UNIVERSIDADE DE SÃO PAULO

ESCOLA POLITÉCNICA

Ciência dos Dados

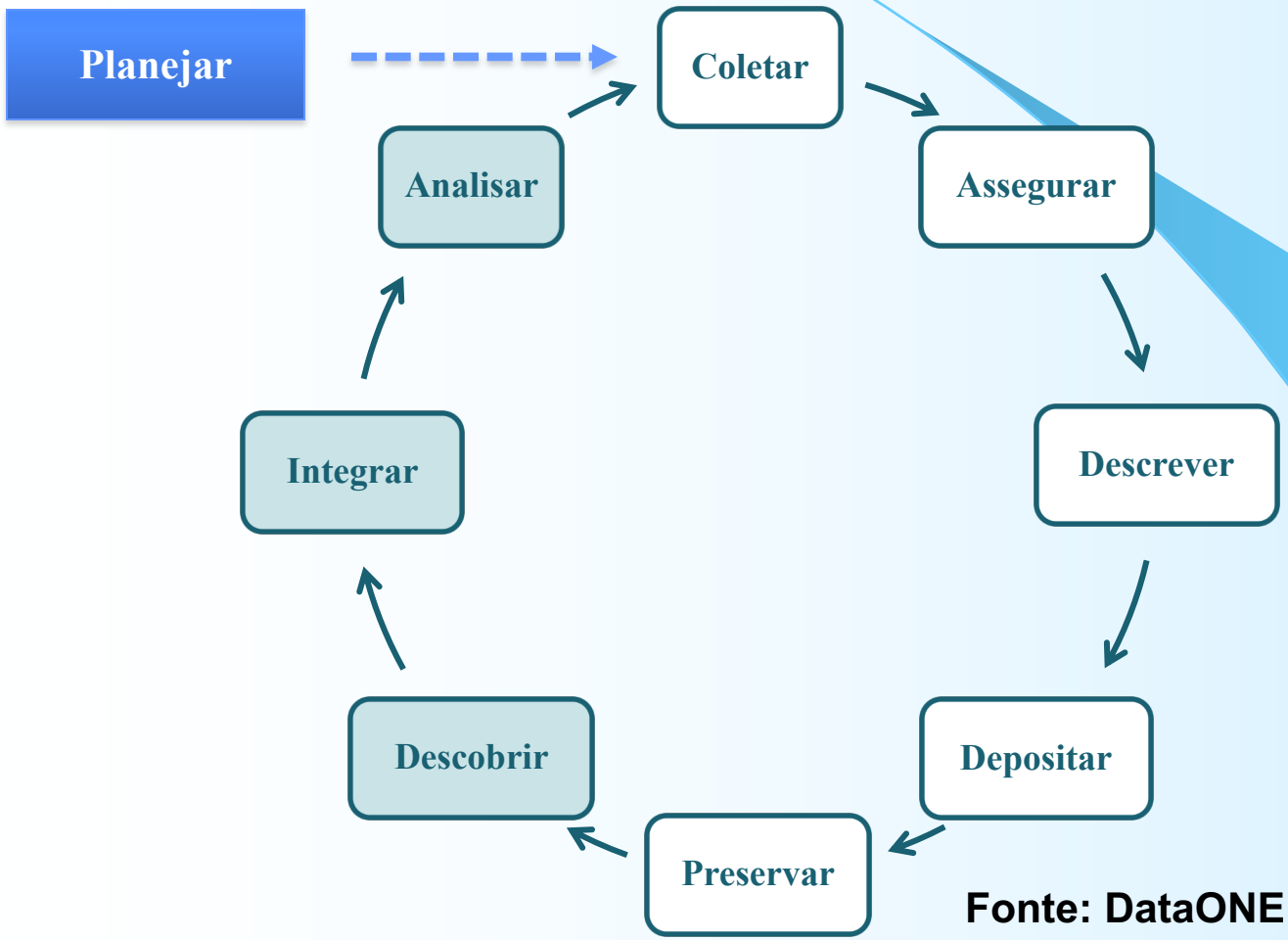
PCS5787
Seleção dos Dados

Pós-Graduação em Engenharia Elétrica

2o. Semestre de 2020

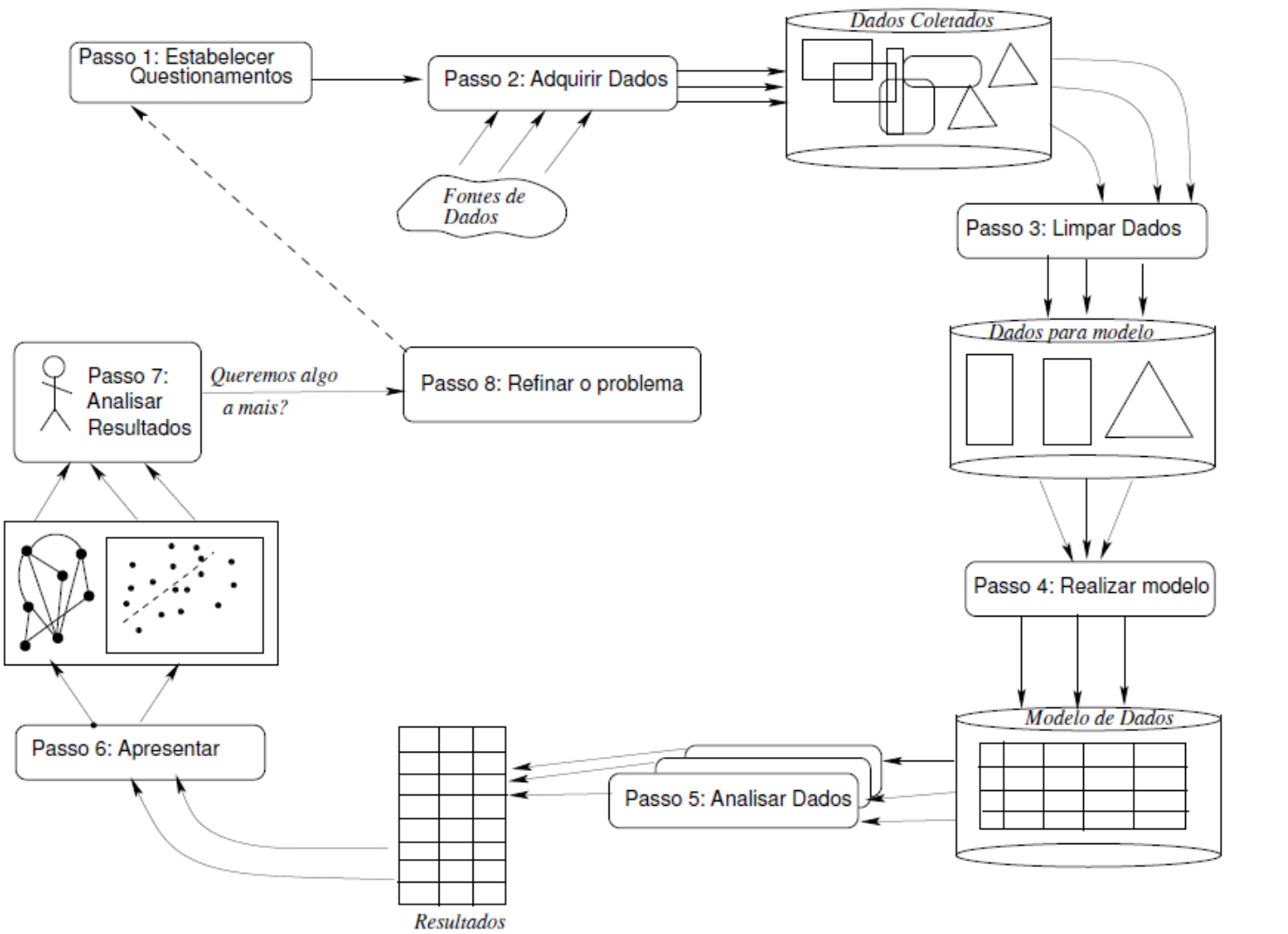
Introdução – Apoio as decisões tomadas desde o Planejamento até Integração

Ciclo de Vida dos Dados – DataONE



Fonte: DataONE Best Practices

Introdução – Passos 2 e 3



Cortesia: Alex Dehktyar

Pré-processamento

- Desempenho de técnicas de Análise é afetado pela qualidade dos dados
  - Conjuntos de dados podem ter diferentes características, dimensões ou formatos
    - Atributos numéricos vs simbólicos
    - Limpos vs com ruídos, imperfeições e incompletos
      - Valores incorretos, inconsistentes, duplicados ou ausentes
    - Atributos independentes vs relacionados
    - Poucos vs muitos objetos e/ou atributos

Pré-processamento: minimizar/eliminar problemas nos dados; tornar dados mais adequados para uso por um determinado algoritmo de AM

Pré-processamento

- Benefícios:
  - Facilitar o posterior uso de técnicas de Análise
    - Ou tornar mais adequado para a técnica
      - Ex. algumas trabalham somente com entradas numéricas
  - Obtenção de modelos mais fiéis à distribuição dos dados
    - Melhorar qualidade da Análise
  - Redução de complexidade computacional
    - Tempo e custo

Pré-processamento

Grupos de tarefas de pré-processamento:

- **Eliminação manual de atributos**
- **Integração de dados**
- **Amostragem de dados**
- **Redução de dimensionalidade**
- **Balanceamento de dados**
- **Limpeza de dados**
- **Transformação de dados**

Observação: **não existe ordem fixa para aplicação das diferentes técnicas de pré-processamento**

Pré-processamento

Grupos de tarefas de pré-processamento:

- **Eliminação manual de atributos**
- **Integração de dados**
- **Amostragem de dados**
- **Redução de dimensionalidade**
- **Balanceamento de dados**
- **Limpeza de dados**
- **Transformação de dados**

Alguns atributos não possuem relação com o problema sendo solucionado

Ex. RG em diagnóstico

Pré-processamento

Grupos de tarefas de pré-processamento:

- **Eliminação manual de atributos**
- **Integração de dados**
- **Amostragem de dados**
- **Redução de dimensionalidade**
- **Balanceamento de dados**
- **Limpeza de dados**
- **Transformação de dados**

***Diferentes conjuntos de dados integrados*: podem levar a inconsistências e redundâncias**

Pré-processamento

Grupos de tarefas de pré-processamento:

- **Eliminação manual de atributos**
- **Integração de dados**
- **Amostragem de dados**
- **Redução de dimensionalidade**
- **Balanceamento de dados**
- **Limpeza de dados**
- **Transformação de dados**

Algoritmos de AM podem ter dificuldades quando precisam lidar com uma grande quantidade de dados (objetos, atributos ou ambos)

Ex. redundância e inconsistência

Pré-processamento

Grupos de tarefas de pré-processamento:

- **Eliminação manual de atributos**
- **Integração de dados**
- **Amostragem de dados**
- **Redução de dimensionalidade**
- **Balanceamento de dados**
- **Limpeza de dados**
- **Transformação de dados**

Conjunto de dados desbalanceado: proporção de exemplos em algumas classes pode ser muito maior do que em outras

Alguns algoritmos de AM têm dificuldade neste cenário

Pré-processamento

Grupos de tarefas de pré-processamento:

- **Eliminação manual de atributos**
- **Integração de dados**
- **Amostragem de dados**
- **Redução de dimensionalidade**
- **Balanceamento de dados**
- **Limpeza de dados**
- **Transformação de dados**

Presença de ruídos, dados incompletos e inconsistentes pode afetar desempenho dos algoritmos de AM

Alguns são incapazes de lidar com dados incompletos

Pré-processamento

Grupos de tarefas de pré-processamento:

- **Eliminação manual de atributos**
- **Integração de dados**
- **Amostragem de dados**
- **Redução de dimensionalidade**
- **Balanceamento de dados**
- **Limpeza de dados**
- **Transformação de dados**

Vários algoritmos de Análise têm dificuldades em usar os dados em seu formato original

Ex. transformação de valores simbólicos para numéricos

Integração de Dados

- Dados podem vir de diferentes fontes
  - integração de diferentes conjuntos de dados
    - Cada um pode ter atributos diferentes para os mesmos objetos

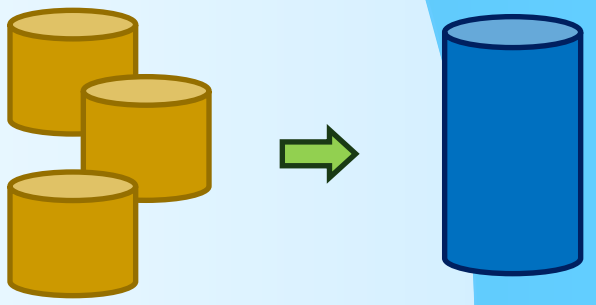
- Identificação de entidade
  - Identificar os objetos em comum
    - Normalmente por busca por atributos comuns nos conjuntos
      - Que tenham valor único para cada objeto
      - Ex1. identificação de paciente,
      - Ex2: eventos/objetos relacionados pelo tempo e espaço (localização).

Integração de Dados

- Dificuldades:
  - Atributos correspondentes com nomes diferentes
  - Dados podem ter sido atualizados em momentos diferentes (lembrar das etapas do Ciclo de Vida dos Dados).

Comum usar metadados para minimizar esses problemas

Metadados: dados sobre os dados,
que descrevem suas principais características



Integração de Dados

Do ponto de vista de Banco de Dados

– Heterogeneidade dos esquemas:

- Heterogeneidade estrutural:
  - Conflito de tipos;
  - Conflito de chaves;
  - Conflito comportamental.
- Heterogeneidade semântica:
  - Identificação de Sinônimos, homônimos;
  - Diferentes ontologias (contextos) – correlações espaço/temporal

Eliminação manual de atributos

- Há atributos que claramente não contribuem para o análise
  - 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

Não contribuem para estimar se um paciente tem doença ou não

Eliminação manual de atributos

- Normalmente, o conjunto de atributos é definido de acordo com a experiência de especialista
  - Ex. conjunto de dados `hospital`

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
28	M	79	Grandes	38,0	2	SP	Doente
18	F	67	Pequenas	39,5	4	MG	Doente
49	M	92	Grandes	38,0	2	RS	Saudável
18	M	43	Grandes	38,5	20	MG	Doente
21	F	52	Médias	37,6	1	PE	Saudável
22	F	72	Pequenas	38,0	3	RJ	Doente
19	F	87	Grandes	39,0	6	AM	Doente
34	M	67	Médias	38,4	2	GO	Saudável

Médico pode decidir que atributo associado ao estado de origem do paciente também não é relevante para seu diagnóstico clínico

Eliminação manual de atributos

- Ex. conjunto de dados `hospital`
  - Após eliminação manual dos atributos

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
28	M	79	Grandes	38,0	2	Doente
18	F	67	Pequenas	39,5	4	Doente
49	M	92	Grandes	38,0	2	Saudável
18	M	43	Grandes	38,5	20	Doente
21	F	52	Médias	37,6	1	Saudável
22	F	72	Pequenas	38,0	3	Doente
19	F	87	Grandes	39,0	6	Doente
34	M	67	Médias	38,4	2	Saudável

Eliminação manual de atributos

- Outro atributo irrelevante facilmente detectado:
  - Atributo que possui o mesmo valor para todos objetos
    - Não traz informação para ajudar a distinguí-los
- Há ainda atributos irrelevantes de identificação não tão clara
  - Técnicas de seleção de atributos podem ajudar a identificar

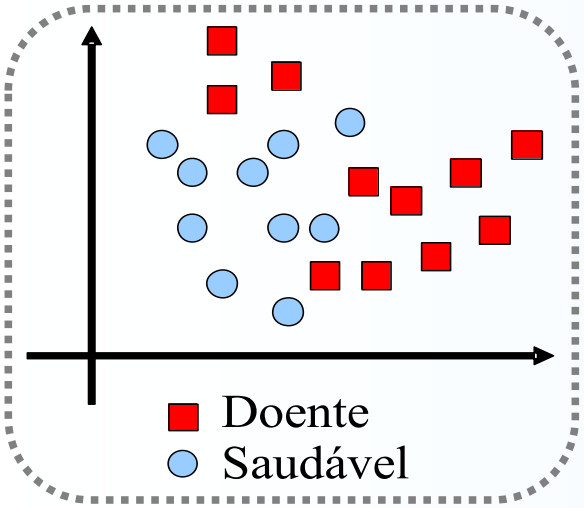
Amostragem de dados

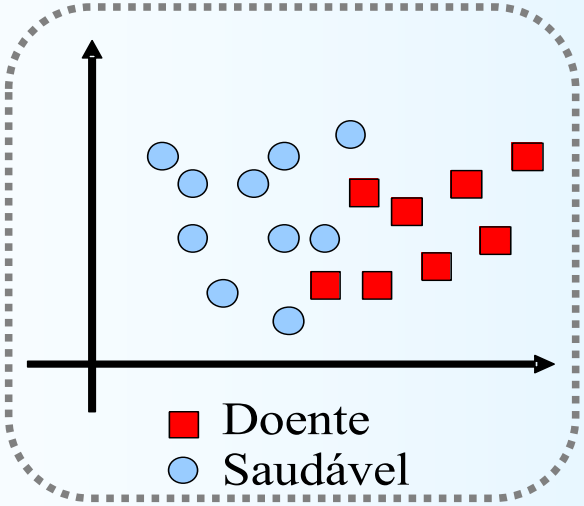
- Algoritmos de Análise podem ter dificuldades em lidar com um número grande de objetos
  - Saturação de memória
  - Aumento do tempo computacional para ajustar os parâmetros do modelo

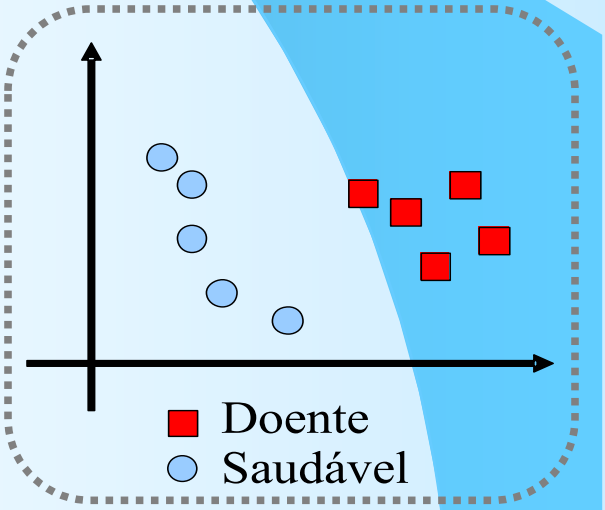
- Contudo, quanto mais dados, maior tende a ser a acurácia do modelo

Procurar balanço entre eficiência computacional e acurácia do modelo

Amostragem de dados







Dados Desbalanceados

- Tópico quando envolve classificação de dados
  - Número de objetos varia para as diferentes classes
    - Típico da aplicação
      - Ex. 80% dos pacientes que vão a um hospital estão doentes
    - Problema na geração/coleta dos dados

Classe majoritária	Classe minoritária
• **Contém a maior parte dos exemplos**	• **Tem o menor número de exemplos no conjunto**

Dados Desbalanceados

- Acurácia preditiva de classificador deve ser maior que a obtida atribuindo um novo objeto à classe majoritária
  - Vários algoritmos de Análise têm o desempenho prejudicado para dados muito desbalanceados
    - Tendem a favorecer a classificação na classe majoritária

Dados Desbalanceados

- Alternativas para lidar com dados desbalanceados:
  - Obter novos dados para a classe minoritária
    - Na maioria dos casos não é possível...
  - Balancear artificialmente o conjunto de dados:
    - Redefinir o tamanho do conjunto de dados
    - Induzir um modelo para uma classe

Limpeza de Dados

- Qualidade dos dados:
  - Em geral, dados não foram produzidos para a sua análise específica;
  - Exemplos de problemas:
    - Ruídos: erros ou valores diferentes do esperado
    - Inconsistências: não combinam/contradizem valores de outros atributos no mesmo objeto
    - Redundâncias: objetos/atributos com mesmos valores
    - Dados incompletos: ausência de valores de atributos

Principal dificuldade: detecção de dados ruidosos

Limpeza de Dados

- Exemplos de causas de erros:
  - Falha humana
  - Falha no processo de coleta de dados
  - Limitações do dispositivo de medição
  - Má fé

Alguns erros são sistemáticos e mais fáceis de detectar e corrigir

Dados incompletos

- Ausência de valores para alguns atributos de alguns objetos
  - Ex. conjunto de dados `hospital`

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
--	M	79	--	38,0	--	Doente
18	F	67	Pequenas	39,5	4	Doente
49	M	92	Grandes	38,0	2	Saudável
18	--	43	Grandes	38,5	20	Doente
21	F	52	Médias	37,6	1	Saudável
22	F	72	Pequenas	38,0	3	Doente
--	F	87	Grandes	39,0	6	Doente
34	M	67	Médias	38,4	2	Saudável

Dados incompletos

- Alternativas para lidar com valores ausentes:
  - Eliminar os objetos com valores ausentes
  - Definir e preencher manualmente os valores ausentes
  - Utilizar método/heurística para definir valores automaticamente

Dados incompletos

- Usando média/moda
  - Ex. conjunto de dados `hospital`

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
27	M	79	Grandes	38,0	4	Doente
18	F	67	Pequenas	39,5	4	Doente
49	M	92	Grandes	38,0	2	Saudável
18	F	43	Grandes	38,5	20	Doente
21	F	52	Médias	37,6	1	Saudável
22	F	72	Pequenas	38,0	3	Doente
27	F	87	Grandes	39,0	6	Doente
34	M	67	Médias	38,4	2	Saudável

Pode gerar inconsistências. Ex. paciente de 2 anos com 60 kg

Dados inconsistentes

- Possuem valores conflitantes em seus atributos
  - Nos atributos de entrada
    - Ex. 3 anos de idade e 120 kg
  - Entre entradas iguais e saída diferente
    - Ex. conjunto de dados `hospital`

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
28	M	79	Grandes	38,0	2	Doente
18	F	67	Pequenas	39,5	4	Doente
49	M	92	Grandes	38,0	2	Saudável
18	M	43	Grandes	38,5	20	Doente
21	F	52	Médias	37,6	1	Saudável
22	F	72	Pequenas	38,0	3	Doente
19	F	87	Grandes	39,0	6	Doente
22	F	72	Pequenas	38,0	3	Saudável

Dados redundantes

- Valores que não trazem informação nova
  - Objetos redundantes
    - Muito semelhante a outro(s) no conjunto de dados
      - Ex.: Pessoas em diferentes BDs com mesmo endereço e pequenas diferenças nos nomes
  - Atributos redundantes
    - Valor pode ser deduzido a partir do valor de um ou mais atributos
- Possíveis causas:
  - Problemas na coleta, entrada, armazenamento, integração ou transmissão

Dados redundantes

- Ex. conjunto de dados `hospital`

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
28	M	79	Grandes	38,0	2	Doente
18	F	67	Pequenas	39,5	4	Doente
49	M	92	Grandes	38,0	2	Saudável
18	F	67	Pequenas	39,5	4	Doente
21	F	52	Médias	37,6	1	Saudável
22	F	72	Pequenas	38,0	3	Doente
19	F	87	Grandes	39,0	6	Doente
34	M	67	Médias	38,4	2	Saudável

Duplicação

Dados redundantes

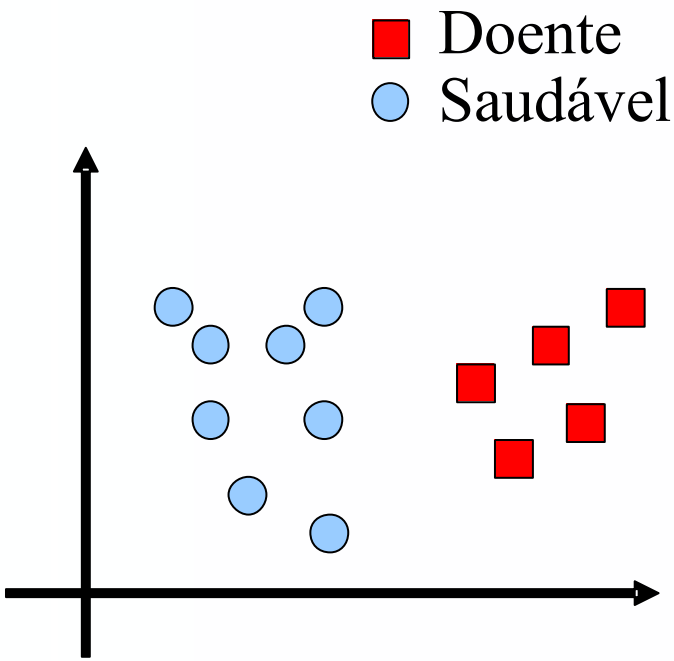
- Redundância de atributo está relacionada à sua correlação com um ou mais dos demais atributos
  - Dois atributos estão correlacionados quando têm perfil de variação semelhante para diferentes objetos
    - Ex. conjunto de dados `hospital`

Idade	Sexo	Peso	Manchas	Temp.	# Int.	# Vis.	Diagnóstico
28	M	79	Grandes	38,0	2	2	Doente
18	F	67	Pequenas	39,5	4	4	Doente
49	M	92	Grandes	38,0	2	2	Saudável
18	M	43	Grandes	38,5	20	20	Doente
21	F	52	Médias	37,6	1	1	Saudável
22	F	72	Pequenas	38,0	3	3	Doente
19	F	87	Grandes	39,0	6	6	Doente
34	M	67	Médias	38,4	2	2	Saudável

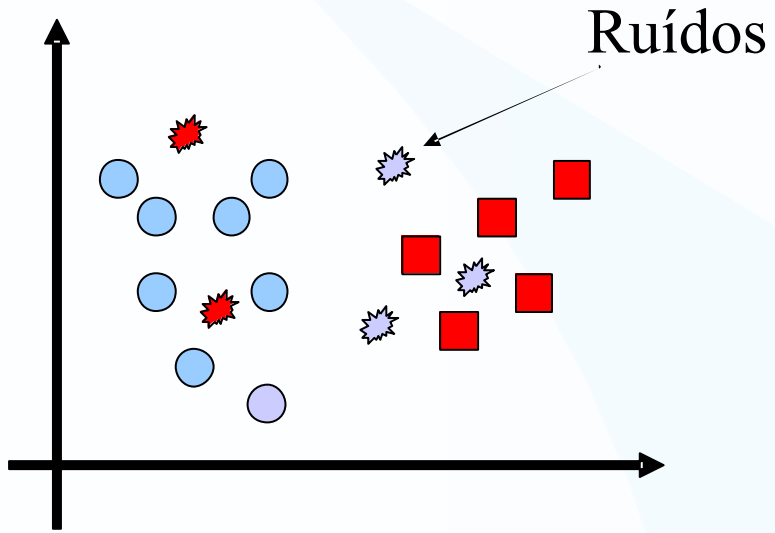
Ruídos

- Objetos que aparentemente não pertencem à distribuição que gerou os dados
- Várias causas possíveis
- Podem levar a superajuste do modelo
  - Algoritmo pode se ater às especificidades dos ruídos
- Mas eliminação pode levar à perda de informação importante
  - Algumas regiões do espaço de atributos podem não ser consideradas

Ruídos



Dados sem ruído



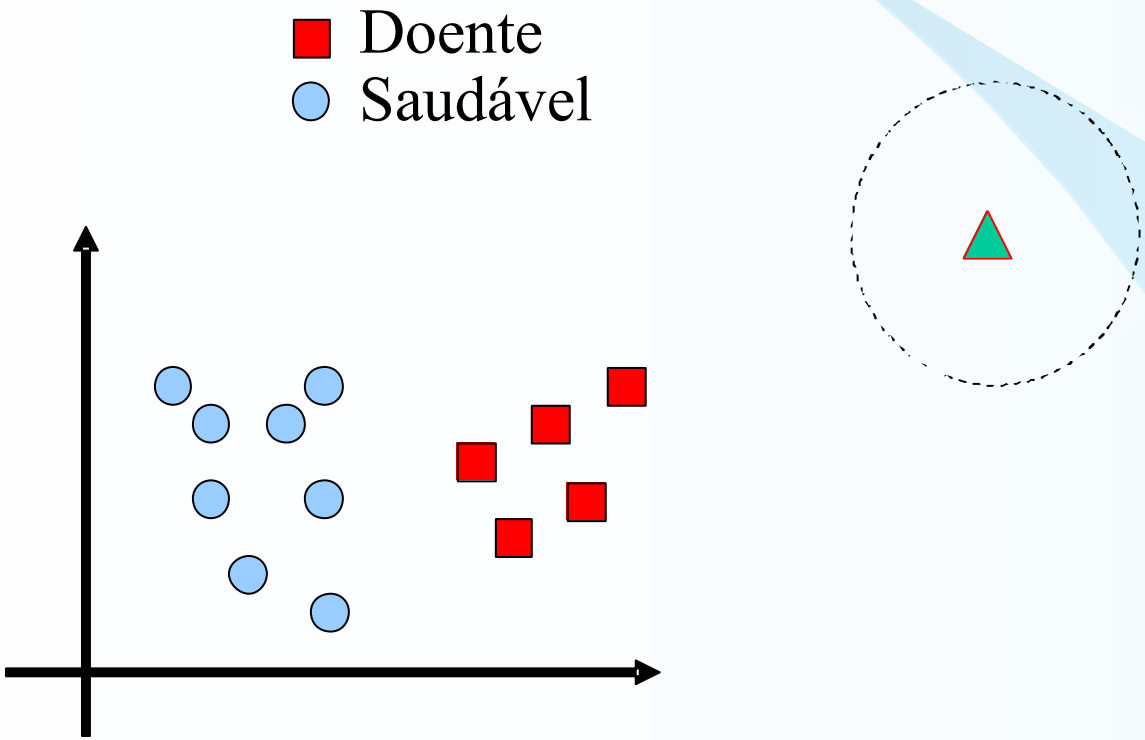
Dados com ruído

Outliers

- Valores que estão além dos limites aceitáveis ou são muito diferentes dos demais (exceções)
  - Podem ser valores legítimos
    - Ex. conjunto de dados `hospital`

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
28	M	79	Grandes	38,0	2	Doente
18	F	300	Pequenas	39,5	4	Doente
49	M	92	Grandes	38,0	2	Saudável
18	M	43	Pequenas	38,5	20	Doente
21	F	52	Médias	37,6	1	Saudável
22	F	72	Pequenas	38,0	3	Doente
19	F	87	Grandes	39,0	6	Doente
34	M	67	Médias	38,4	2	Saudável

Outliers



Ruídos

- Algumas técnicas de pré-processamento:
  - Técnicas baseadas em distribuição
  - Técnicas de encestamento
  - Técnicas baseadas em agrupamento dos dados
  - Técnicas baseadas em distância
  - Técnicas baseadas em regressão ou classificação

Ruídos

- Técnicas:

Baseadas em distribuição

- **Ruídos identificados como observações que diferem de uma distribuição usada na modelagem dos dados**
- **Problema: distribuição dos dados normalmente não é conhecida *a priori***

Encestamento

- **Suavizam valor de atributo**
- **1º: Ordena valores de atributo;**
- **2º:divide em cestas (faixas), cada uma com o mesmo número de valores**
- **3º: Substitui valores em uma mesma cesta, por ex., por média/moda**

Ruídos

- Técnicas:

Agrupamento

- **Agrupa objetos/atributos de acordo com semelhança**
- **Atributos/objetos que não formam grupo são ruídos ou *outliers***
- **Objetos colocados em um grupo que pertence a outra classe também são considerados ruídos**

Baseadas em distâncias

- **Presença de ruído em atributo frequentemente faz com que ele se distancie dos demais objetos de sua classe**
- **verificar a que classe pertencem os vizinhos mais próximos de x**
- **Se são de classe diferente, x pode ser ruído ou *borderline* (próximo à fronteira de separação das classes, podem ser inseguros)**

Exercício

- Identificar problemas no seguinte conjunto de dados:

Nome	Profissão	Nível	Peso	Altura	Salário	Situação
João	Encanador	Médio	70	180	3000	adimplente
Lia	Médico	Superior	200	174	7000	inadimplente
Maria	Advogado	Médio	90	180	600	adimplente
José	Médico	Superior	100	-6	2000	inadimplente
Sérgio	Bancário	Superior	82	178	5000	inadimplente
Ana	Professor	Fundam.	77	188	1800	adimplente
Luísa	Médico	Superior	100	-6	2000	inadimplente

Exercício

- Identificar problemas no seguinte conjunto de dados:

Nome	Profissã	Nível	Peso	Altura	Salário	Situação
João	Encanador	Médio	70	180	3000	adimplente
Lia	Médico	Superior	200	174	7000	inadimplente
Maria	Advogado	Médio	90	180	600	adimplente
José	Médico	Superior	100	-6	2000	inadimplente
Sérgio	Bancário	Superior	82	178	5000	inadimplente
Ana	Professor	Fundam.	77	188	1800	adimplente
Luísa	Médico	Superior	100	-6	2000	inadimplente

Transformação de Dados

- Algumas técnicas de Análise são limitadas à manipulação de valores de determinado tipo
  - Apenas numéricos ou simbólicos
- Algumas técnicas de Análise têm desempenho influenciado pela variação dos valores numéricos

Conversão simbólico-numérico

- Atributo simbólico com dois valores
  - Um dígito binário é suficiente
    - Ex. presença/ausência = 1/0
    - Se ordinal, 0 indica o menor valor e 1 o maior valor
- Atributo simbólico com mais valores
  - Conversão depende se o atributo é nominal ou ordinal

Conversão simbólico-numérico

- Ex. conjunto de dados `hospital`
  - Conversão de atributo Sexo para numérico

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
28	0	79	Grandes	38,0	2	Doente
18	1	67	Pequenas	39,5	4	Doente
49	0	92	Grandes	38,0	2	Saudável
18	0	43	Grandes	38,5	20	Doente
21	1	52	Médias	37,6	1	Saudável
22	1	72	Pequenas	38,0	3	Doente
19	1	87	Grandes	39,0	6	Doente
34	0	67	Médias	38,4	2	Saudável

M = 0
F = 1

Conversão simbólico-numérico

- Atributo nominal com mais que dois valores
  - Ex. codificação canônica (1-para-c ou topológica)

Atributo	Código 1-para-c
Azul	100000
Amarelo	010000
Verde	001000
Preto	000100
Marrom	000010
Branco	000001

Dependendo do número de valores nominais, pode gerar cadeias muito grandes de bits. Ex.: 193 nomes de países

Conversão simbólico-numérico

- Atributo ordinal com mais que dois valores
  - Relação de ordem deve ser preservada
  - Ordenar valores ordinais e codificar cada um de acordo com sua posição na ordem com inteiro ou real

Atributo	Valor inteiro
Primeiro	0
Segundo	1
Terceiro	2
Quarto	3
Quinto	4
Sexto	5

Distância entre valores varia de acordo com proximidade entre eles

Conversão simbólico-numérico

- Ex. conjunto de dados `hospital`
  - Conversão de atributo ordinal Manchas

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
28	M	79	3	38,0	2	Doente
18	F	67	1	39,5	4	Doente
49	M	92	3	38,0	2	Saudável
18	M	43	3	38,5	20	Doente
21	F	52	2	37,6	1	Saudável
22	F	72	1	38,0	3	Doente
19	F	87	3	39,0	6	Doente
34	M	67	2	38,4	2	Saudável

Grandes = 3
Médias = 2
Pequenas = 1

Conversão simbólico-numérico

- Atributo ordinal com mais que dois valores
  - Se for necessário usar valores binários, pode ser utilizado:
    - Código cinza: correção de erros em comunicação
    - Código termômetro: aumento de valores se assemelha a aumento de temperatura em termômetro

Atributo	Código cinza	Código termômetro
Primeiro	000	00000
Segundo	001	00001
Terceiro	011	00011
Quarto	010	00111
Quinto	110	01111
Sexto	100	11111

Conversão numérico-simbólico

- Atributo discreto e binário ⇒ conversão é trivial
  - Associa um nome a cada valor
    - Também se são sequências binárias sem relação de ordem
- Demais casos: discretização
  - Transforma valores numéricos em intervalos (categorias)
  - Existem vários métodos diferentes para discretização
    - Paramétricos: usuário pode influenciar definição dos intervalos
    - Não paramétricos: usam apenas informações presentes nos valores dos atributos

Transformação de atributos numéricos

- Algumas vezes é necessário transformar o valor de um atributo numérico em outro valor numérico
  - Quando o intervalo de valores são muito diferentes, levando a grande variação
  - Quando vários atributos estão em escalas diferentes
    - Para evitar que um atributo predomine sobre outro
- Porém, em alguns casos pode ser importante preservar a variação

Transformação de atributos numéricos

- Transformação é aplicada aos valores de um dado atributo para todos os objetos

- Uma transformação muito usada: normalização
  - Faz com que conjunto de valores de um atributo tenha uma determinada propriedade
  - Quando escalas de valores de atributos distintos são muito diferentes
    - Evita que um atributo predomine sobre o outro
      - A menos que isso seja importante

Normalização por reescala

- Ex. conjunto de dados `hospital`
  - Normalização de Idade entre 0 (min) e 1 (max)

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
28	M	79	Grandes	38,0	2	Doente
18	F	67	Pequenas	39,5	4	Doente
49	M	92	Grandes	38,0	2	Saudável
18	M	43	Grandes	38,5	20	Doente
21	F	52	Médias	37,6	1	Saudável
22	F	72	Pequenas	38,0	3	Doente
19	F	87	Grandes	39,0	6	Doente
34	M	67	Médias	38,4	2	Saudável

Maior = 49
Menor = 18

Normalização por reescala

- Ex. conjunto de dados `hospital`
  - Normalização de Idade entre 0 (min) e 1 (max)

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
28	M	79	Grandes	38,0	2	Doente
18	F	67	Pequenas	39,5	4	Doente
49	M	92	Grandes	38,0	2	Saudável
18	M	43	Grandes	38,5	20	Doente
21	F	52	Médias	37,6	1	Saudável
22	F	72	Pequenas	38,0	3	Doente
19	F	87	Grandes	39,0	6	Doente
34	M	67	Médias	38,4	2	Saudável

$$v_{novo} = \frac{v_{atual} - 18}{49 - 18}$$

Normalização por reescala

- Ex. conjunto de dados `hospital`
  - Normalização de Idade entre 0 (min) e 1 (max)

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
0,32	M	79	Grandes	38,0	2	Doente
0	F	67	Pequenas	39,5	4	Doente
1	M	92	Grandes	38,0	2	Saudável
0	M	43	Grandes	38,5	20	Doente
0,1	F	52	Médias	37,6	1	Saudável
0,13	F	72	Pequenas	38,0	3	Doente
0,03	F	87	Grandes	39,0	6	Doente
0,52	M	67	Médias	38,4	2	Saudável

Normalização por reescala

- Ex. conjunto de dados `hospital`
  - Efeito de *outlier*

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
0,32	M	79	Grandes	38,0	0,05	Doente
0	F	67	Pequenas	39,5	0,16	Doente
1	M	92	Grandes	38,0	0,05	Saudável
0	M	43	Grandes	38,5	1	Doente
0,1	F	52	Médias	37,6	0	Saudável
0,13	F	72	Pequenas	38,0	0,11	Doente
0,03	F	87	Grandes	39,0	0,26	Doente
0,52	M	67	Médias	38,4	0,05	Saudável

Normalização por padronização

- Para padronizar valores de atributos basta:
  - Adicionar/subtrair por uma medida de localização
  - Multiplicar/dividir por uma medida de escala
- Lida melhor com *outliers*
- Ex. atributos com média 0 e variância 1:

$$v_{novo} = \frac{v_{atual} - méd(x^i)}{desv\_pad(x^i)}$$

Diferentes atributos podem ter limites superiores e inferiores diferentes, mas terão os mesmos valores para as medidas de escala e espalhamento

Transformação de atributos numéricos

- Outro tipo de transformação: tradução
  - Valor é traduzido por um mais facilmente manipulável
    - Ex. converter data de nascimento para idade
    - Ex. converter temperatura de F para C
    - Ex. localização por GPS para código postal

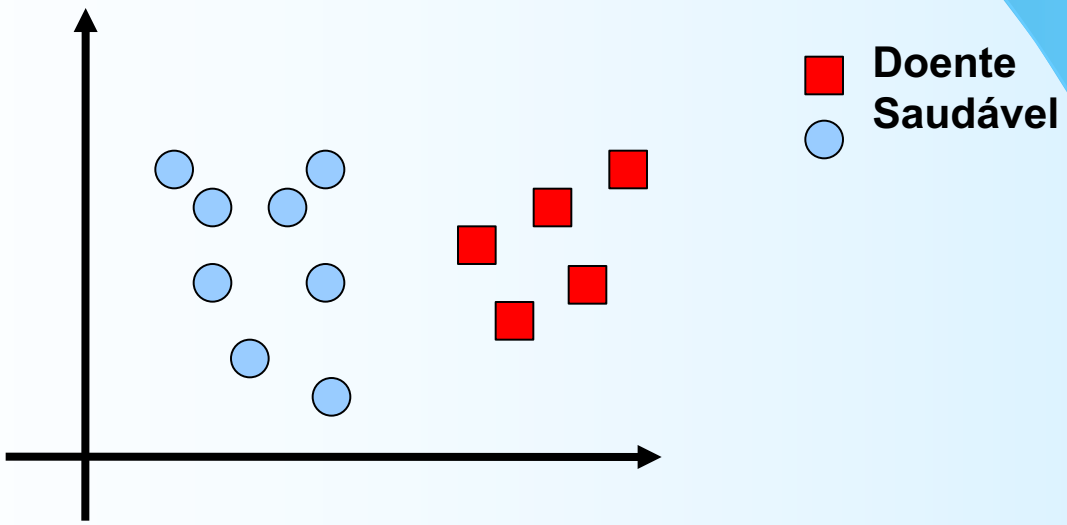
Redução de dimensionalidade

- Muitos problemas possuem número elevado de atributos
  - Ex. reconhecimento de imagens
    - Se cada pixel for considerado um atributo

Problema: maldição da dimensionalidade

Maldição da dimensionalidade

- Supor dados representados por pontos em um hipervolume
  - Valores de atributos dão as coordenadas



Maldição da dimensionalidade

- Hipervolume cresce exponencialmente com a adição de novos atributos
  - 1 atributo com 10 possíveis valores $\Rightarrow$ 10 possíveis objetos
  - 5 atributos com 10 possíveis valores $\Rightarrow$ 10^5 possíveis objetos
  - $\Rightarrow$ problemas com poucos exemplos e muitos atributos:
    - Dados se tornam muito esparsos

Sem exemplos em várias das regiões do espaço de objetos

Instâncias parecem equidistantes (dificultando encontrar padrões)

Redução de dimensionalidade

- Vantagens:
  - Alguns algoritmos de Análise que têm dificuldades em lidar com número elevado de atributos
  - Melhorar desempenho do modelo gerado
    - Identificação e eliminação de ruídos nos atributos
  - Reduzir custo computacional do modelo
  - Resultados mais compreensíveis

Redução de dimensionalidade

- Técnicas podem ser divididas em duas abordagens:

Agregação

- **Combinação dos atributos originais por funções lineares ou não lineares**
- **Ex. PCA (*Principal Component Analysis*), que elimina redundâncias por correlação**
- **Levam à perda dos valores originais**

Seleção de atributos

- **Identificar os atributos mais importantes**
- **Manter os relevantes**
- **Remover os redundantes e inconsistentes**
- **Diferentes critérios podem ser usados para medir importância**

PCA

- Conjunto de *d* atributos ($x^1, x^2, \ldots, x^d$)
  - Transformação linear para um novo conjunto de *d* atributos pode ser calculada como:

$$z^1 = a_{11}\, x^1 + a_{21}\, x^2 + \ldots + a_{d1}\, x^p$$
$$z^2 = a_{12}\, x^1 + a_{22}\, x^2 + \ldots + a_{d2}\, x^p$$
$$\ldots$$
$$z^d = a_{1d}\, x^1 + a_{2d}\, x^2 + \ldots + a_{dd}\, x^p$$

  - Componentes principais (PCs) são tipos específicos de combinações lineares

PCA

- Propriedades das Componentes Principais:
  - As *d* componentes principais são não correlacionadas (independentes)
  - As CPs são ordenadas de acordo com a quantidade de variância dos dados originais que elas contêm
    - Primeira componente "explica" (contém) a maior variabilidade do conjunto de dados
    - Segunda componente define próxima parte, e assim por diante

Em geral apenas algumas das primeiras CPs são responsáveis pela maior parte da variabilidade nos dados
O restante das PCs tem contribuição insignificante e pode ser eliminada

Considerações finais

- Pré-processamento:
  - Amostragem
  - Limpeza de dados
  - Transformação de dados
  - Redução do número de atributos

Referências

- FACELI, K.; LORENA, A.C.;GAMA J.; CARVALHO, A.C.L.F. Inteligência Artificial: uma abordagem de aprendizado de máquina. Capítulos 1,2 e 3.
- JAIN R. The Art of Computer Systems Performance Analysis, John Wiley & Sons, 1991. Capítulos:1, 2, 3 e 5.
- slides baseados em apresentações (cordialmente cedidos):
  - Prof Dr André C. P. L. F. Carvalho, ICMC-USP e Profa. Dra. Ana Carolina Lorena