

**UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO**

**Tópicos Especiais em Computação: Aprendizado
de Máquina**

Cap 2: Análise e Pré-processamento de dados

**Prof. Jefferson Moraes
Email: jmorais@ufpa.br**

Pré-processamento

- Desempenho dos algoritmos de AM é geralmente afetado pelo estado dos dados
 - Conjunto de dados podem apresentar diferentes características, dimensões ou formatos
 - Atributos numéricos ou simbólicos
 - Limpos ou com ruídos e imperfeições
 - Valores incorretos, inconsistentes, duplicados ou ausentes (*missing*)
 - Atributos independentes ou relacionados
 - Poucos ou muitos dados e/ou atributos
- Técnicas de pré-processamento são úteis: minimizar/eliminar problemas nos dados; tornar os dados mais adequados para uso por um algoritmo de AM

Pré-processamento

- Grupo de operações 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 uma ordem fixa para aplicação das diferentes técnicas de pré-processamento

Eliminação Manual de Atributos

- Alguns atributos não possuem relação com o problema a ser solucionado e por isso podem ser descartados
 - Há atributos que claramente não contribuem para o aprendizado
 - Ex: Os atributos **ID** e **Nome** no conjunto de dados hospital 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 do especialista
 - Ex: o especialista (médico) pode decidir que atributos associados à identificação do paciente, nome do paciente e ao estado de origem do paciente não são relevantes para o seu diagnóstico clínico

Eliminação Manual de Atributos

- Conjunto de dados hospital sem atributos considerados irrelevantes

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

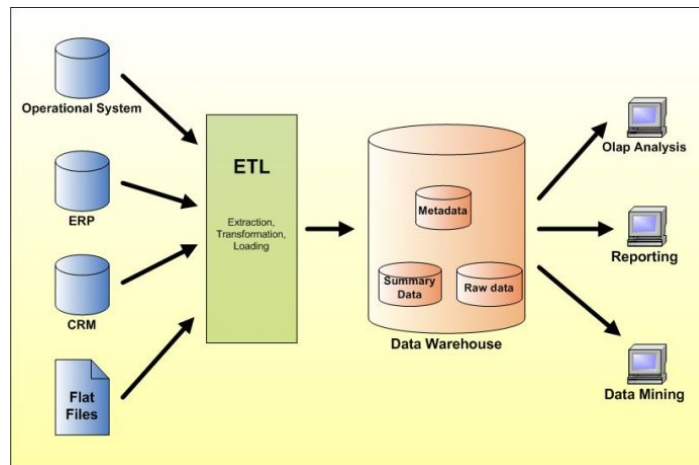
- Outro atributo irrelevante facilmente detectado:
 - Atributo que possui o mesmo valor para todos os dados
 - Não carregar informação suficiente para ajudar a distingui-los
- Há ainda atributos irrelevantes de identificação não tão clara
 - Técnicas de seleção de atributos (*feature selection*) podem ajudar a descobrir

Integração de dados

- Dados podem vir de diferentes fontes
 - ➔ Integração de diferentes conjuntos de dados
 - Cada um pode ter atributos diferentes para os mesmos dados
- Assim, é necessário identificar quais são os dados que estão presentes nos diferentes conjuntos de dados a serem combinados
 - Problema da identificação de entidade
 - Normalmente realizada por meio da busca por atributos comuns nos conjuntos a serem combinados que tenham valor único para cada dado
 - Ex: identificação de pacientes

Integração de dados

- Alguns aspectos podem dificultar a integração
 - Atributos correspondentes podem ter nomes diferentes em diferentes bases de dados
 - Dados podem ter sido atualizados em momentos diferentes
- Para minimizar esses problemas usa-se metadados
 - Metadados: dados sobre dados que descrevem suas principais características
- O processo de integração origina um depósito ou repositório de dado (*data warehouse*), que funciona como uma base de dados centralizada



Amostragem de dados

- Algoritmos de AM podem ter dificuldades em lidar com um número grande de dados
 - Saturação da memória
 - Aumento do tempo computacional para ajustar os parâmetros do modelo
- Por outro lado, quanto mais dados, maior tende a ser a acurácia do modelo
 - Busca-se balanço entre eficiência computacional e acurácia do modelo

Amostragem de dados

- Amostra ou subconjunto de dados

- ☐ Pode levar ao mesmo desempenho do conjunto completo, a menor custo computacional

- Deve ser representativa

- ☐ A amostra deve representar aproximadamente as mesmas propriedades do conjunto de dados original
- ☐ Fornecer uma estimativa da informação contida na população original
- ☐ Uso deve ter efeito semelhante ao de toda a população
- ☐ Permitir conclusão do todo a partir de uma parte



Amostragem de dados

- Existem basicamente três abordagens para amostragem
 - Amostragem aleatória simples
 - Amostragem estratificada
 - Amostragem progressiva

Amostragem de dados

- Amostragem aleatória simples

- Possui duas variações

- Sem reposição de exemplos: exemplos são extraídos do conjunto original e cada exemplo pode ser selecionado uma vez
 - Com reposição: quando uma cópia dos exemplos selecionados é mantida no conjunto de dados original
 - As duas formas são semelhantes quando tamanho da amostra é bem menor que o conjunto original

Amostragem de dados

■ Amostragem estratificada

- Usada quando as classes apresentam propriedades diferentes (ex. Número de dados diferentes)
- Na classificação cuidado na amostragem diz respeito à distribuição dos dados nas diferentes classes
 - A existência de classes como uma quantidade significativamente maior de exemplos pode levar o algoritmo de AM a induzir as classes majoritárias
- Variações: manter o mesmo número de objetos para cada classe ou manter o número proporcional ao original

Amostragem de dados

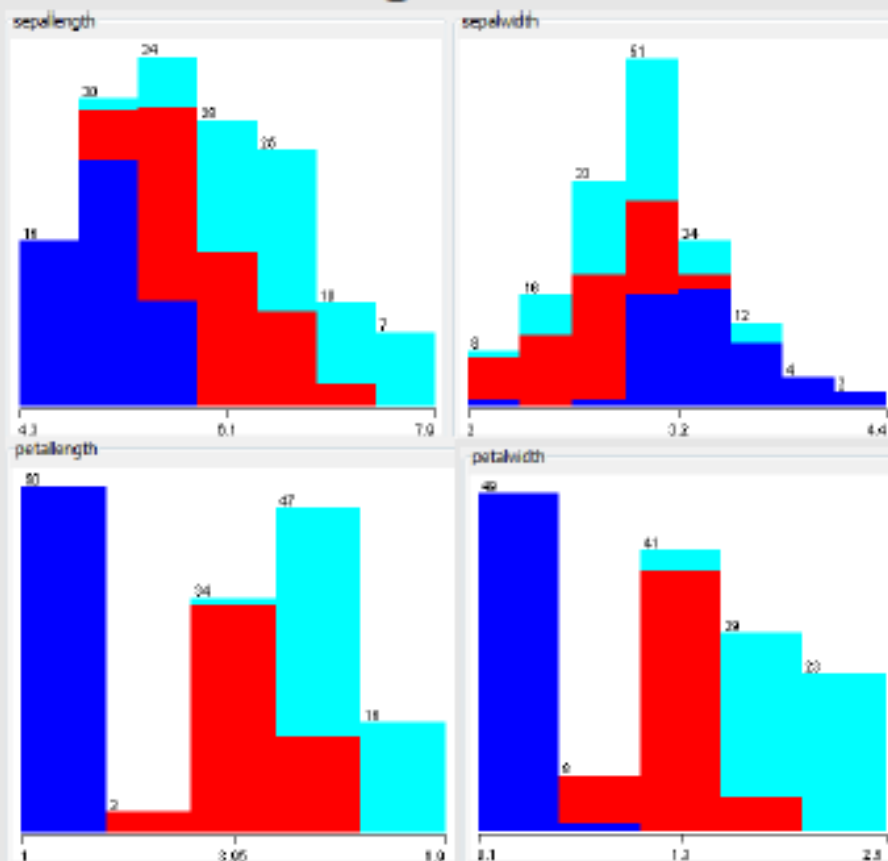
■ Amostragem progressiva

- Começa com amostra pequena e vai aumentando enquanto acurácia preditiva continuar a melhorar
- Resultado: é possível definir a menor quantidade de dados necessária, reduzindo ou eliminando a perda de acurácia
- O tamanho pode ser confirmado com outras amostras de tamanho semelhante
- Fornece uma boa estimativa para o tamanho da amostra

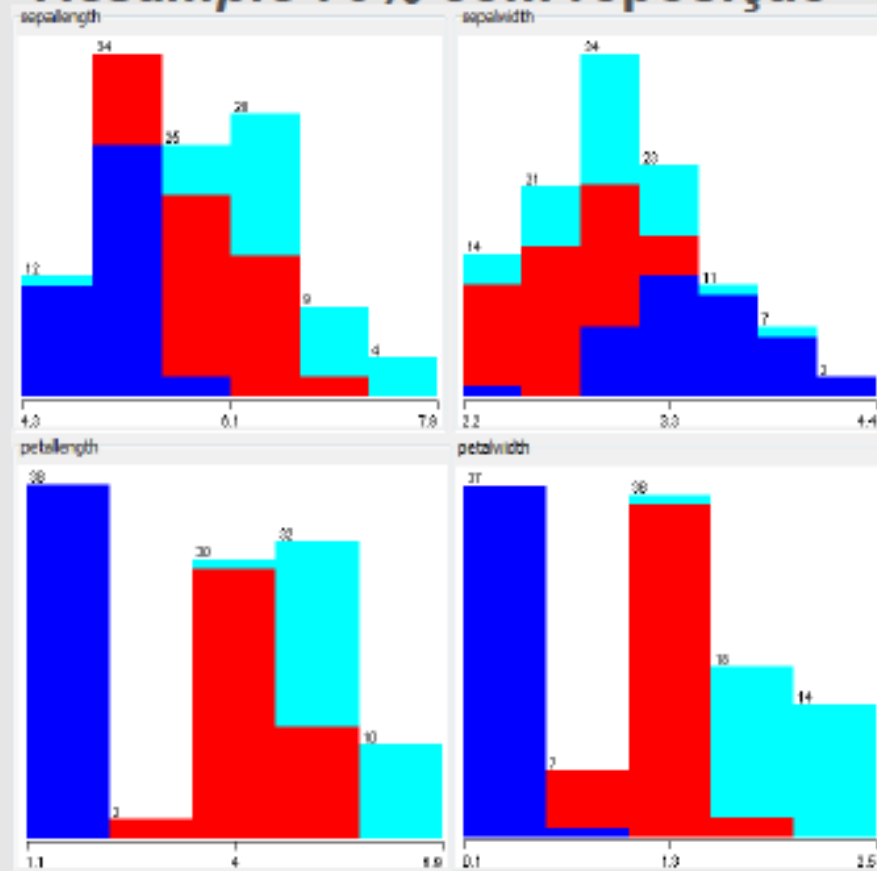
Amostragem de dados

- Ex: conjunto de dados iris

Original



Resample 70% com reposição





Amostragem de dados

- Especialista do domínio também pode auxiliar a decidir subconjuntos de dados a serem usados
 - Ex: em uma análise de pacientes em um hospital, podem ser utilizados apenas os dados referentes ao sexo feminino

Dados Desbalanceados

- Tópico da área de classificação de dados
 - Em vários conjuntos de dados reais, o número de dados 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
 - Contém a maior parte dos exemplos
 - Classe minoritária
 - 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 AM têm o desempenho prejudicado para dados muito desbalanceados
 - Tendem a favorecer a classificação na classe majoritária
- 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; Usar diferentes custos de classificação para as classes; induzir um modelo para uma classes

Dados Desbalanceados

■ Técnicas de rebalanceamento

□ Redefinir o tamanho do conjunto de dados

- Acréscimo/eliminação de exemplos na classe minoritária/majoritária
- Acréscimo: risco de objetos que não representam situações reais e *overfitting*
- Eliminação: risco de perda de objetos importantes e *underfitting*



Dados Desbalanceados

■ Técnicas de rebalanceamento

- ☐ Usar custos de classificação diferentes para as classes
- ☐ Dificuldades: definição dos custos, incorporar custos em alguns algoritmos de AM
- ☐ Pode apresentar baixo desempenho quando muitos objetos da classe majoritária são semelhantes

Dados Desbalanceados

■ Técnicas de rebalanceamento

- ☐ Induzir modelo para uma única classe
- ☐ Técnicas de classificação para uma classe, treinadas usando somente exemplos de uma classe
- ☐ Aprendem classe(s) separadamente
- ☐ Outros uso de *one-class classification*
 - Detecção de novidades (ex. falhas em máquinas)
 - Detecção de *outliers*
 - Comparação de conjuntos de dados (evitar retrainar classificadores para dados semelhantes)

Limpeza de dados

■ Qualidade dos Dados

- Em geral, dados não foram produzidos para uso em AM
- Exemplos de problemas
 - Ruídos: erros ou valores diferentes do esperado
 - Inconsistências: não combinam/contradizem valores de outros atributos no mesmo dado
 - Redudâncias: dados/atributos com mesmo 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
- ☐ Valor do atributo com o tempo
- ☐ Alguns erros são sistemáticos e mais fáceis de detectar e corrigir

Limpeza de dados

■ Consequências

- Valores ou objetos inteiros podem ser perdidos
- Objetos espúrios ou duplicados podem ser obtidos
 - Ex: diferentes registros para mesma pessoa que morou em endereços diferentes
- Inconsistências
 - Ex: pessoa com 2m pesando 10Kg



Limpeza de Dados

- Algumas técnicas de AM conseguem lidar com algumas imperfeições nos dados
 - Outras não conseguem ou apresentam dificuldades
- Porém de forma geral, qualidade das análises pode ser deteriorada
- Todas as técnicas se beneficiam de melhora na qualidade dos dados, que pode ser obtida por meio de etapa de limpeza

Dados incompletos

- Ausência de valores para alguns atributos de alguns dados
 - 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

■ Possíveis causas:

- ☐ Atributo não era importante quando primeiros dados foram coletados
 - Ex: e-mail do paciente que não era comum na década de 90
- ☐ Desconhecimento do valor do atributo
 - Ex. não saber tipo sanguíneo de paciente em seu cadastro
- ☐ Falta de necessidade/obrigação de apresentar valor
 - Ex. Renda de um paciente
- ☐ Inexistência de valor para o atributo
 - Ex. número de partos para pacientes do sexo masculino
- ☐ Problema com equipamento para coleta, transmissão e armazenamento de dados

Dados incompletos

- Algumas técnicas de AM são incapazes de lidar com valores ausentes
 - Geram erro de execução
- 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
- Empregar algoritmos de AM que lidam internamente com valores ausentes

Dados incompletos

■ Técnicas

☐ Eliminar objetos

- Mais empregada quando classe está ausente
- Não indicada quando número de atributos com valores ausentes varia muito entre os objetos ou quando muitos objetos têm valores ausentes

☐ Definir/preencher manualmente

- Não é factível para muitos valores ausentes

☐ Usar heurística

- Alternativa mais usada

Dados incompletos

- Técnicas para definição automática de valores
 - Criar valor “desconhecido”
 - Comum a todos ou diferente para cada atributo
 - Utilizar média/moda/ mediana dos valores conhecidos
 - Usando todos os objetos ou somente aqueles da mesma classe
 - **Variação**: usar valor mais frequente entre k vizinhos mais próximos
 - Usar indutor para estimar o valor
 - Valor a ser definido passa a ser o atributo alvo
 - Usa informação dos outros atributos para inferir o ausente

Dados incompletos

- Usando a 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

- **Observação:** pode gerar inconsistências. Ex: paciente com 2 anos com 70 Kg

Dados inconsistentes

- Possuem valores conflitantes em seus atributos
 - Nos atributos de entrada
 - Ex: 3 anos de idade e 100 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 inconsistentes

- Possíveis causas:

- Erro/engano

- Presença de ruídos nos dados

- Proposital (fraude)

- Problemas na integração dos dados

- Ex. conjuntos de dados com escalas diferentes para uma mesma medida

Dados inconsistentes

- Algumas inconsistências são de fácil detecção:
 - Violação de relações conhecidas entre atributos
 - Ex.: Valor de atributo A é sempre menor que valor de atributo B
 - Valor inválido para o atributo
 - Ex.: altura com valor negativo
 - Em outros casos, informações adicionais precisam ser verificadas

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 reduntantes

- 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



Dados reduntantes

- Objetos redundantes participam mais de uma vez do ajuste do modelo
 - ☐ Pode assim ser considerado um perfil mais importante que o dos outros
 - ☐ Pode também aumentar custo computacional
- Passos para eliminar objetos redundantes:
 - ☐ Identificar as redundâncias
 - ☐ Eliminar as redundâncias
 - ☐ Remoção ou combinação dos valores

Dados redundantes

- Atributo redundante: valor pode ser estimado a partir de pelo menos um dos demais atributos
- Atributos com a mesma informação preditiva
 - Ex. atributos idade e data de nascimento
 - Ex. atributos quantidade de vendas, valor por venda e venda total
- Atributo redundante pode supervalorizar um dado aspecto dos dados
- Pode também tornar mais lento o processo de indução
- Atributos redundantes são geralmente eliminados por técnicas de **seleção de atributos** (*feature selection*)

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

Outliers

- Valores que estão além dos limites aceitáveis ou são muito diferentes dos demais (exceções)

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



Ruídos

- Algumas técnicas de pré-processamento para detecção e remoção de ruídos
 - 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
 1. Ordena valores de atributo;
 2. Divide em cestas (faixas), cada uma com o mesmo número de valores
 3. 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)



Ruídos

■ Técnicas

□ Baseadas em regressão/classificação

- Usam função de regressão ou classificação para, dado um valor com ruído, estimar seu valor verdadeiro (regressão para atributo contínuo e classificação para simbólico)



Transformação de Dados

- Várias técnicas de AM estão limitadas à manipulação de valores de determinados tipos
 - Apenas numéricos ou simbólicos
- Algumas técnicas de AM têm seu desempenho influenciado pelo intervalo de variação dos valores numéricos

Conversão Simbólico-Numérico

- Técnicas com SVM e ANN lidam apenas com dados numéricos
- Atributo nominal com dois valores
 - Ex: presença/ausência = 1/0
 - Se ordinal, 0 indica o menor valor e 1 o maior valor
- Atributo nominal com mais valores
 - Conversão depende se o atributo é **nominal** ou **ordinal**

Conversão Simbólico-Numérico

- Atributo simbólico com mais de dois valores
 - Inexistência de relação de ordem deve ser mantida
 - A diferença entre quaisquer dois valores numéricos deve ser a mesma
 - Codificação canônica (ou topológica): uso de *c bits para c* valores
 - Cada posição na sequência binária corresponde a um valor possível do atributo nominal
 - Cada sequência possui apenas um bit com valor 1
 - Distância de Hamming entre quaisquer dois valores é 2

Conversão simbólico-numérico

- Ex: conjunto de dados hospital
 - Conversão do atributo Sexo para numérico (M=0 e F=1)

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

Conversão simbólico-numérico

- Atributo nominal com mais de 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

Conversão simbólico-numérico

- Ex: conjunto de dados hospital
 - Conversão atributos ordinal Manchas (Grandes=3, Médias=2 e Pequenas =1)

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

Conversão numérico-simbólico

- Algumas técnicas de AM trabalham apenas com valores qualitativos. Ex: algoritmos de associação
- 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

Conversão numérico-simbólico

- Métodos de discretização podem ser:
 - Supervisionados: usa informação da classe
 - Melhor resultado, não leva a mistura de classes
 - Ex. escolher pontos de corte que maximizam pureza dos intervalos (entropia)
 - Não supervisionados
- Método de discretização deve definir:
 - Como mapear valores quantitativos para qualitativos
 - Tamanho dos intervalos
 - Quantidade de valores nos intervalos

Conversão numérico-simbólico

■ Algumas estratégias

☐ Larguras iguais

- Dividir valores em sub-intervalos com mesma largura
- Problema: desempenho afetado pela presença de *outliers*

☐ Frequências iguais

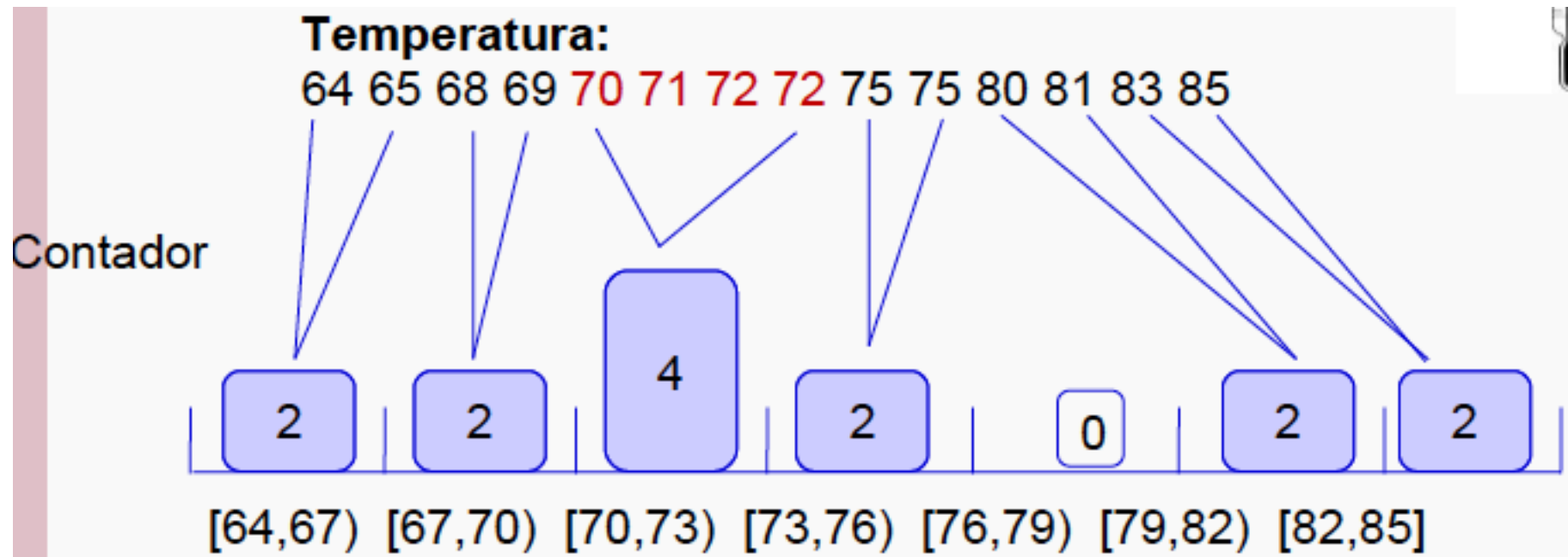
- Mesmo número de objetos em cada intervalo
- Problema: pode gerar intervalos de tamanhos muito diferentes

☐ Inspeção visual

☐ Uso de um algoritmo de agrupamento

Conversão numérico-simbólico

- Ex: discretização com larguras iguais

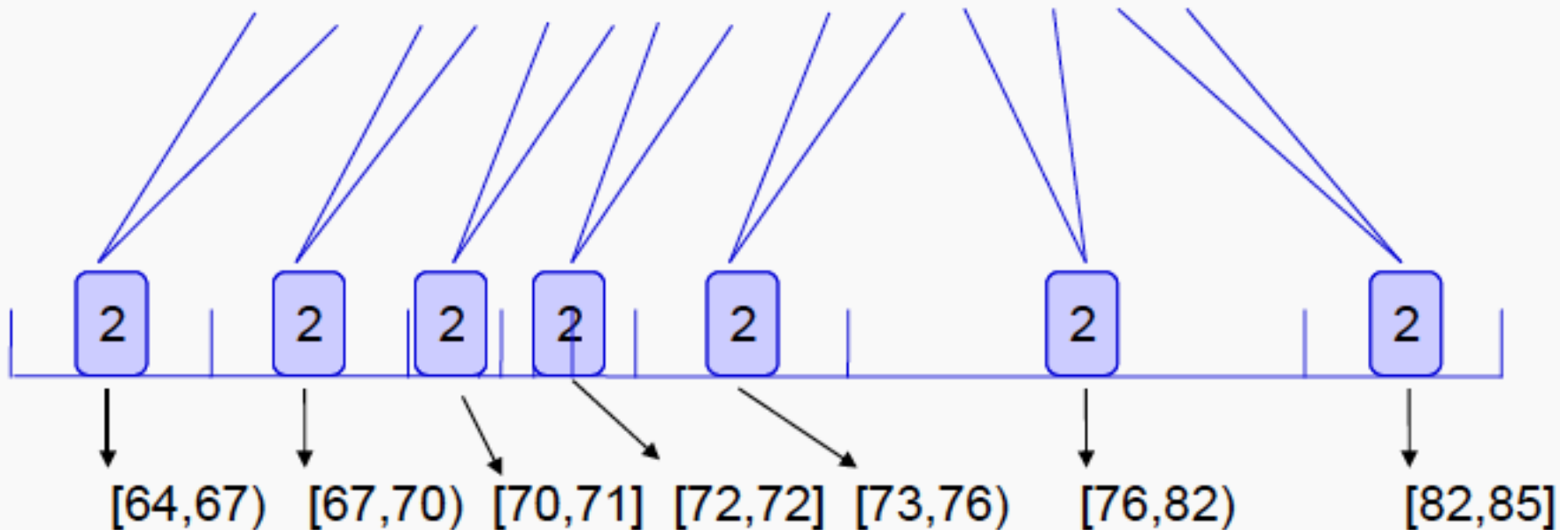


Conversão numérico-simbólico

- Ex: discretização com frequências iguais

Temperatura:

64 65 68 69 70 71 72 72 75 75 80 81 83 85



Dealing with numeric attributes

- Discretize numeric attributes
- Divide each attribute's range into intervals
 - Sort instances according to attribute's values
 - Place breakpoints where the class changes

(the majority class)

- This minimizes

- Example: *tennis*

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	75	80	False	Yes
...

64 65 68 69 70 71 72 72 75 75 80 81 83 85
 Yes | No | Yes Yes Yes | No No Yes | Yes Yes | No | Yes Yes | No

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 a todos os objetos de um dado atributo
- 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

- Deve ser aplicada a cada atributo individualmente

- Duas formas

- Por amplitude

- Por reescala: define nova escala (máximo e mínimo) de valores para atributos
 - Por padronização: define um valor central e de espalhamento comuns para todos os atributos

- Por distribuição

- Muda a escala de valores
 - Ex: Ordena valores dos atributos e substitui cada valor por sua posição no *ranking* (valores 1, 5, 9 e 3 viram 1, 3, 4 e 2)
 - Se valores originais forem distintos, resultado é distribuição uniforme

Normalização por reescala

- Reescalar:
adicionar/subtrair/multiplicar/dividir por uma constante
- Normalização min-max
 - São definidos inicialmente mínimo e máximo para os novos valores

$$v_{\text{novo}} = \text{min} + \frac{v_{\text{atual}} - \text{menor}}{\text{maior} - \text{menor}} \cdot (\text{max} - \text{min})$$

Normalização por reescala

- Ex: conjunto de dados hospital

- Normalização da Idade entre 0 (min) e 1 (max). Maior = 49 e Menor = 18

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

$$V_{\text{novo}} = \frac{V_{\text{atual}} - 18}{49 - 18}$$

Normalização por reescala

- Ex: conjunto de dados hospital

- ☐ Normalização da Idade entre 0 (min) e 1 (max). Maior = 49 e Menor = 18

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

- Normalização da #Int entre 0 (min) e 1 (max). Maior = 20 e Menor = 1

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

$$V_{\text{novo}} = \frac{V_{\text{atual}} - 1}{20 - 1}$$

Normalização por reescala

- Ex: conjunto de dados hospital
 - Normalização da #Int entre 0 (min) e 1 (max). Maior = 20 e Menor = 1

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 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

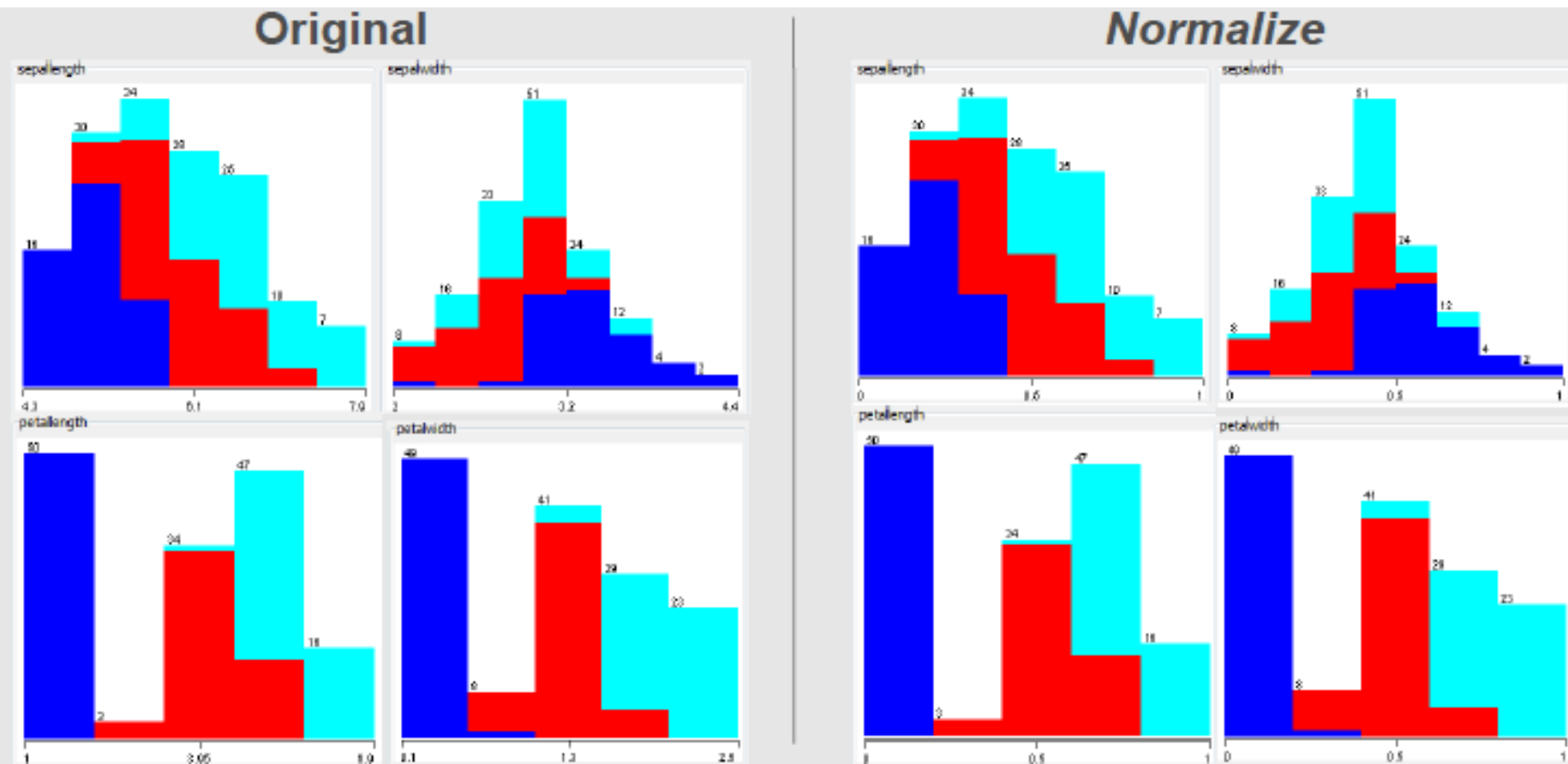
Normalização por reescala

- 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 reescala

- Conjunto de dados iris



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:

$$V_{\text{novo}} = \frac{V_{\text{atual}} - \text{méd}(\mathbf{x}^i)}{\text{desv_pad}(\mathbf{x}^i)}$$

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

Normalização por padronização

■ Ex: Conjunto de dados hospital

□ Padronização da Idade com média 0 e variância 1.

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


Média = 21,5
Desv_pad = 10,79

Normalização por padronização

■ Ex: Conjunto de dados hospital

□ Padronização da Idade com média 0 e variância 1.

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


$$V_{\text{novo}} = \frac{V_{\text{atual}} - 21,5}{10,79}$$

Normalização por padronização

■ Ex: Conjunto de dados hospital

□ Padronização da Idade com média 0 e variância 1.

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
0,6	M	79	Grandes	38,0	2	Doente
-0,32	F	67	Pequenas	39,5	4	Doente
2,55	M	92	Grandes	38,0	2	Saudável
-0,32	M	43	Grandes	38,5	20	Doente
-0,05	F	52	Médias	37,6	1	Saudável
0,05	F	72	Pequenas	38,0	3	Doente
-0,23	F	87	Grandes	39,0	6	Doente
1,16	M	67	Médias	38,4	2	Saudável

Média = 0
Desv_pad = 1

Normalização por padronização

■ Ex: Conjunto de dados hospital

- Padronização da #Int com média 0 e variância 1.

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
0,6	M	79	Grandes	38,0	2	Doente
-0,32	F	67	Pequenas	39,5	4	Doente
2,55	M	92	Grandes	38,0	2	Saudável
-0,32	M	43	Grandes	38,5	20	Doente
-0,05	F	52	Médias	37,6	1	Saudável
0,05	F	72	Pequenas	38,0	3	Doente
-0,23	F	87	Grandes	39,0	6	Doente
1,16	M	67	Médias	38,4	2	Saudável

Média = 2,5
Desv_pad = 6,26

Normalização por padronização

■ Ex: Conjunto de dados hospital

- Padronização da #Int com média 0 e variância 1.

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
0,6	M	79	Grandes	38,0	2	Doente
-0,32	F	67	Pequenas	39,5	4	Doente
2,55	M	92	Grandes	38,0	2	Saudável
-0,32	M	43	Grandes	38,5	20	Doente
-0,05	F	52	Médias	37,6	1	Saudável
0,05	F	72	Pequenas	38,0	3	Doente
-0,23	F	87	Grandes	39,0	6	Doente
1,16	M	67	Médias	38,4	2	Saudável

$$v_{\text{novo}} = \frac{v_{\text{atual}} - 2,5}{6,26}$$

Normalização por padronização

■ Ex: Conjunto de dados hospital

- Padronização da #Int com média 0 e variância 1.

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
0,6	M	79	Grandes	38,0	-0,08	Doente
-0,32	F	67	Pequenas	39,5	0,24	Doente
2,55	M	92	Grandes	38,0	-0,08	Saudável
-0,32	M	43	Grandes	38,5	2,8	Doente
-0,05	F	52	Médias	37,6	-0,24	Saudável
0,05	F	72	Pequenas	38,0	0,08	Doente
-0,23	F	87	Grandes	39,0	0,56	Doente
1,16	M	67	Médias	38,4	-0,08	Saudável

Média = 0
Desv_pad = 1

Normalização por padronização

■ Ex: Conjunto de dados hospital

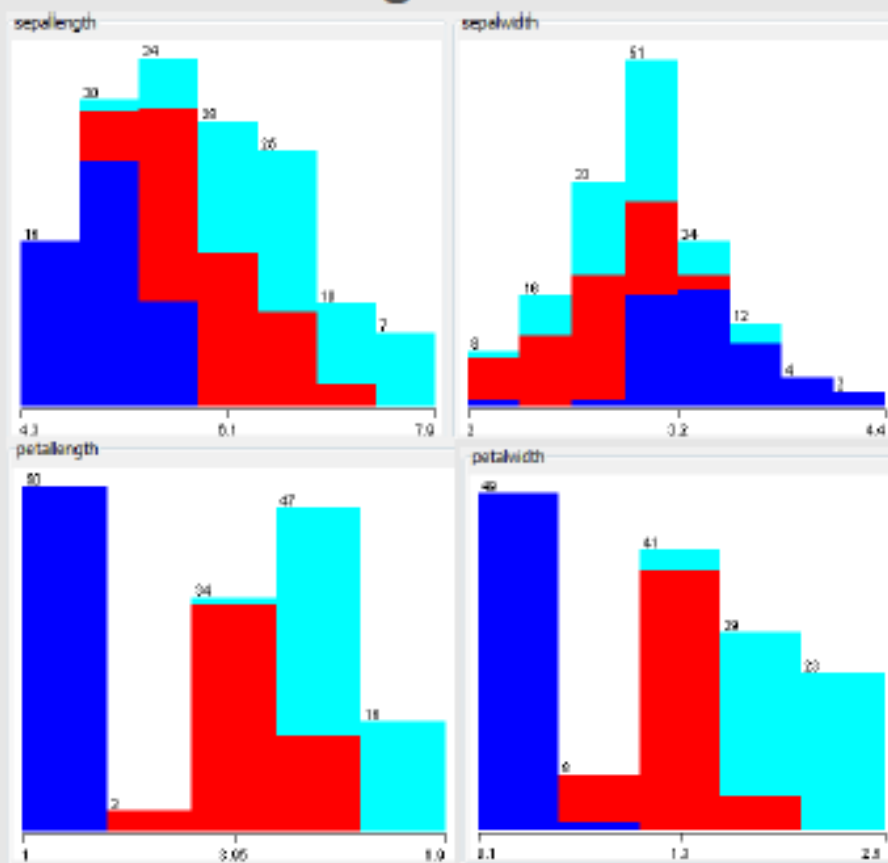
□ Efeito de *outlier*

Idade	Sexo	Peso	Manchas	Temp.	# Int.	Diagnóstico
0,6	M	79	Grandes	38,0	-0,08	Doente
-0,32	F	67	Pequenas	39,5	0,24	Doente
2,55	M	92	Grandes	38,0	-0,08	Saudável
-0,32	M	43	Grandes	38,5	2,8	Doente
-0,05	F	52	Médias	37,6	-0,24	Saudável
0,05	F	72	Pequenas	38,0	0,08	Doente
-0,23	F	87	Grandes	39,0	0,56	Doente
1,16	M	67	Médias	38,4	-0,08	Saudável

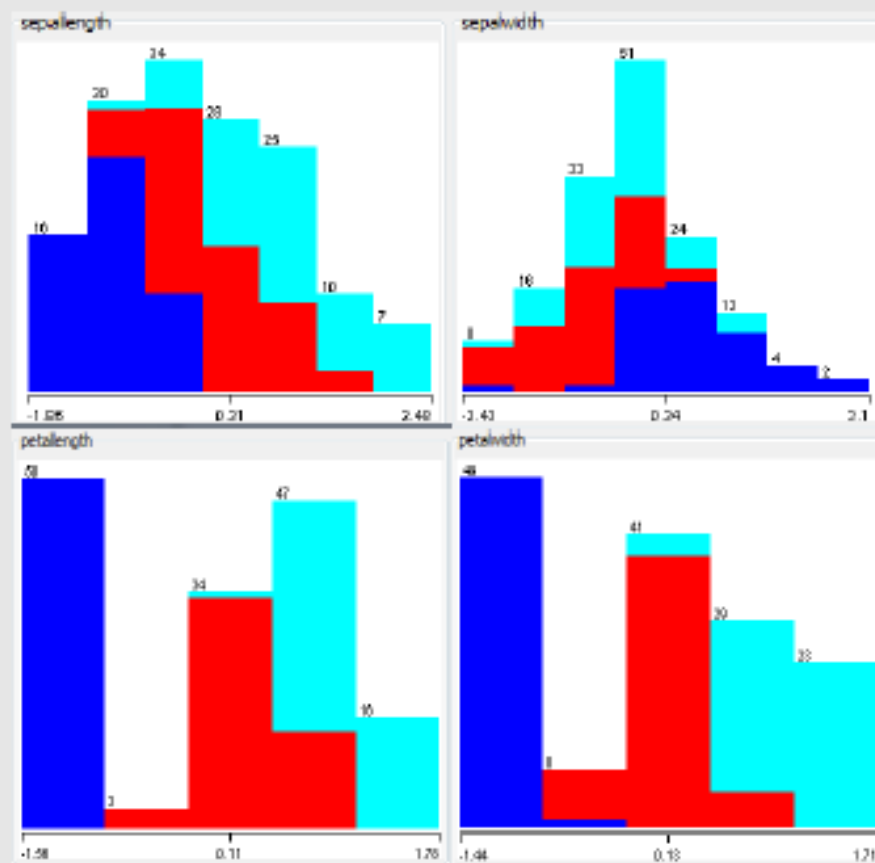
Normalização por padronização

- Ex: Conjunto de dados iris

Original



Standardize



Transformação de atributos numéricos

- Outro tipo de transformação: tradução
 - Valor de um atributo 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

Transformação de atributos numéricos

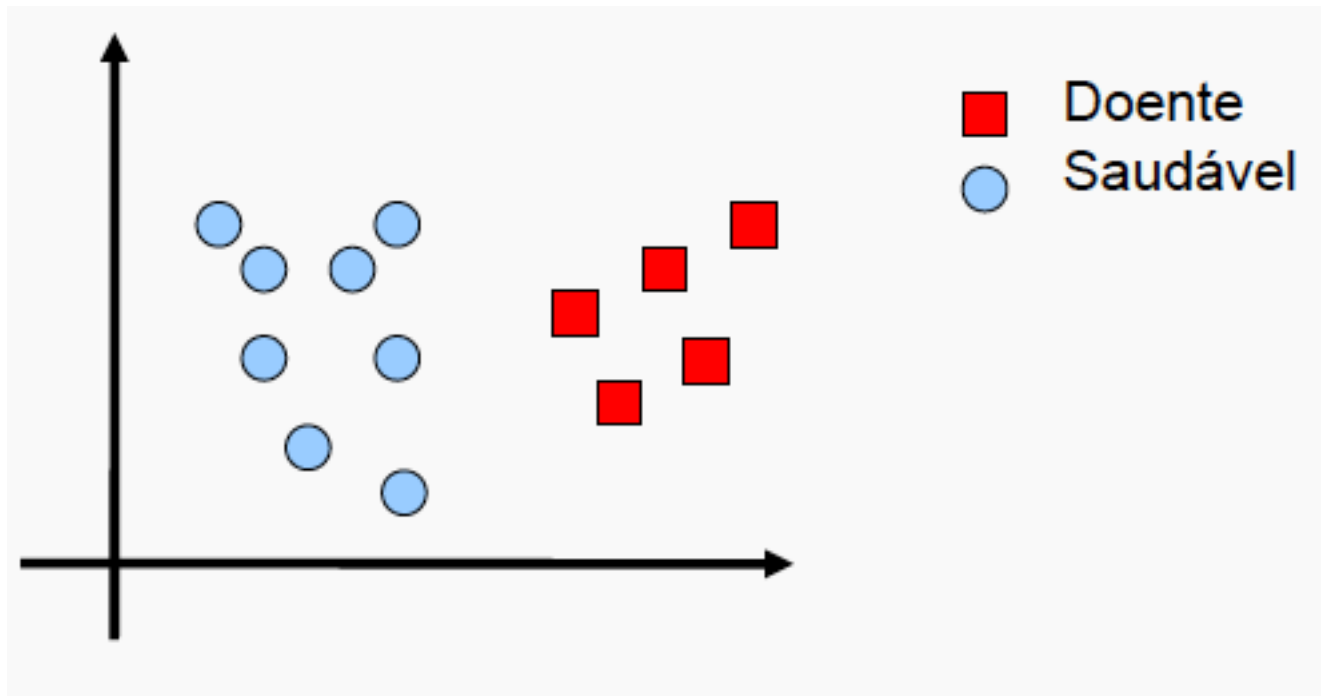
- Outro tipo de transformação: aplicação de função simples
 - Aplicação a cada valor do atributo
 - Ex. log, exp, raiz, seno, $1/x$, abs
 - Ex. apenas magnitude dos valores é importante fi converter para valor absoluto
- Funções raiz, log e $1/x$: aproximam uma distribuição Gaussiana
- Função log: comprimir dados com grande intervalo de valores

Redução de dimensionalidade

- Muitos problemas tratados por técnicas de AM apresentam um número elevado de atributos
 - Ex: reconhecimento de imagens e dados de expressão gênica
 - Cada pixel ou cada gene for considerado um atributo
- Problema da maldição da dimensionalidade (*curse of dimensionality*)
 - À medida que cresce o número de parâmetros (atributos) diminui a capacidade de generalização do classificador

Maldição da dimensionalidade

- Supor que os dados são 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 105 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)

Maldição da dimensionalidade

- Número de exemplos necessários para manter o desempenho cresce exponencialmente com o número de atributos
 - Na prática, o número de exemplos de treinamento é fixo
 - => Necessidade de **redução de dimensionalidade**

Redução da dimensionalidade

■ Vantagens:

- Alguns algoritmos de AM que têm dificuldades em lidar com número elevado de parâmetros
- Melhorar desempenho do modelo induzido
 - Identificação e eliminação de ruídos nos atributos
- Reduzir o custo computacional do modelo
- Resultados mais compreensíveis

■ Duas abordagens para reduzir a dimensionalidade

- Extração de parâmetros (*Feature Extraction*)
 - Faz um mapeamento do espaço original um espaço de menor dimensão
 - Ex: PCA (*Principal Component Analysis*), que elimina redundâncias por correlação
- Seleção de parâmetros (*Feature Selection*)
 - Consiste em identificar e reter apenas os parâmetros que mais contribuem para execução de uma tarefa
 - Não modifica os parâmetros apenas escolhe um subconjunto dentre o total de subconjuntos possíveis

PCA

- Conjunto de d atributos ($\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^d$)
 - Transformação linear para um novo conjunto de d atributos pode ser calculado como:

$$\begin{aligned} z^1 &= a_{11} x^1 + a_{21} x^2 + \dots + a_{d1} x^d \\ z^2 &= a_{12} x^1 + a_{22} x^2 + \dots + a_{d2} x^d \\ &\dots \\ z^d &= a_{1d} x^1 + a_{2d} x^2 + \dots + a_{dd} x^d \end{aligned}$$

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

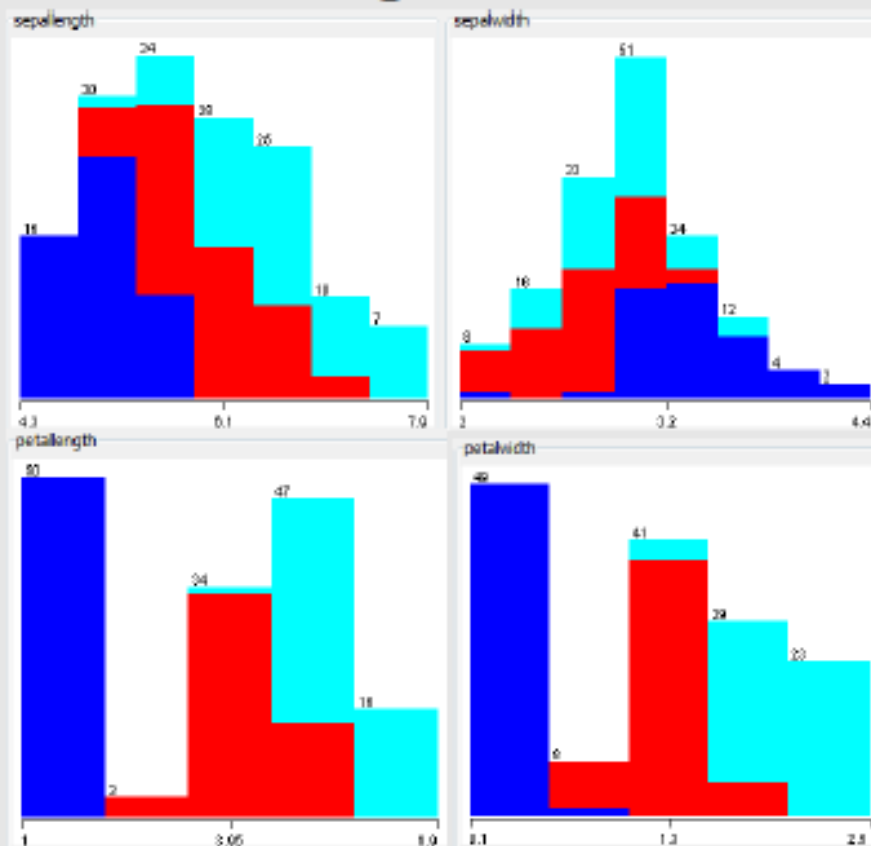
PCA

- Propriedades das componentes principais:
 - As d componentes principais não são correlacionadas (independentes)
 - As PCs 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 PCs são responsáveis pela maior parte da variabilidade nos dados
 - O restante das PCs tem contribuição insignificante e pode ser eliminada

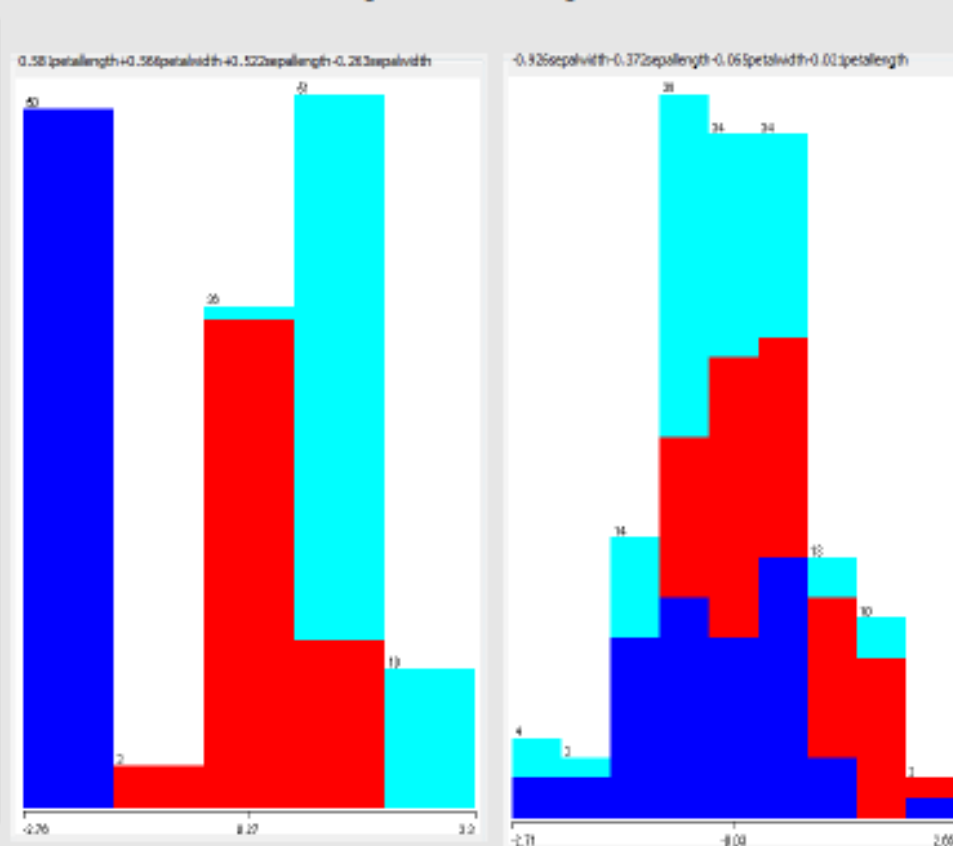
PCA

- Ex: conjunto de dados iris

Original



PrincipalComponents



Seleção de Atributos

■ Permite

- ☐ Identificar atributos importantes
- ☐ Melhorar o desempenho de várias técnicas de AM
- ☐ Reduzir necessidade de memória e tempo de processamento
- ☐ Eliminar atributos irrelevante e reduzir ruído
- ☐ Lidar com maldição da dimensionalidade
- ☐ Simplificar o modelo gerado
- ☐ Tornando mais fácil sua compreensão
- ☐ Facilitar a visualização dos dados
- ☐ Reduzir o custo de coleta dos dados

Seleção de Atributos

- Na prática, é difícil identificar atributos passíveis de eliminação
 - Redundantes
 - Irrelevantes
- Algumas razões
 - Número grande de exemplos
 - Número grande de atributos
 - Relações complexas entre atributos
- Neste caso necessidade de técnicas automáticas para seleção de atributos

Seleção de Atributos

- Objetivo: encontrar subconjunto ótimo de atributos de acordo com algum critério
- Técnicas podem ser classificadas de diferentes formas:
 - Quanto à maneira de avaliar os atributos selecionados
 - Interação com o algoritmo de AM
 - Quanto considerar cada atributo individualmente ou em subconjuntos
 - Quanto à medida de importância usada
 - Quanto ao uso ou não da classe
 - Em conjuntos de dados supervisionados

Seleção de Atributos

- Quanto à avaliação dos atributos:
- Técnicas podem estar integradas a um algoritmo de indução ou serem independentes do algoritmo

Embutida

- Seleção é embutida ou integrada no próprio algoritmo de AM

Filtro

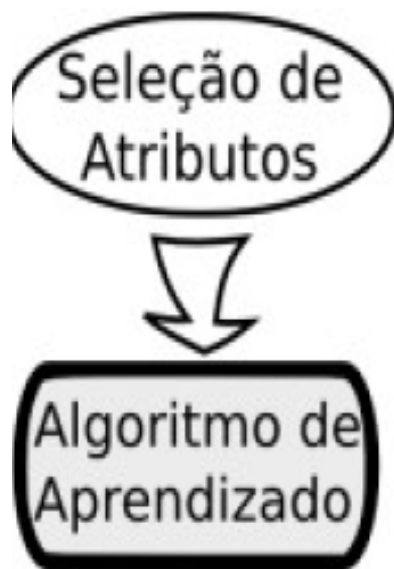
- Filtra atributos, sem levar em consideração algoritmo de AM que os utilizará
- Verificam características dos dados

Wrapper

- Usa algoritmo de AM como “caixa-preta” para a seleção, que avalia os subconjuntos

Seleção de atributos

■ Abordagens de avaliação de atributos



Filtro



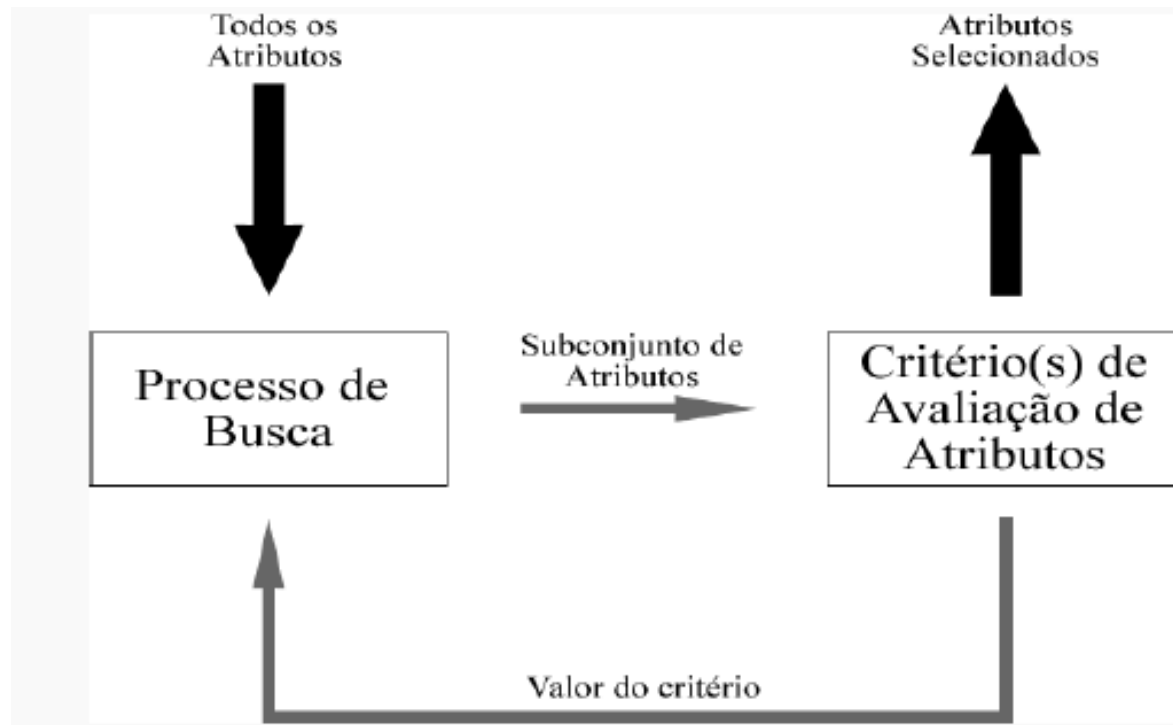
Wrapper



Embutida

Seleção de atributos:filtro

- Seleção de atributos independente do algoritmo AM
 - Ex: verificando a correlação entre os atributos



Seleção de atributos: filtro

■ Vantagens:

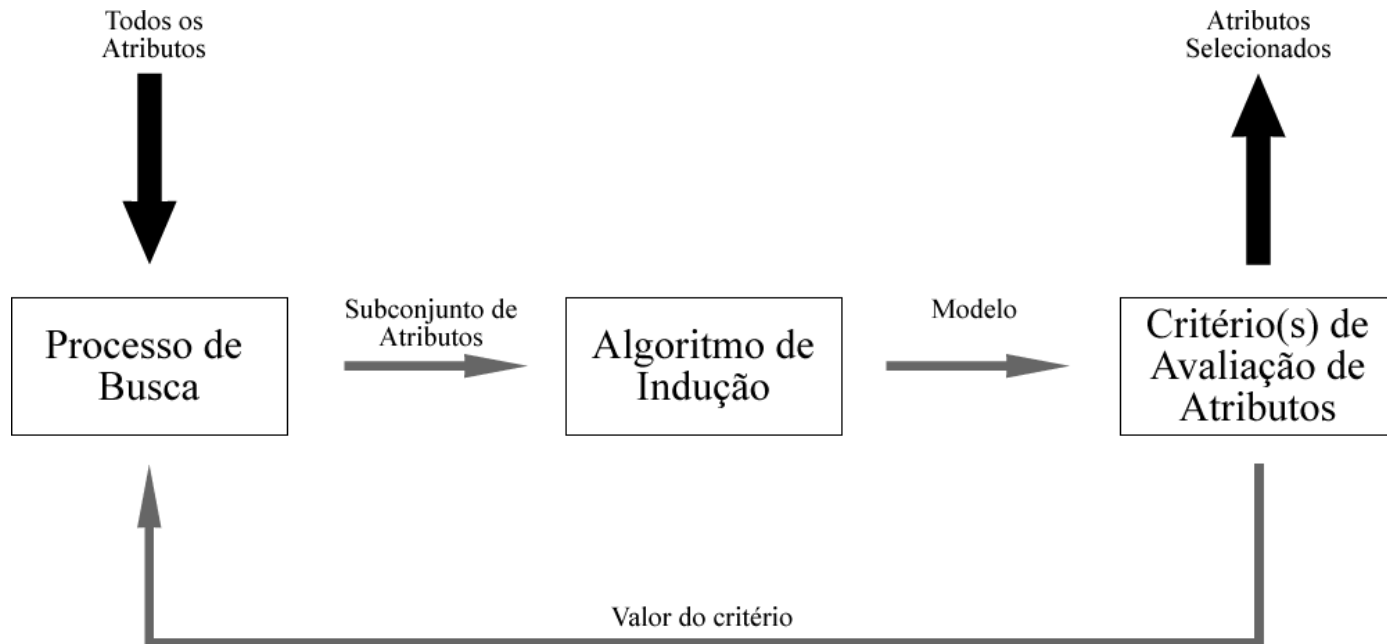
- ☐ Geralmente mais rápida
- ☐ Heurísticas para avaliar os subconjuntos são pouco custosas
- ☐ Características selecionadas podem ser usadas por diferentes algoritmos de AM
- ☐ Conseguem lidar com uma grande quantidade de dados

■ Desvantagens

- ☐ Independência do algoritmo de AM
 - Não considerar o viés do algoritmo de AM pode levar a modelos pouco eficientes

Seleção de atributos: *wrapper*

- Usa algoritmo de AM para avaliar os atributos
 - Ex. desempenho preditivo usando o subconjunto



Seleção de atributos: *wrapper*

■ Vantagens:

- ☐ Melhor subconjunto para cada algoritmo de AM
- ☐ Pode selecionar menos atributos também
- ☐ Geralmente leva a modelos com melhor desempenho preditivo/descritivo

■ Desvantagens:

- ☐ Risco de overfitting
- ☐ Subconjunto depende do algoritmo de AM
- ☐ Para cada novo algoritmo, deve ser repetido
- ☐ Custo computacional elevado

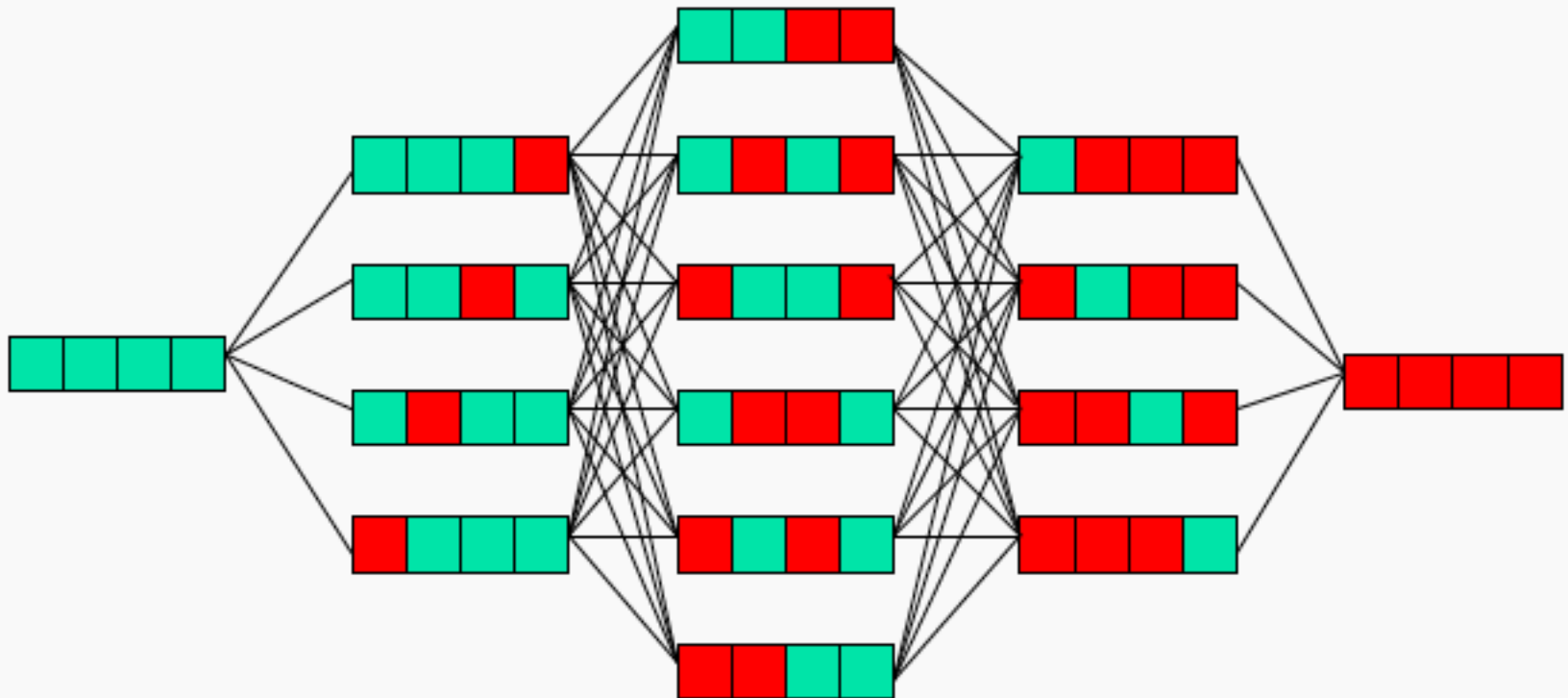
Seleção de atributos

- Quanto a seleção ser individual ou coletiva:
 - Ordenação
 - Seleção é individual
 - Atributos são ordenados de acordo com relevância segundo algum critério e atributos no topo são selecionados (ranking)
 - Tende a selecionar atributos correlacionados
 - Necessidade de definir limiar de seleção
 - Seleção de subconjunto
 - Seleciona subconjunto dos atributos originais que melhor represente
 - Verifica como atributos atuam de forma coletiva, em conjunto
 - É computacionalmente mais cara
 - Pode ser formulada como problema de busca

Seleção de subconjuntos de atributos

■ Busca:

- Cada ponto no espaço de busca pode ser visto como um possível subconjunto de atributos



Seleção de subconjuntos de atributos

- Deve-se definir na busca:
 - ☐ Ponto(s) de partida ou direção da busca
 - ☐ Estratégia de busca
 - ☐ Critério usado na avaliação dos subconjuntos
 - ☐ Abordagens filtro, wrapper ou embutida
 - ☐ Critério de parada

Seleção de subconjuntos de atributos

■ Pontos de partida/direção da busca

Para trás (*backward*)

- Começa com todos atributos
- Remove um por vez

Para frente (*forward*)

- Começa sem nenhum atributo
- Inclui um por vez

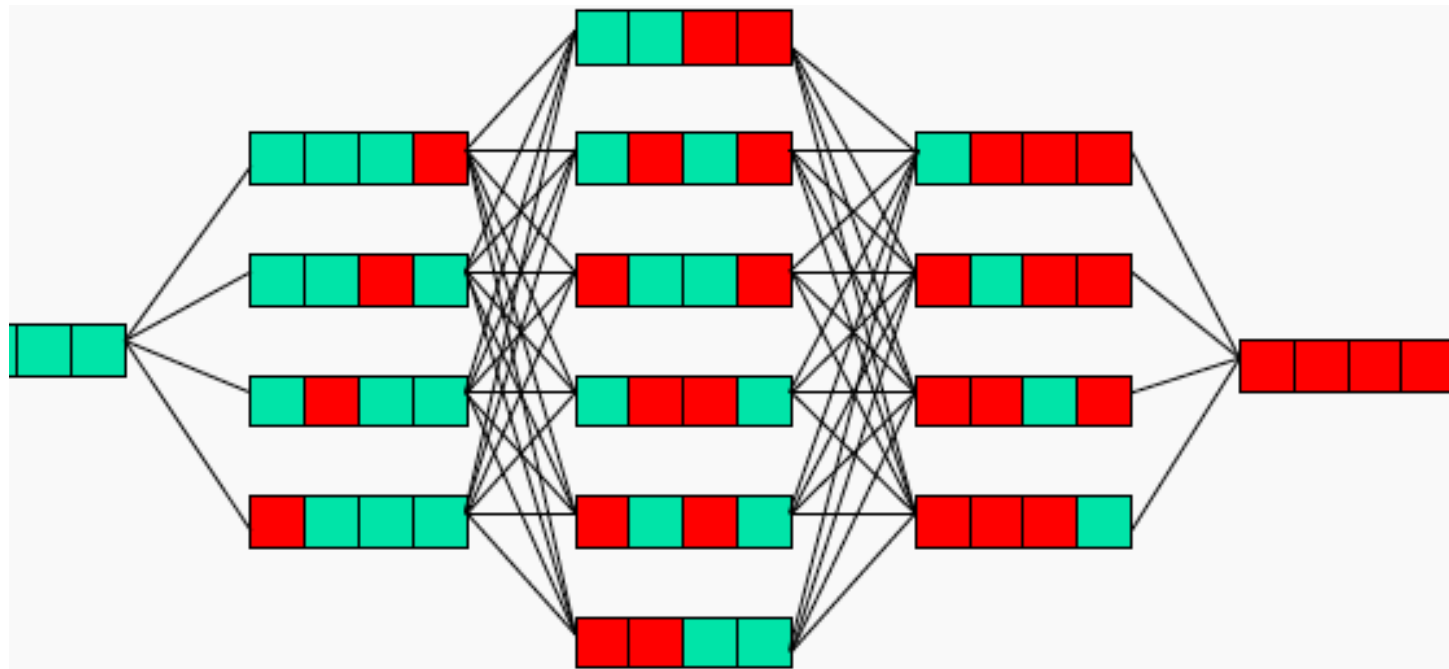
Bidirecional

- Pode começar em qualquer ponto
- Atributos podem ser adicionados ou removidos

Estocástica (*random*)

- Ponto de partida da busca e atributos a serem adicionados/removidos são decididos estocasticamente

Seleção de subconjuntos de atributos



Backward
Feedforward
Bidirecional

Seleção de subconjuntos de atributos

■ Estratégias de busca possíveis

Busca completa

- Avalia todos os possíveis subconjuntos
- Pode não necessitar visitar todos exhaustivamente

Busca heurística

- Utiliza regras e métodos para conduzir a busca
- Não garante encontrar a solução ótima

Busca não determinística

- Relacionada com a geração estocástica
- Boa solução pode ser encontrada antes do final da busca
- Não garante encontrar o ótimo

Seleção de subconjuntos de atributos

- Relações entre sentidos e estratégias de busca

Sentido	Estratégia		
	Completa	Heurística	Não determinística
<i>Forward</i>	Sim	Sim	Não
<i>Backward</i>	Sim	Sim	Não
Estocástico	Não	Sim	Sim

Seleção de subconjuntos de atributos

- Critérios de parada possíveis:
 - Quando todos os subconjuntos forem testados (exaustiva)
 - Quando um número máximo de alternativas é testado
 - Quando atinge um número de atributos desejado
 - Enquanto adição/remoção de atributos não deteriora desempenho do modelo de AM



Seleção de atributos

- Tipos de critérios para avaliação da importância dos atributos

Consistência

- Indicam se subconjunto permite construir projeção consistente dos dados
- Possibilitar a construção de hipóteses lógicas consistentes em um conjunto de dados

Dependência

- Mensuram capacidade de prever o valor de um atributo a partir do valor de outro atributo

Seleção de atributos

- Tipos de critérios para avaliação da importância dos atributos

Distância

- Exemplos próximos têm alguma relação

Informação

- Atributos mais importantes leva a maior ganho de informação

Precisão

- Atributos mais importantes levam a melhor desempenho preditivo do modelo
- Geralmente associada à abordagem *wrapper*

Seleção de atributos

■ Ex conjunto de dados da iris

```
Search Method:
  Attribute ranking.
Attribute Evaluator (supervised, Class (nominal): 5 class):
  Information Gain Ranking Filter

Ranked attributes:
1.418  3 petallength
1.378  4 petalwidth
0.698  1 sepallength
0.376  2 sepalwidth

Selected attributes: 3,4,1,2 : 4
```



Considerações Finais

■ Pré-processamento

- ☐ Integração de dados
- ☐ Amostragem
- ☐ Dados desbalanceados
- ☐ Limpeza de dados
- ☐ Transformação de dados
- ☐ Redução do número de atributos

Referências

- Softwares utilizados
 - Weka
- Material de aula profa. Ana Carolina Lorena
- Livro:
 - Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina