Pré-processamento de Dados

Conteúdo

- Motivação
- Atividades
  - Limpeza de dados
  - Integração de dados
  - Transformação de dados
  - Redução de dados
- Considerações Finais

Motivação

Cenário atual

- Bancos de dados atuais são grandes, ocupam vários gigabytes de espaço em disco
  - Muitas vezes contém inconsistências
- Como pré-processar os dados de forma a melhorar a qualidade dos dados e, consequentemente, do processo de Aprendizado?

Técnicas existentes

- Limpeza de dados
  - Remove “ruído” e inconsistência
- Integração de dados
  - Agrupa dados heterogêneos em um formato homogêneo
- Transformação de dados
  - Melhora a eficiência ou permite a aplicação de algoritmos de mineração
- Redução de dados
  - Agrupamento ou eliminação de dados irrelevantes

Limpeza de Dados

Limpeza de dados

- Bancos de dados são suscetíveis a armazenar informações irrelevantes, inconsistentes ou mesmo não armazenar dados importantes
  - Informações não disponíveis no momento de inserção
  - Informações não consideradas importantes
  - Falha no sistema ou equipamentos
  - Falha humana

Ausência de Dados

- Ignorar a tupla
  - Necessário apenas quando a tupla tem vários atributos não preenchidos
- Completar manualmente
  - Nem sempre é possível
- Uso de um valor especial (“unkown” ou ∞)
  - O algoritmo pode equivocadamente considerá-los como um novo valor
- Com estas estratégias não há possibilidade de se inserir dados errôneos ou tendenciosos (“*bias*”)

Ausência de Dados

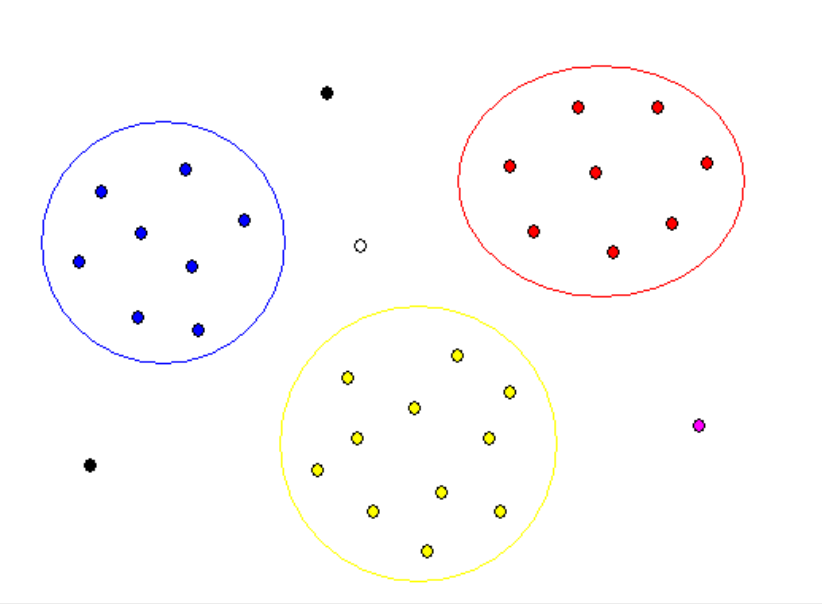
- Uso do valor médio do atributo
- Uso da média para todas tuplas pertencentes a uma mesma classe
- Uso do valor mais provável
  - Valor este obtido a partir de uma árvore de decisão que utiliza valores de outros atributos
- Com estas estratégias há possibilidade de se inserir dados incorretos, pois utiliza valores estimados

Ruído

- Ruído é qualquer erro ou variação em medições
  - Outliers: valor que distoa significativamente do esperado
- Pode-se “suavizar” os dados a fim de amenizar suas consequências
- Estratégias
  - Agrupamento: divisão em grupos
  - Regressão linear: função tal que uma variável pode ser obtida a partir de outra

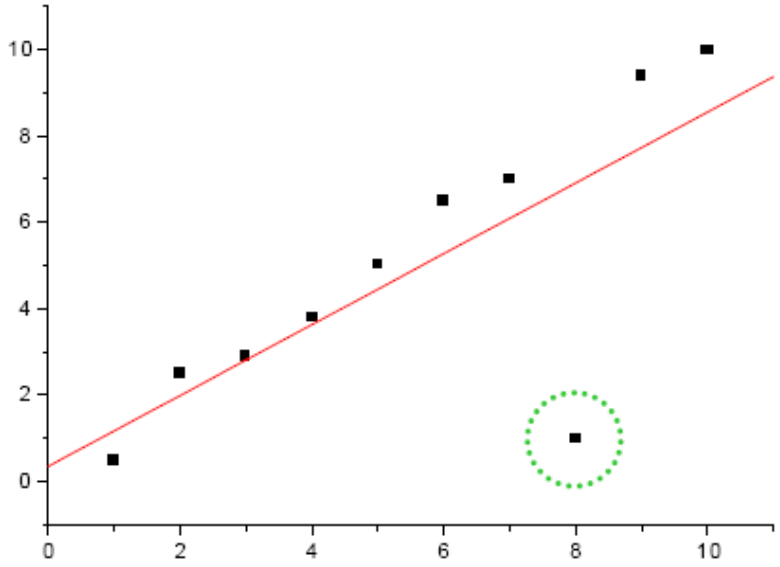
Ruído

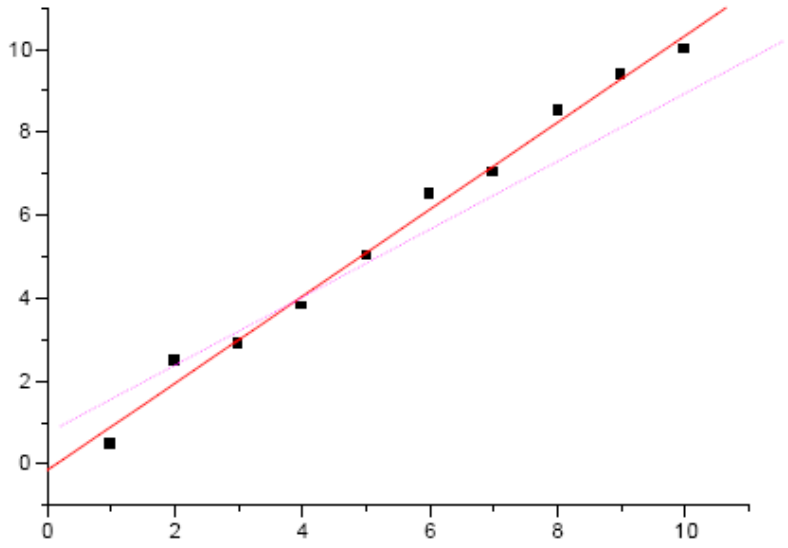
- Agrupamento
  - Valores não agrupados são considerados outliers



Ruído

- Regressão Linear
  - Valores não pertencentes a função com certo grau de incerteza são considerados outliers





Integração de Dados

Integração de dados

- Envolve união de dados de fontes distintas
  - BD relacionais, arquivos de texto, arquivos XML...
- Como mapear entidades entre BD distintos
  - Ontologias
  - Uso de metadados

Integração de dados

- Algumas vezes, pode haver redundância de dados na integração entre fontes, quando um valor de uma fonte pode ser inferido a partir de outra tabela

Integração de dados

- Redundâncias podem ser detectadas por uma análise de correlação
  - Se houver uma correlação positiva, o valor de A aumenta se o valor de B aumenta
  - Se o valor é zero, os dados são independentes
  - Se a correlação é negativa, o valor de um aumenta quando o do outro diminui

Integração de dados

Análise de correlação

$$r_{A,E} = \frac{\sum (A - \hat{A})(E - \hat{E})}{(n - 1)\,\sigma_A\,\sigma_E}$$

$\hat{A}$ e $\hat{E}$ → Valores médios dos atributos A e E

σ_A e σ_E → Desvio padrão de A e E

Integração de dados

- Outros fatores a serem considerados
  - Duplicação
    - Tuplas repetidas provenientes de fontes distintas
  - Resolução de conflitos de valores
    - Ex: valores em moedas distintas
    - Ex: custos de diárias de hotéis na mesma moeda, mas com taxas de serviços distintos inclusas

Transformação de Dados

Transformação de dados

- Alguns algoritmos só trabalham com certos tipos de dados
  - Alguns algoritmos probabilísticos não podem ser usados com valores numéricos contínuos
- Necessário fazer transformação nos dados

Transformação de dados

- Tipos de Transformações
  - Normalização
  - Discretização
  - Adaptação
    - Valores nominais para binarios ou ordinais
    - Datas para intervalos nominais ou ordinais

Normalização

- Transformação de valores para uma escala determinada (por ex., de 0.0 a 1.0)
- Importante para algoritmos de redes neurais, pois impede que os valores com faixa de valores grandes (p. ex., salário) se sobreponham aos valores menores (como atributos binários ou idade)
- É importante armazenar os parâmetros para que os dados futuros possam ser normalizados

Estratégias de normalização

- Max Absoluto
  - Transformação linear baseada nos valores máximo absoulto do atributo. Se valores fora desta faixa forem inseridos futuramente, serão mapeados para valores fora da faixa alvo [-1.0, 1.0]
- Exemplo

Suponha que o máximo absoluto de um atributo seja 98000.

O novo valor para 73600 será

$$\frac{73600}{98000} = 0.751$$

Estratégias de normalização

- Min-max
  - Transformação linear baseada nos valores mínimo e máximo do atributo. Se valores fora desta faixa forem inseridos futuramente, serão mapeados para valores fora da faixa alvo [0.0, 1.0]
- Exemplo

Suponha que os valores mínimo e máximo de um atributo são R$12000 e R$98000. O atributo deve ser mapeado para [0.0, 1.0]

O novo valor para R$73600 será

$$\frac{73600 - 12000}{98000 - 12000}(1.0 - 0) + 0 = 0.716$$

Estratégias de normalização

- Z-score
  - Os valores são normalizados baseados na média e desvio padrão da caracerística
    - Média normalizada é zero
    - Desvio padrão normalizado é 1
- Exemplo

  O valor médio de uma característica é R$54000 e o desvio padrão é R$16000. O novo valor para R$73600 será então

  $$\frac{73.600 - 54.000}{16.000} = 1,225$$

Discretização

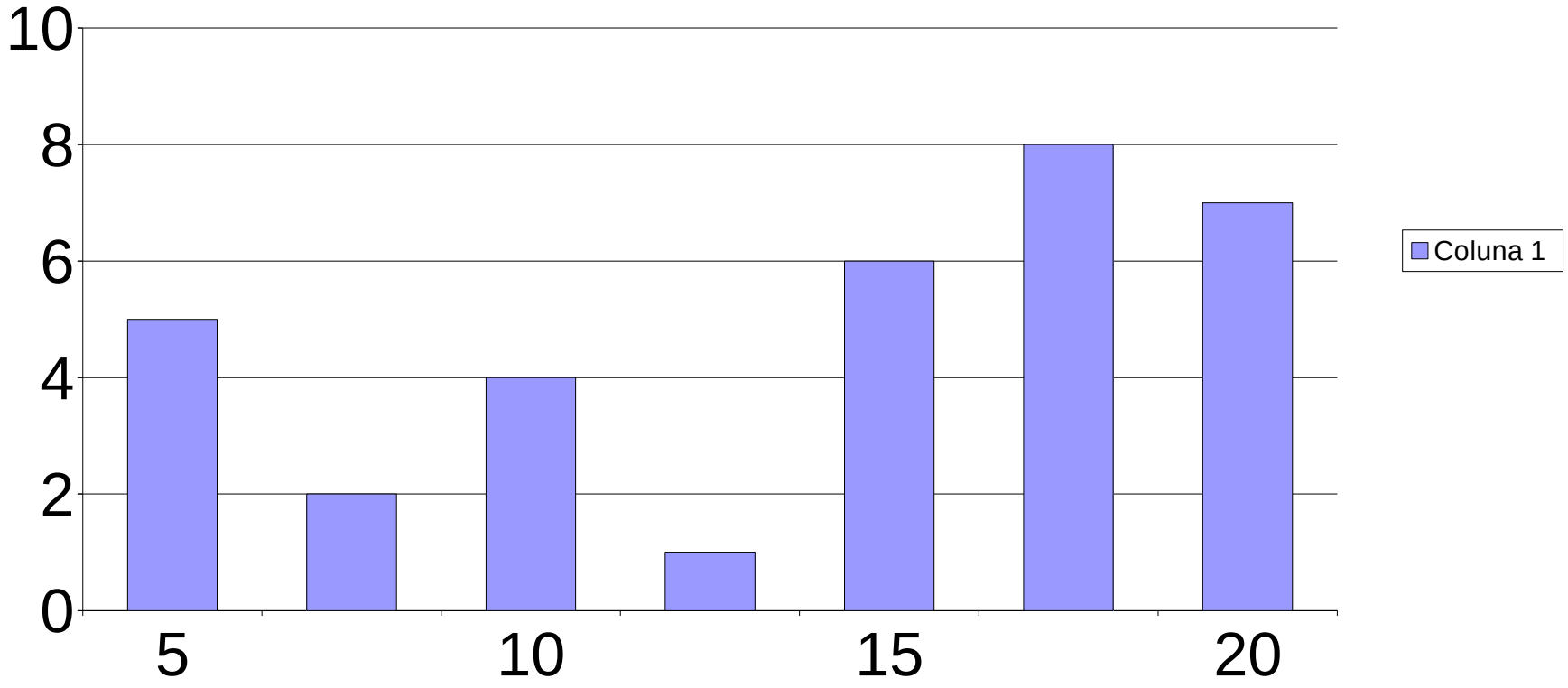
- Transformação de valores contínuos para representação discreta
  - Intervalo
  - Valor representativo
- Há perda de informação
  - Valor real não é usado
  - Pode perder a noção de ordem

Discretização

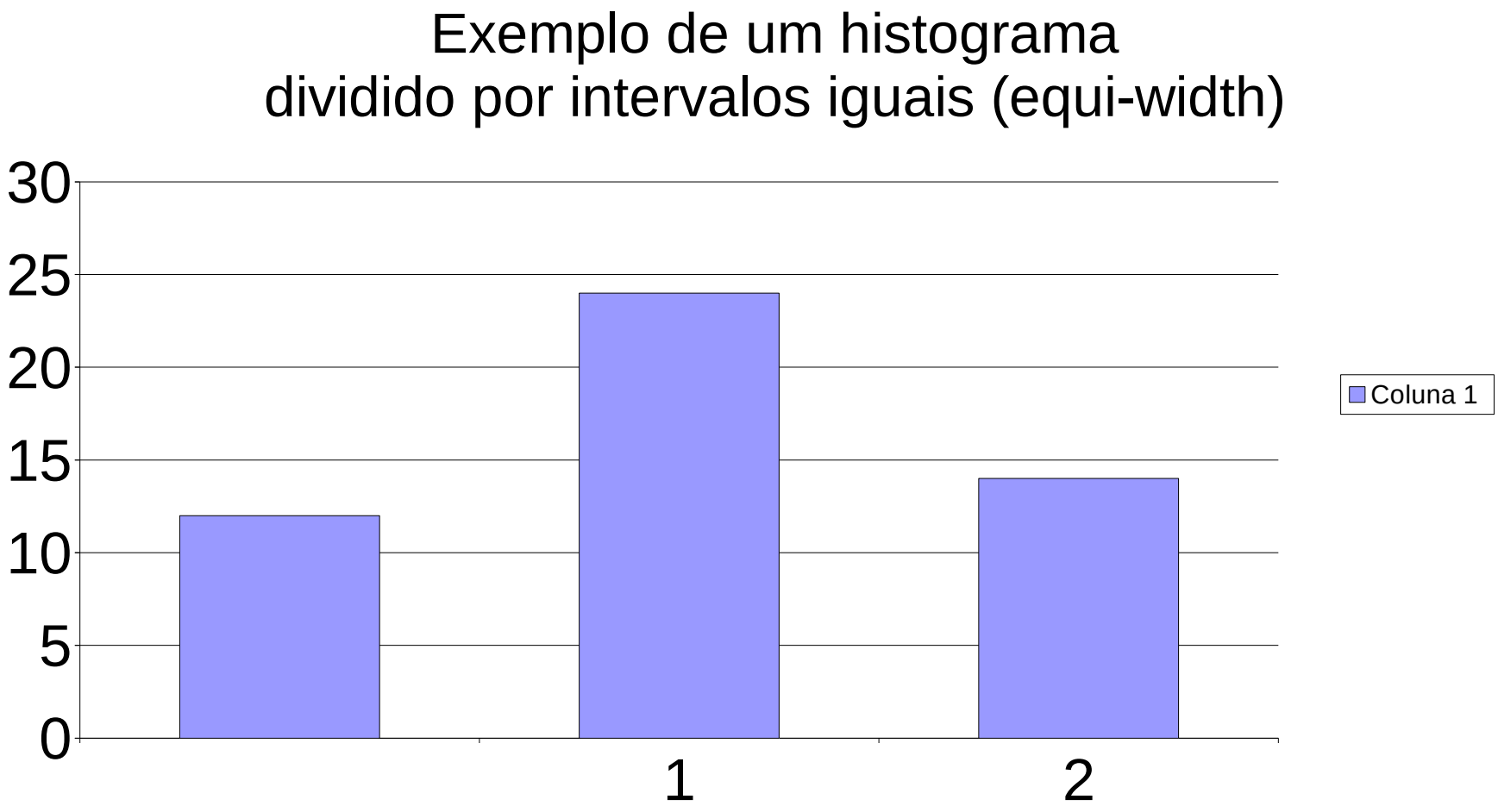
- Tipos
  - Hierarquia de Conceitos
    - Intervalos definidos manualmente segundo conhecimento do domínio
      - 0-50 KWh → Baixo Consumo
      - 50-200 KWh → Médio Consumo
      - 200-~ KWh → Alto Consumo
  - Particionamento por Larguras Iguais
    - Divisão em N intervalos de mesma largura
  - Particionamento por Freqüências Iguais
    - Divisão em N intervalos com o mesmo número de exemplos

Histogramas

Exemplo de um histograma "singleton":
cada coluna corresponde a um par de valor-frequência



Histogramas



Discretização

- Tipos
  - Hierarquia de Conceitos
    - Intervalos definidos manualmente segundo conhecimento do domínio
      - 0-50 KWh → Baixo Consumo
      - 50-200 KWh → Médio Consumo
      - 200-~ KWh → Alto Consumo
  - Particionamento por Larguras Iguais
    - Divisão em N intervalos de mesma largura
  - Particionamento por Freqüências Iguais
    - Divisão em N intervalos com o mesmo número de exemplos

Adaptação

- Mapeando nominais para binários
- Cores (Verde, Azul, Amarelo)
  - Variáveis Binárias: Verde, Azul, Vermelho
  - Azul: 0, 1, 0
  - Vermelho: 0, 0, 1
- Mapeando nominais para ordinais
  - Tamanho da Camisa
    - P, M, G, XG
      - P → 0
      - M → 1
      - G → 2
      - XG → 3

Redução de Dados

Redução de dados

- Diminui o volume de dados, agilizando a aplicação dos algoritmos e melhorando eventualmente seu desempenho
- Técnicas
  - Redução no número de características
  - Redução no número de exemplos
  - Redução de valores das características

Redução no número de Características

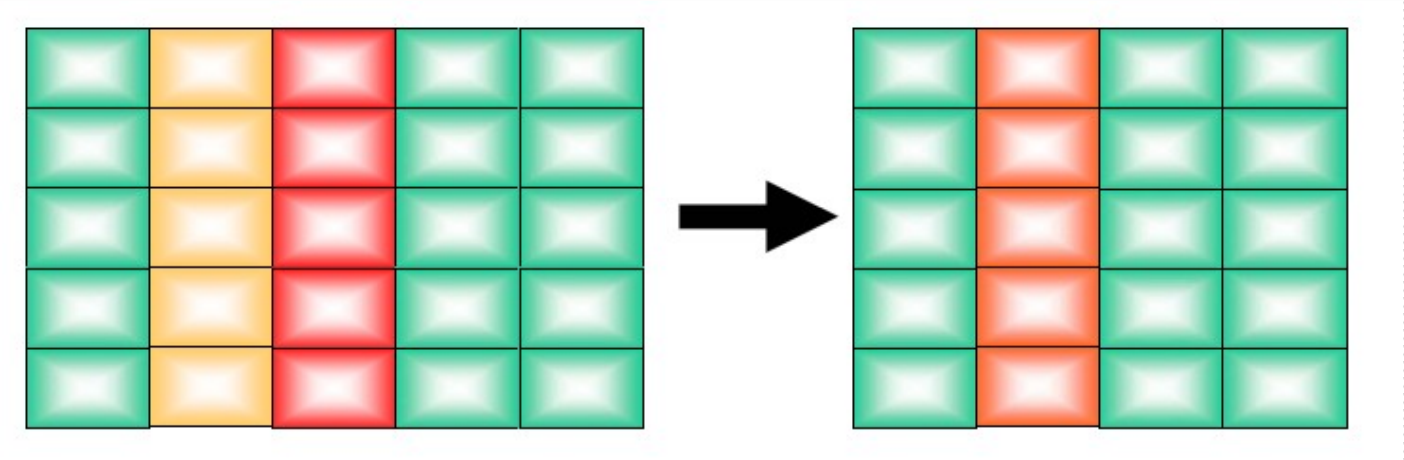
- Também chamada Redução de Dimensionalidade
  - Remoção de Características irrelevantes para a análise
  - Extração de Características
  - Seleção de Características

Redução de Dimensionalidade

- Remoção de características irrelevantes para a análise
- Exemplo
  - O número do telefone de um cliente provavelmente é irrelevante para se descobrir preferências

Redução de Dimensionalidade

- Extração de Características
  - Característica que combina outros
    - Sexo e Faixa Etária X Combinado
    - (M, Idoso) X SenhorIdoso



Extração de Características

cabeça	*corpo*	classe
quadrada	quadrada	amigo
triangular	triangular	amigo
redonda	triangular	inimigo
quadrada	redonda	inimigo
triangular	quadrada	inimigo
triangular	redonda	inimigo

cabeça	*corpo*	*mesma_forma*	classe
quadrada	quadrada	v	amigo
triangular	triangular	v	amigo
redonda	redonda	f	inimigo
quadrada	quadrada	f	inimigo
triangular	triangular	f	inimigo
redonda	redonda	f	inimigo

Se mesma_forma = v
então amigo.
Se mesma_forma = f
então inimigo.

Extração de Características

- Compressão de dados
  - Transformadas Wavelet ou Fourier
  - Análise de componentes principais

Extração de Características

- Transformadas Wavelet ou Fourier
  - Transforma dados originais em um vetor de mesmo comprimento
    - Uma vantagem está no fato de poder ser definido um “valor de corte”, onde os dados abaixo desse valor não são armazenados
    - Vetor esparso, mais fácil de ser processado
    - Extração de Características
      - » Banda de Frequências
        - – Picos, Média, RMS

Extração de Características

- Análise de componentes principais
  - Busca de vetores que melhor representam os dados
    - Dados são normalizados
    - Vetores perpendiculares são criados, cada um correspondendo a um atributo
    - Ordena-se componentes por “significância”

Redução de Dimensionalidade

- Seleção de Características
  - Correlação de Características
  - Análise de Valor Preditivo
  - Seleção Embutida no Método
    - Árvore de decisão
    - Redes Neurais
  - Busca
  - Híbrida

Seleção de Características

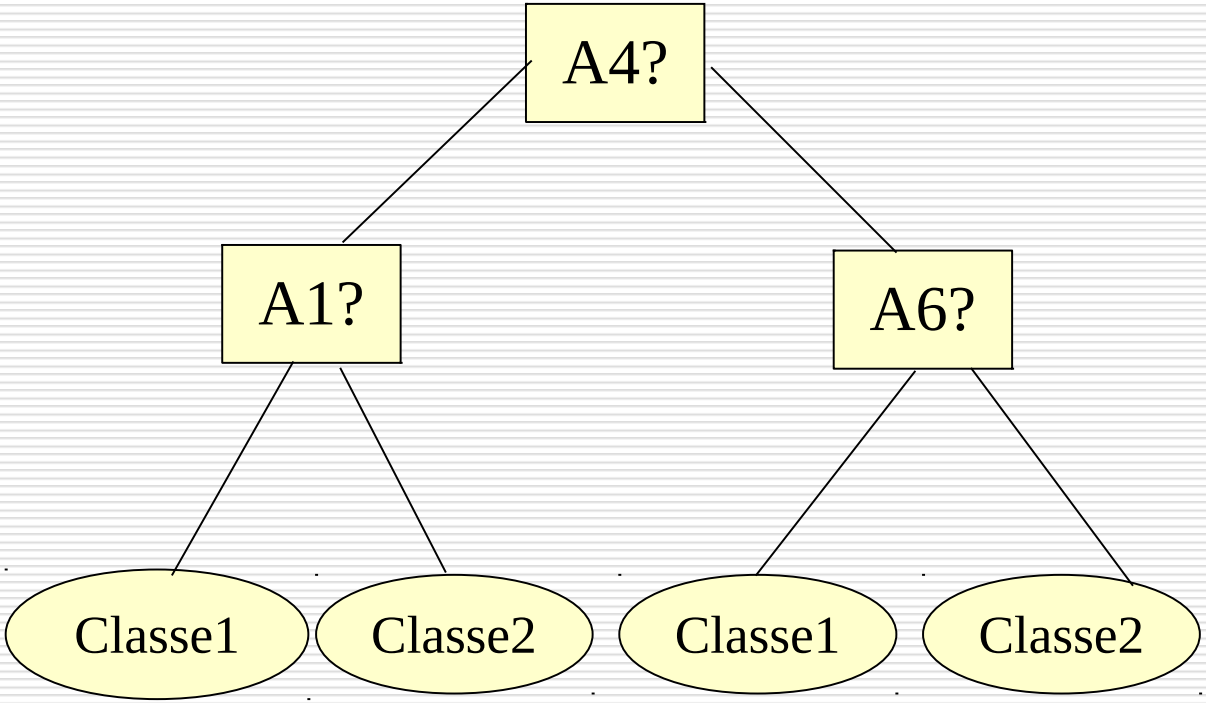
- Características redundantes correlacionados
  - Idade e Problemas de Saúde
  - Uso de métricas estatísticas
    - Correlação
    - Quiquadrado

Análise de Valor Preditivo

- Uso de métricas estatísticas
  - Correlação
  - Métricas derivadas de tabela de contigência
  - Ordenação dos atributos e seleção dos que possuem mais poder preditivo
  - Pode acontecer de combinação de atributos pouco preditivos ser melhor do que combinação de atributos muito preditivos

Seleção Embutida no Método

- Árvore de Decisão



Árvore de decisão

Atributo-valor					classe
sorri	*segura*	*tem-gravata*	*cabeça*	*corpo*	
sim	balão	sim	quadrada	quadrada	amigo
sim	bandeira	sim	triangular	triangular	amigo
sim	espada	sim	redonda	triangular	inimigo
sim	espada	sim	quadrada	redonda	inimigo
não	espada	não	triangular	quadrada	inimigo
não	bandeira	não	triangular	redonda	inimigo

Árvore de decisão

Regras:

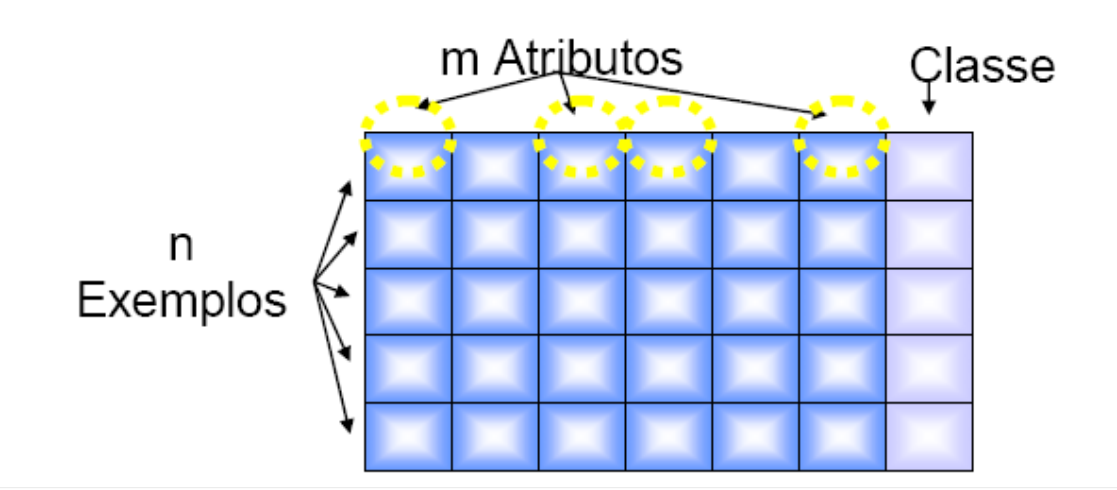
Se sorri = sim *e* segura = espada
    então inimigo.

Se sorri = sim *e* segura = balão ou bandeira
    então amigo.

Se sorri = não
    então inimigo.

Seleção de atributos

- Busca
  - Há 2^n possibilidades de subconjuntos quando se tem n atributos
  - Isso torna necessária uma busca heurística



Busca

- Abordagens
  - Filtro
    - Usa análise do valor preditivo para avaliar os subconjuntos de características
    - Avaliação mais rápida e independente do método de aprendizado
  - Wrapper
    - Usa o próprio método classificador para avaliar os subconjuntos de características
    - Seleção é customizada para o método de aprendizado

Busca

- Redução no número de atributos
  - Suponha um conjunto {A1,A2,A3,A4,A5,A6} de atributos a serem avaliados. Temos duas estratégias:
    - Foward selection: começa-se com um conjunto vazio
      - Conjunto inicial: {}
      - {} → {A1} → {A1,A4} → {A1,A4,A6}
    - Backward elimination: começa-se com todos atributos
      - Conjunto inicial: {A1,A2,A3,A4,A5,A6}
      - {A1,A3,A4,A5,A6} → {A1,A4,A5,A6} → {A1,A4,A6}

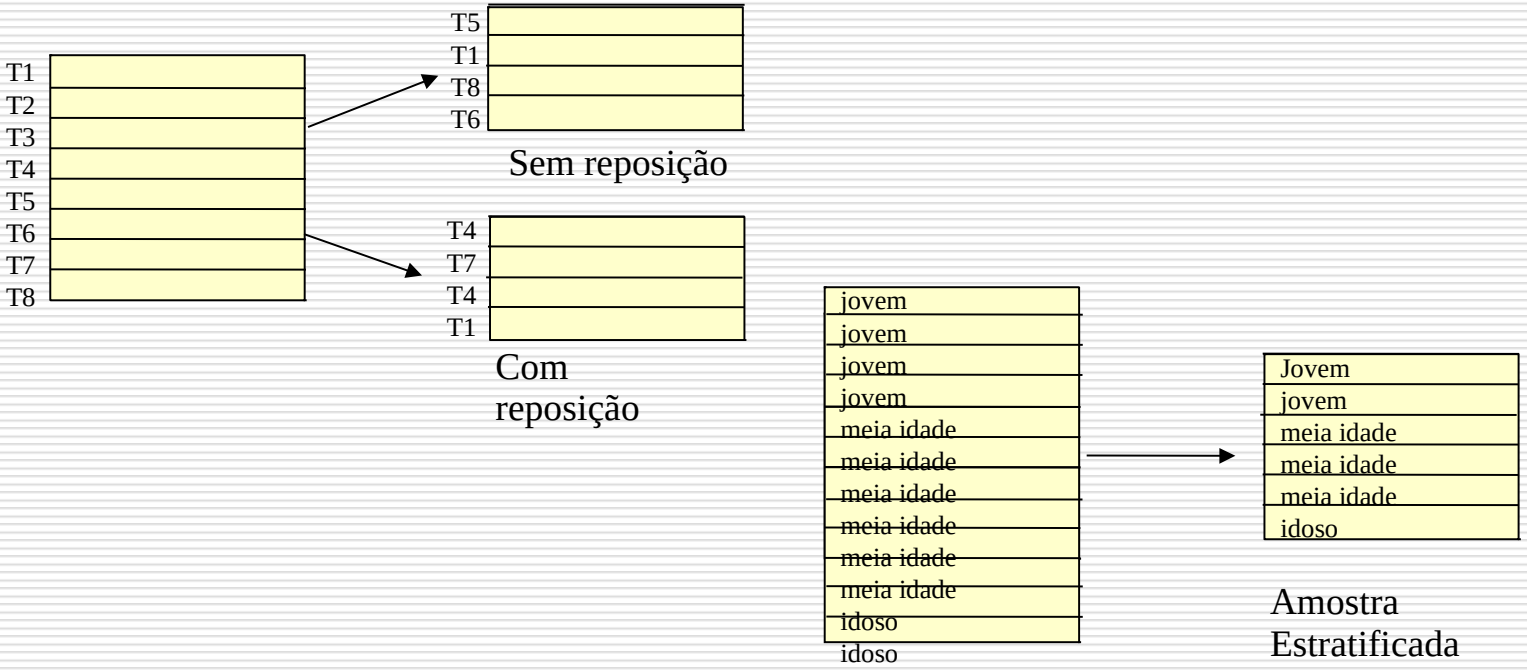
Híbrida

- Uso de análise de valor preditivo para ranqueamento das características
- Seleção do subconjunto de características com melhor valor preditivo
  - Definição arbitrária do número de atributos ou de limiar de valor preditivo
- Aplicação de busca para escolha do subconjunto final de características
  - Normalmente com método wrapper

Redução de exemplos

- Amostragem
  - Escolha aletória
    - Sem reposição
    - Com reposição
  - Estratificação

Amostragem

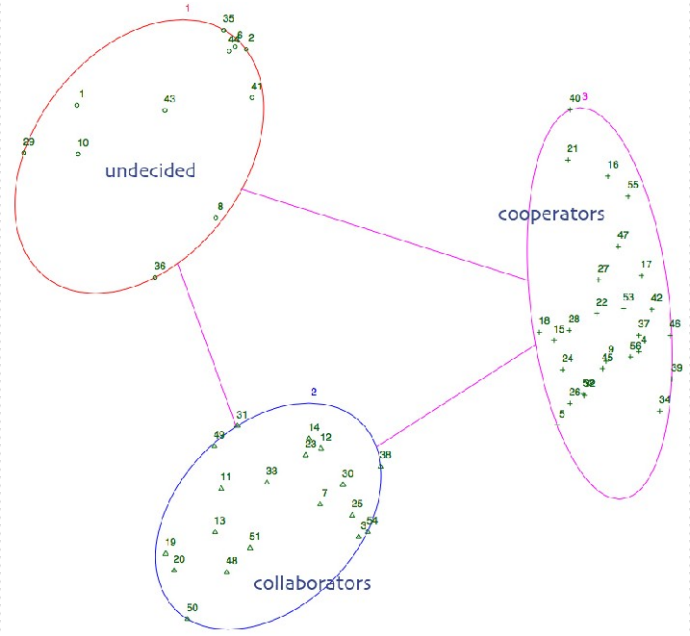

T1
T2
T3
T4
T5
T6
T7
T8
T5
T1
T8
T6
Sem reposição
T4
T7
T4
T1
Com
reposição
jovem
jovem
jovem
jovem
meia idade
meia idade
meia idade
meia idade
meia idade
meia idade
idoso
idoso
Jovem
jovem
meia idade
meia idade
meia idade
idoso
Amostra
Estratificada

Redução dos valores de atributos

- Discretização
- Agrupamento
  - Substituição por valor representativo do grupo
    - Média
    - Mais freqüente

Agrupamento

- Avalia quão similares os valores são
  - Uso de função de distância
    - Euclidiana
  - Medidas de qualidade
    - Diâmetro e distância entre centros



Considerações Finais

- Pré-processamento é etapa decisiva para o sucesso
- Talvez a mais importante
  - “Se dados são bons, qualquer técnica apresenta bons resultados”
- Pode consumir mais de 50% do processo de aprendizado de máquina