarrow 3.3
 - B \rightarrow 3.0
- Porque é importante preservar a ordem natural?

Conversão Ordenado para Numérico

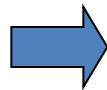
Ordem natural permite comparações com significado, por exemplo, Nota > 3.5

Conversão Nominal com Poucos Valores

Atributos multivalorados desordenados com poucos valores podem ser transformados para binários (*rule of thumb* < 20)

- Exemplo, Color=Red, Orange, Yellow, ..., Violet
- Para cada valor, criar uma “flag” binária em que 1 está presente e 0 caso contrário

ID	Color	...
371	red	
433	yellow	



ID	C_red	C_orange	C_yellow	...
371	1	0	0	
433	0	0	1	

Conversão Nominal com Muitos Valores

- Exemplos:
 - *US State Code* (50 valores)
 - *Profession Code* (7,000 valores, mas apenas poucos frequentes)
- Como tratar:
 - Ignorar os Id-like cujos valores são únicos para cada registro
 - Para os outros campos, agrupar naturalmente os valores:
 - 50 *US States* → 3 ou 5 regiões
 - *Profession Code* → selecionar as mais frequentes e agrupar o restante
- Criar flags binárias para valores selecionados

Transformando Ordinal para Booleano

- Codificar n valores em n-1 atributos booleanos
- Exemplo: atributo “temperatura”

Original data

Temperature
Cold
Medium
Hot



Transformed data

Temperature > cold	Temperature > medium
False	False
True	False
True	True

- Alguns slides foram baseados em apresentações de:
 - Profa. Huei Diana Lee
 - Profa. Maria Carolina Monard
 - Prof. Ronaldo Cristiano Prati.
 - Prof. Walter Nagai
 - Prof. E. Keogh
 - Prof. Nitin Patel
 - Prof. José Augusto Baranauskas
 - Prof. Gustavo E.A.P.A. Batista
 - Prof. Patrick H. Winston
 - Profa. Ana Carolina Lorena
 - Prof. André C. P. L. F. Carvalho
 - Prof. Ricardo Campello
 - Profa. Solange O. Rezende
 - Prof. Marcilio C. P. Souto
 - Prof. Carlos Soares
 - Prof. Paulo Horst
 - Profa. Aurora Trinidad Ramirez Pozo