Aula 6 – Mineração de Dados
Classificação - Parte 1

Profa. Elaine Faria

UFU

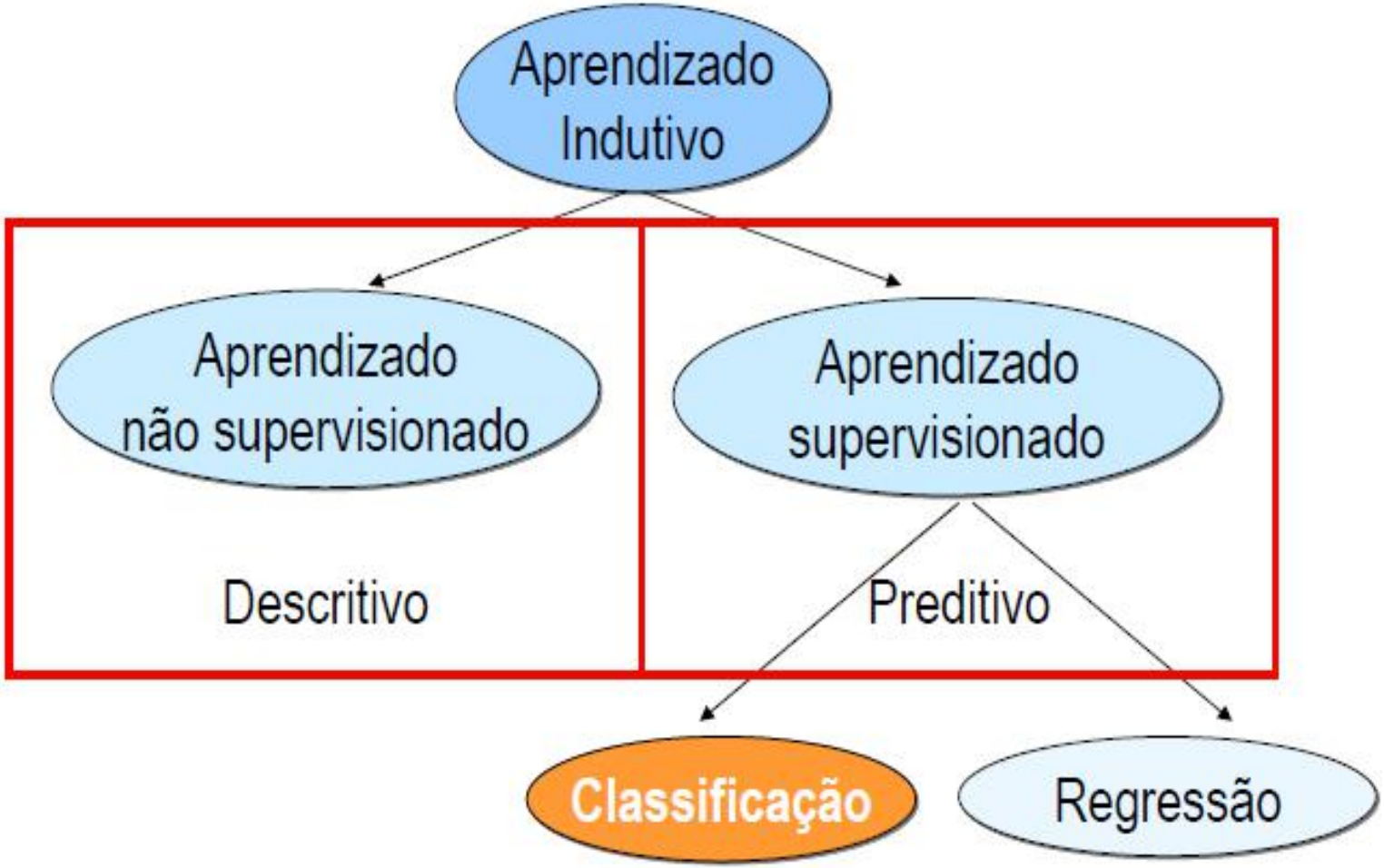
Agradecimentos

Este material é baseado

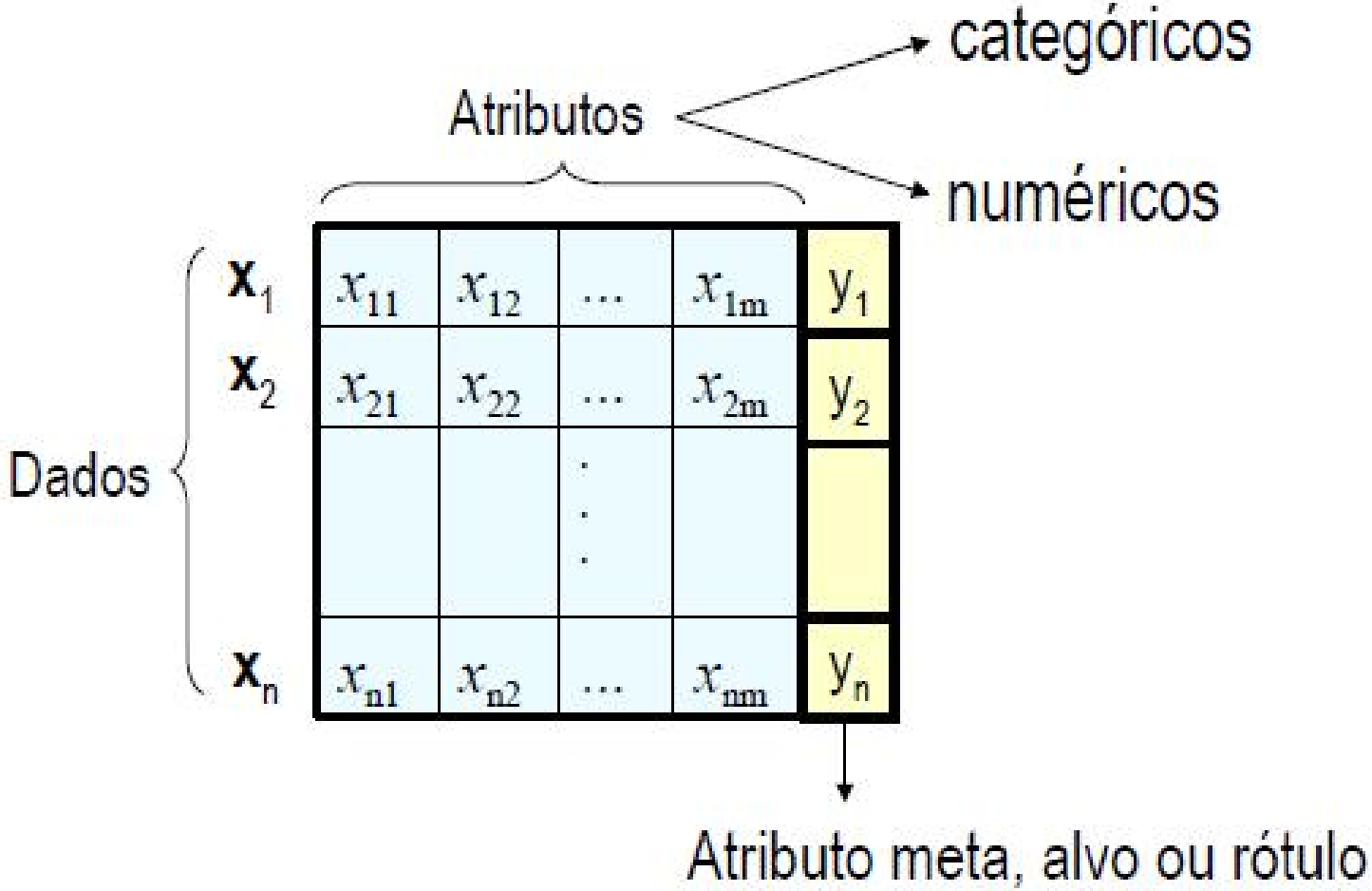
- No livro Tan et al, 2006
- Nos slides do prof. Andre C. P. L. F. Carvalho
- No livro Facelli et al, 2011

- Agradecimentos
  - Ao professor André C. P. L. F. Carvalho que gentilmente cedeu seus slides
  - Ao professor Tan que disponibilizou vários slides do seu livro

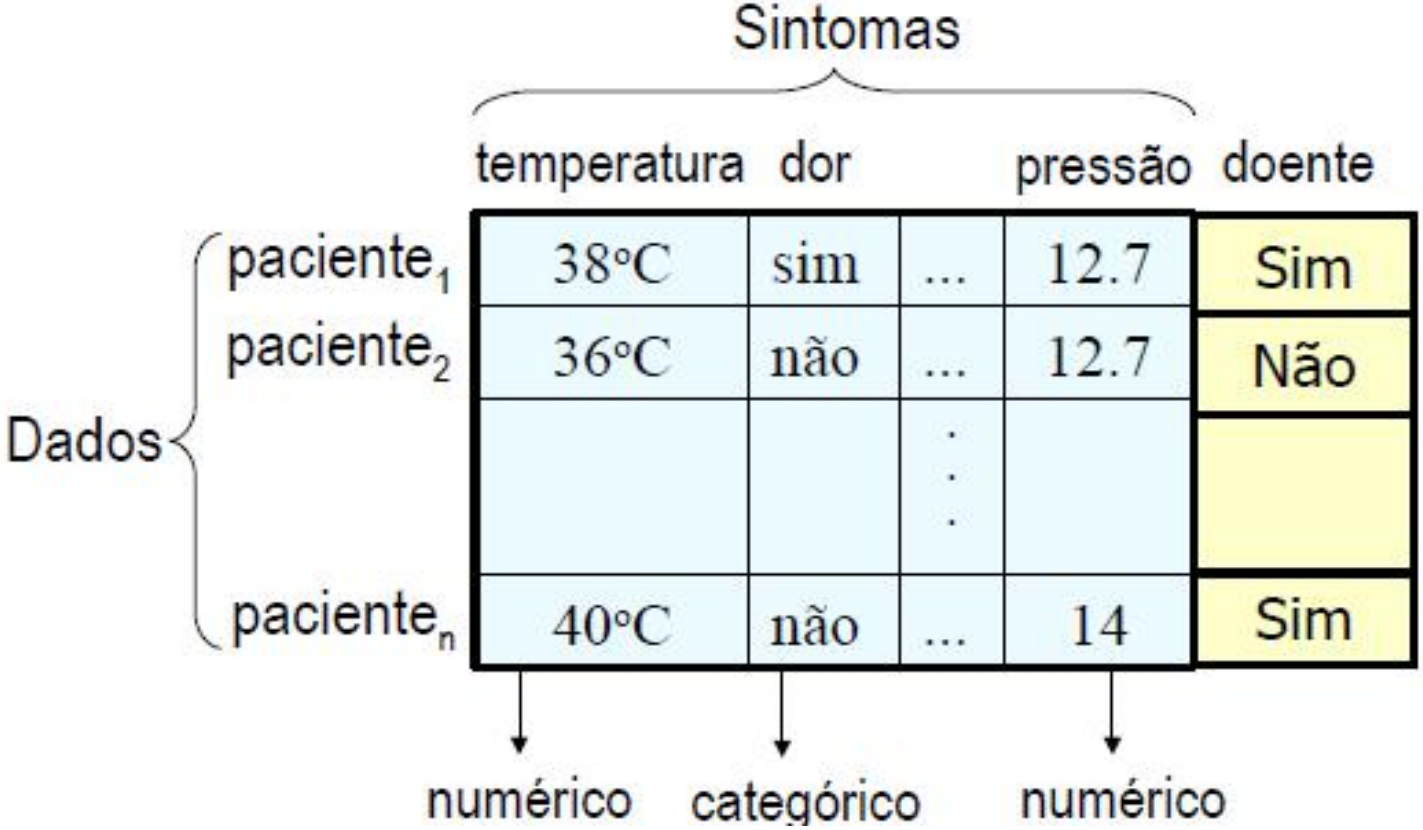
Abordagens de Aprendizado



Dados


categóricos
Atributos
numéricos
Dados
$\mathbf{x}_1$
$\mathbf{x}_2$
$\mathbf{x}_n$
x_{11} x_{12} … x_{1m} y_1
x_{21} x_{22} … x_{2m} y_2
x_{n1} x_{n2} … x_{nm} y_n
Atributo meta, alvo ou rótulo

Dados - Diagnóstico de uma doença



Classificação

- Atribuir objetos a uma dentre várias categorias pré-definidas
- Ex.:
  - Classificação de letras e números
  - Reconhecimento de faces
  - Análise de crédito
  - Diagnóstico médico

Classificação

- Definição
  - Dado um conjunto de treinamento
    - Em que cada exemplo possui um conjunto de atributos
      - Um deles o rótulo ou classe
  - Encontrar um modelo para o atributo classe como uma função dos valores de entrada
    - Função alvo ou modelo de classificação
      - Assume valores em um conjunto discreto

Classificação

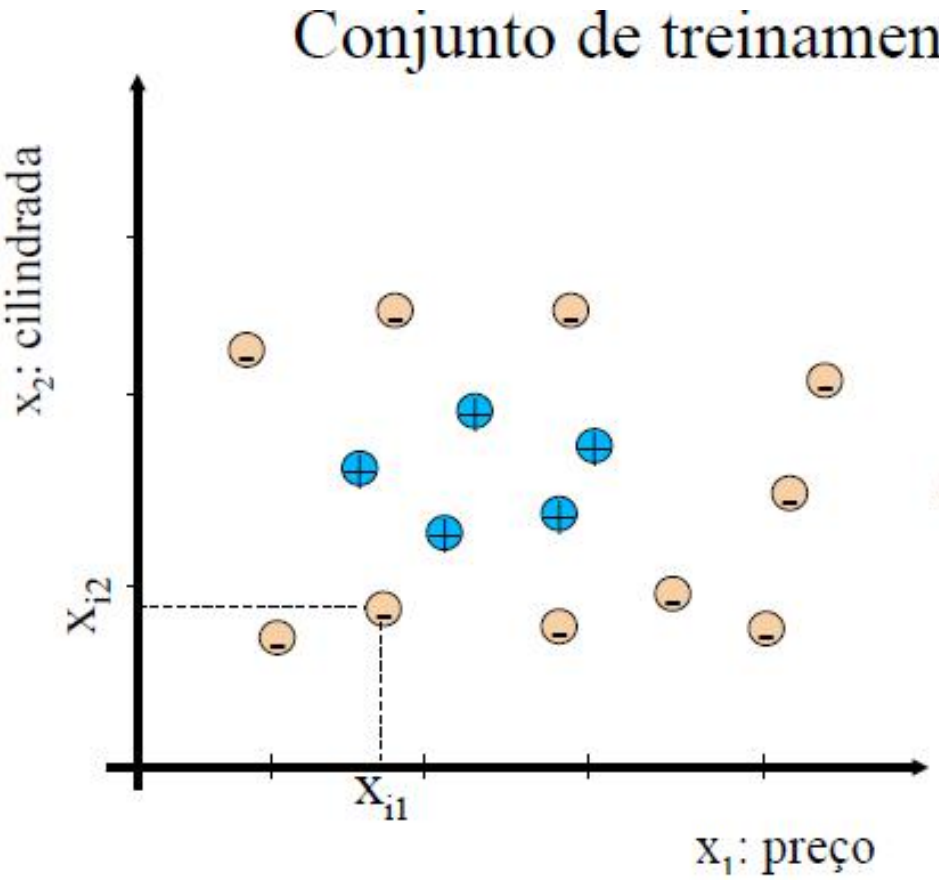
- Supor a tarefa de aprender a classificar carros em duas classes
  - Carro esporte (+)
  - Carro passeio (-)

- Dados de entrada:
  - Características de um carro
  - Preço (x_{i1}) e cilindrada (x_{i2})

Classificação

- Dados exemplos de treinamento, encontrar um modelo
  - Modelagem descritiva
    - O que representa um carro de passeio?
  - Modelagem preditiva
    - Qual a classe de um novo carro?

Classificação



$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$c(X) = \begin{cases} 1 \text{ se } X \text{ for positivo} \\ 0 \text{ se } X \text{ for negativo} \end{cases}$$

$$D = \{X_i, c(X_i)\}_{i=1}^{N}$$

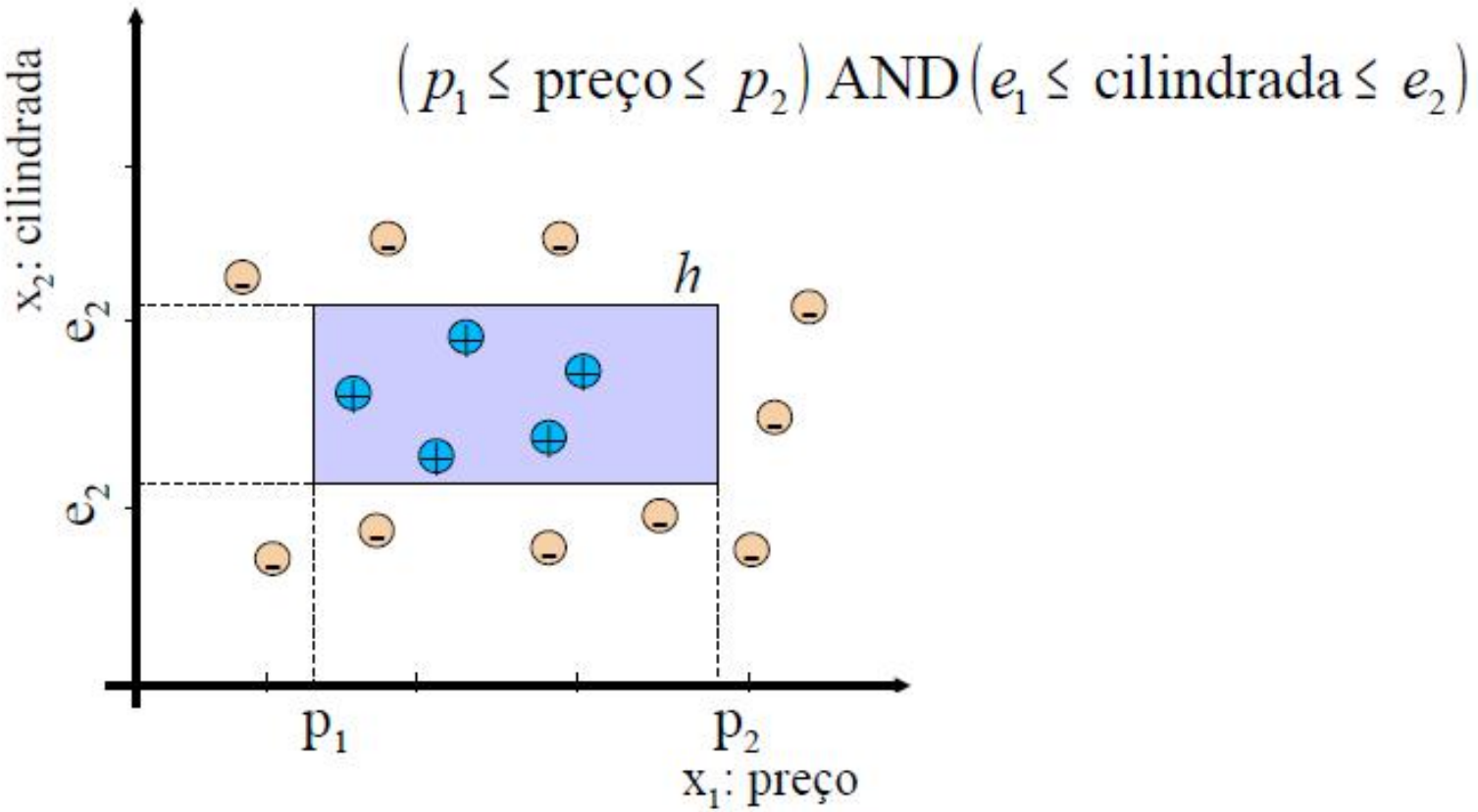
Classificação

- Com os exemplos anteriores, a seguinte hipótese pode ser induzida:

$$\text{Se } (p_1 \leq \text{preço} \leq p_2) \text{ AND } (e_1 \leq \text{cilindrada} \leq e_2)$$
$$\text{Então carro de passeio}$$

  - Para valores adequados de p_1, p_2, e_1 e e_2
  - Assume que pode ser representada por um retângulo no espaço preço X cilindrada

Classificação

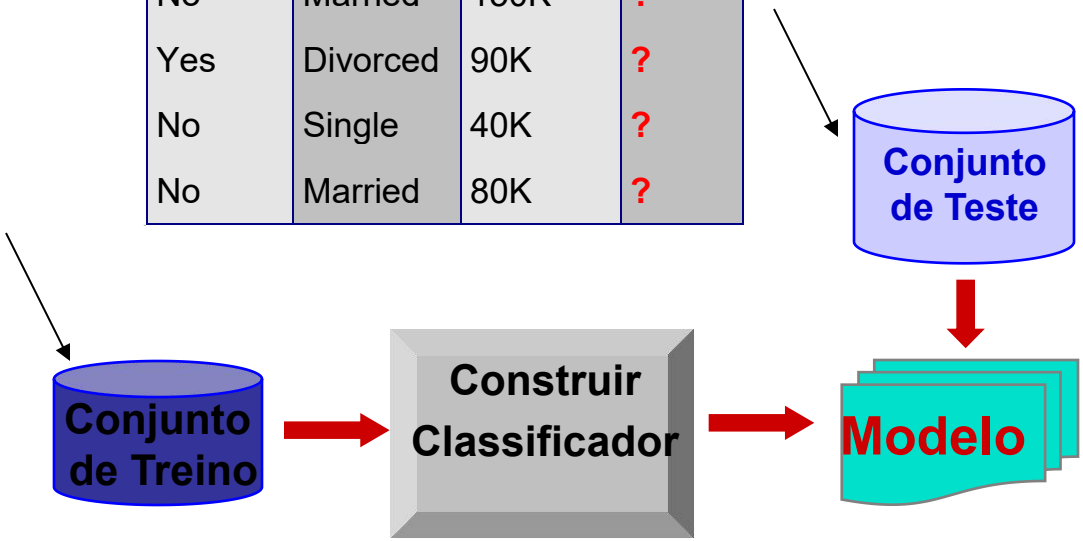


Classificação

classe

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?

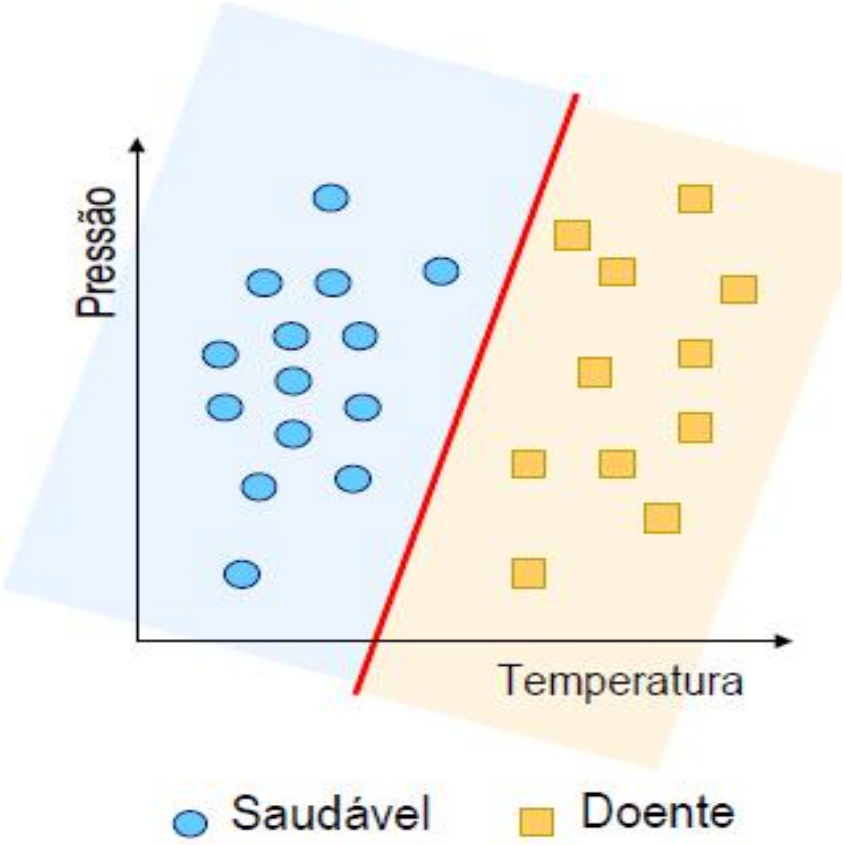


Retirado de: Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Entendo melhor a tarefa de classificação

Classificação Binária

- Mais comum
  - Dados podem pertencer a uma dentre 2 classes
    - Classe positiva
    - Classe negativa



Classificação Avançada

- Classificação com uma única classe
  - Detecção de Novidades
- Classificação Multiclasses
- Classificação Hierárquica
- Classificação Multi-rótulo

Técnicas de classificação

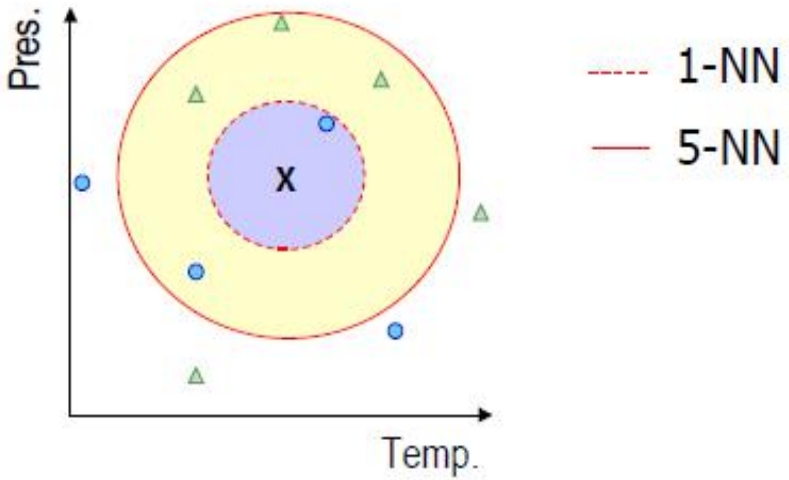
- Exemplos de técnicas de classificação
  - K-vizinhos mais próximos
  - Árvores de Decisão
  - Métodos basedos em regras
  - Redes Neurais
  - *Naive Bayes*
  - Máquinas de Vetores de Suporte (SVM)

K-vizinhos mais próximos (K-NN)

- É um dos algoritmo mais simples de aprendizado de máquina
- Classifica um novo objeto com base nos exemplos próximos a ele
- É um algoritmo preguiçoso (*lazy),* pois não aprende um modelo
- Pode ser usado tanto em classificação quanto regressão

K-vizinhos mais próximos (K-NN)

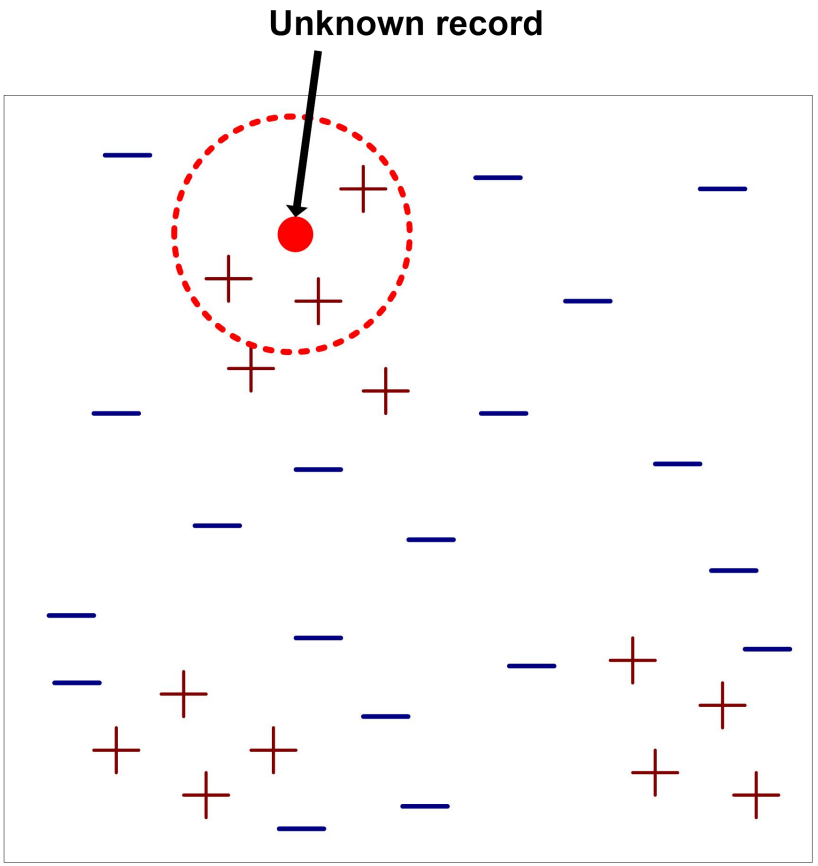
- Aprendizado baseado em instância
  - Classifica de acordo com distância aos vizinhos



Simples

Armazenamento de dados (não há modelo explícito)

K-vizinhos mais próximos (K-NN)



- Requer
  - Um conjunto de instâncias rotuladas
  - Medida de proximidade para calcular a distância/similaridade entre um par de instâncias. Ex: distância Euclidiana
  - O valor de k, o número de vizinhos mais próximos
  - Um método para usar o rótulo das classes dos k vizinhos mais próximos e determinar a classe da nova instância (ex: voto da maioria)

K-vizinhos mais próximos

```
Para cada novo exemplo
Definir a classe dos k exemplos
    mais próximos
Classificar exemplo na classe
    majoritária de seus vizinhos
```

K-vizinhos mais próximos

- Quantos vizinhos?
  - K muito grande
    - Vizinhos podem ser muito diferentes
    - Predição tendenciosa para classe majoritária
    - Custo computacional mais elevado
  - K muito pequeno
    - Não usa informação suficiente
    - Previsão pode ser instável
    - Distâncias podem ser ponderadas

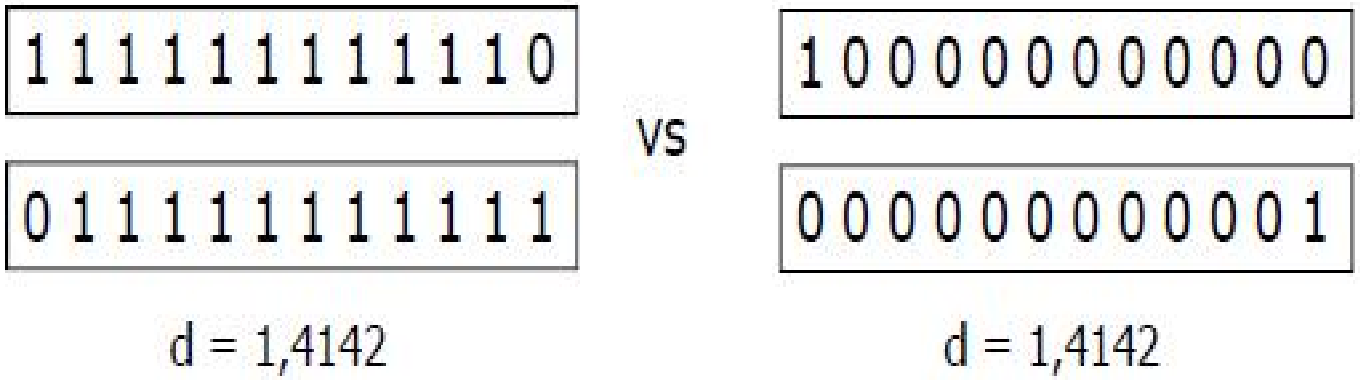
K-vizinhos mais próximos

- Medida de distância
  - KNN tem o desempenho afetado pela medida de distância
    - No caso da Euclidiana, supõe-se atributos numéricos
    - A escala do atributo pode afetar os resultados

Escolha inapropriada da medida de distância

- Exemplo

  Para documentos, cosseno é melhor do que Euclidiana



K-vizinhos mais próximos

- Aspectos positivos
  - O treinamento é simples (armazenar objetos)
  - Constrói aproximações locais da função objetivo diferentes para cada novo dado
  - É aplicável mesmo em problemas complexos
  - Naturalmente incremental

K-vizinhos mais próximos

- Aspectos negativos
  - Por ser *lazy* não obtém uma representação compacta dos dados
    - Não se tem um modelo dos dados
  - Classificar um objeto significa calcular a distância dele a todos os objetos de treinamento
  - É afetado pela presença de atributos redudantes e irrelevantes
    - Como todo algoritmo baseado em distância
  - Problemas em alta dimensionalidade

Entendendo a questão da dimensionalidade

- Considere 100 pontos com distribuição uniforme
  - em um quadrado cujo lado mede 1
  - em um cubo cujo lado mede 1
  - ....
  - calculando a distância média entre pontos temos:

Num. Dimensões	Distância Média
2	0,494
3	0,647
4	0,772
5	0,875
...	...
10	1,280

Aumento da Distância;
Densidade diminui;
Conjunto de Dados esparso

Retirado de: Facelli, Kati et al. Inteligência Artificial: Uma abordagem de Aprendizado de Máquina, LTC, 2011

Árvore de Decisão

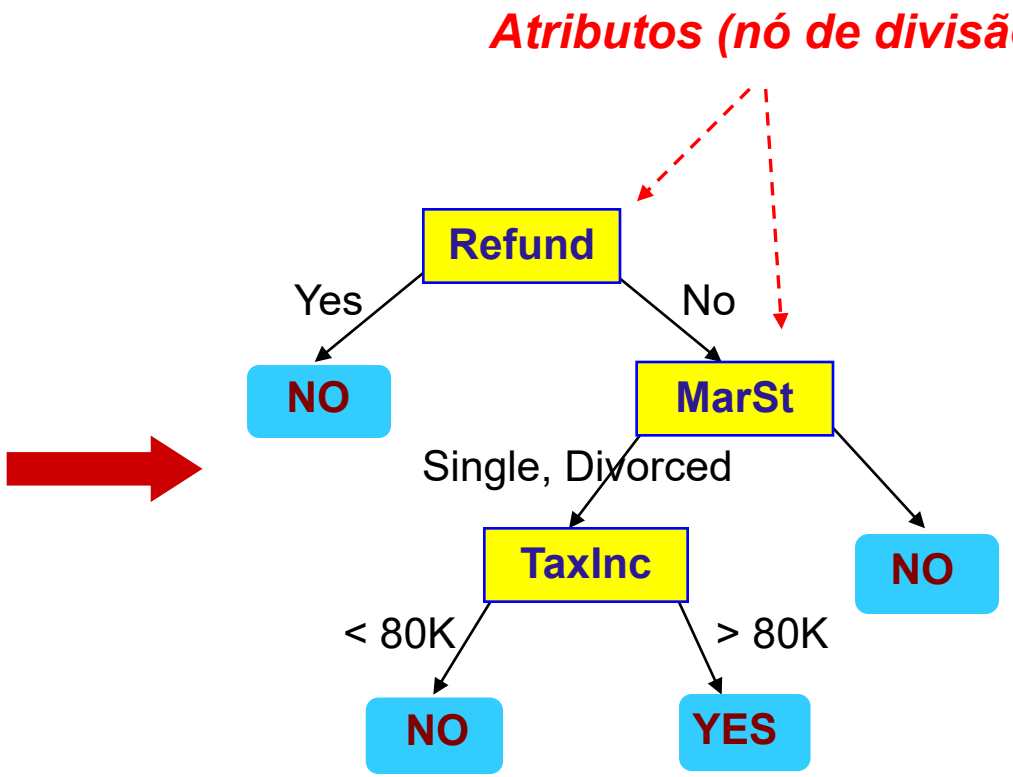
- Utiliza uma estratégia de dividir-para-conquistar
  - Um problema complexo é decomposto em subproblemas mais simples
  - Recursivamente a mesma estratégia é aplicada a cada sub-problema
- A capacidade de discriminação de uma árvore vem da
  - Divisão do espaço definido pelos atributos em subespaços
  - A cada sub-espaço é associada uma classe

Árvore de Decisão

- Representação
  - Cada nó interno testa um atributo (nó de divisão)
    - Contém um teste condicional baseado nos valores do atributo. Ex: idade > 18
  - Cada ramo (aresta) corresponde a um valor do atributo
  - Cada folha representa uma classe
    - Rotulado como uma função
    - Considera-se os valores da variável alvo dos exemplos que chegam no nó folha
    - Ex: função moda
  - Nó raiz: não tem aresta de entrada, só de saída

Árvore de Decisão - Exemplo

	categórico	categórico	contínuo	classe
Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



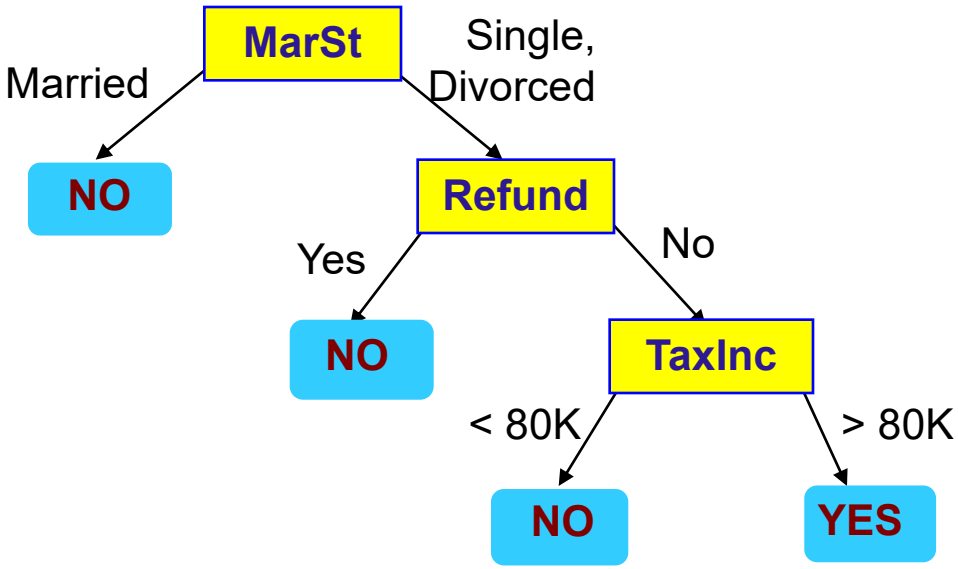
Modelo: Árvore de Decisão

Retirado de: Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Árvore de Decisão - Exemplo

classe

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



Pode existir mais de uma árvore que se adequa aos dados

Retirado de: Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

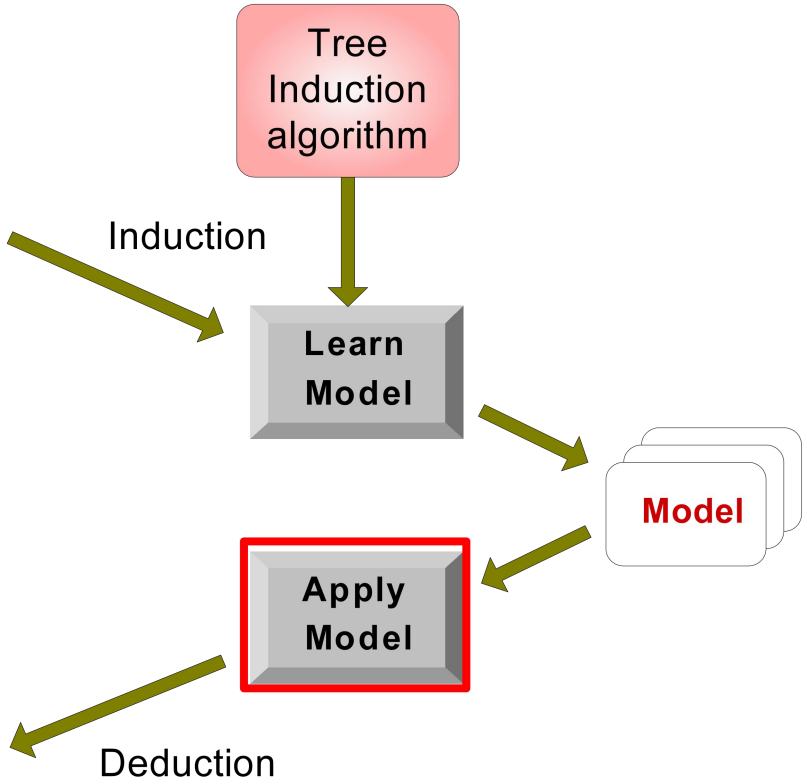
Tarefa de classificação usando árvores de decisão

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

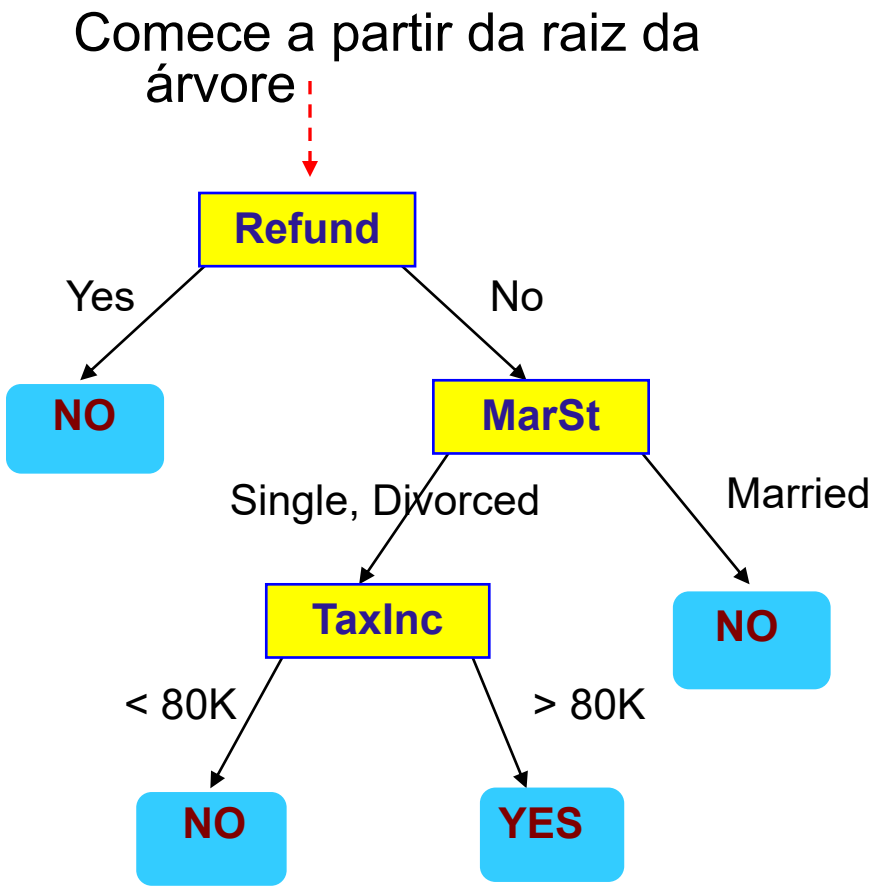
Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Retirado de: Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Aplicação do Modelos aos Dados



Dado de Teste

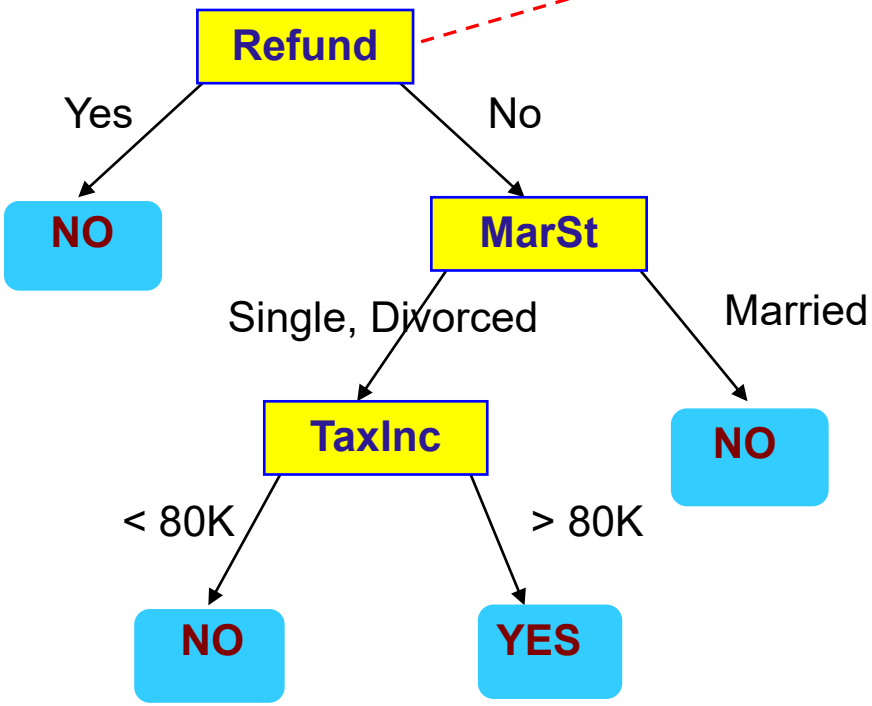
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Retirado de: Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Aplicação do Modelo aos Dados

Dado de Teste

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

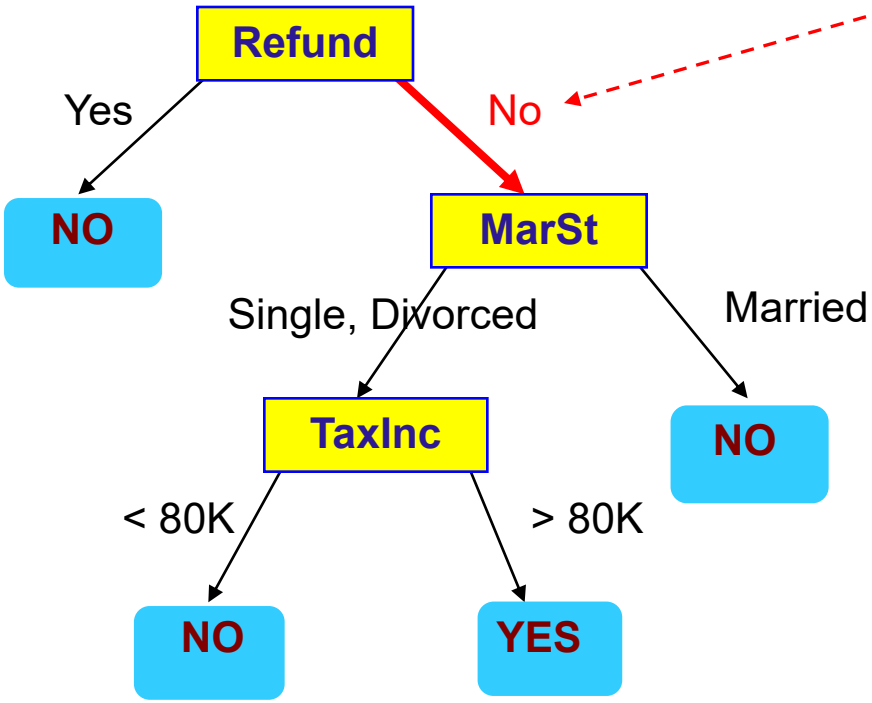


Retirado de: Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Aplicação do Modelo aos Dados

Dado de Teste

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

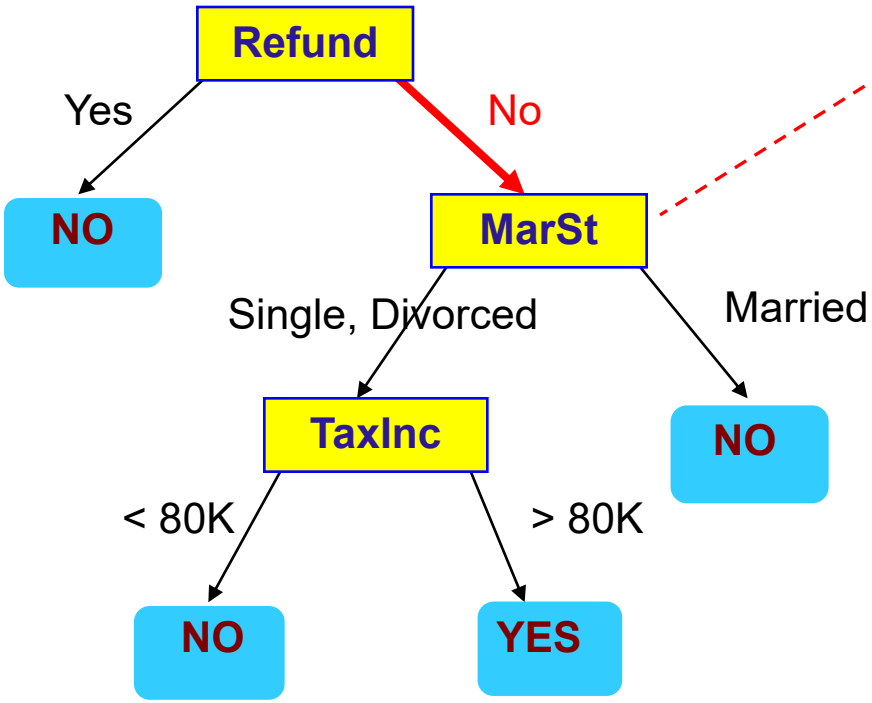


Retirado de: Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Aplicação do Modelo aos Dados

Dado de Teste

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

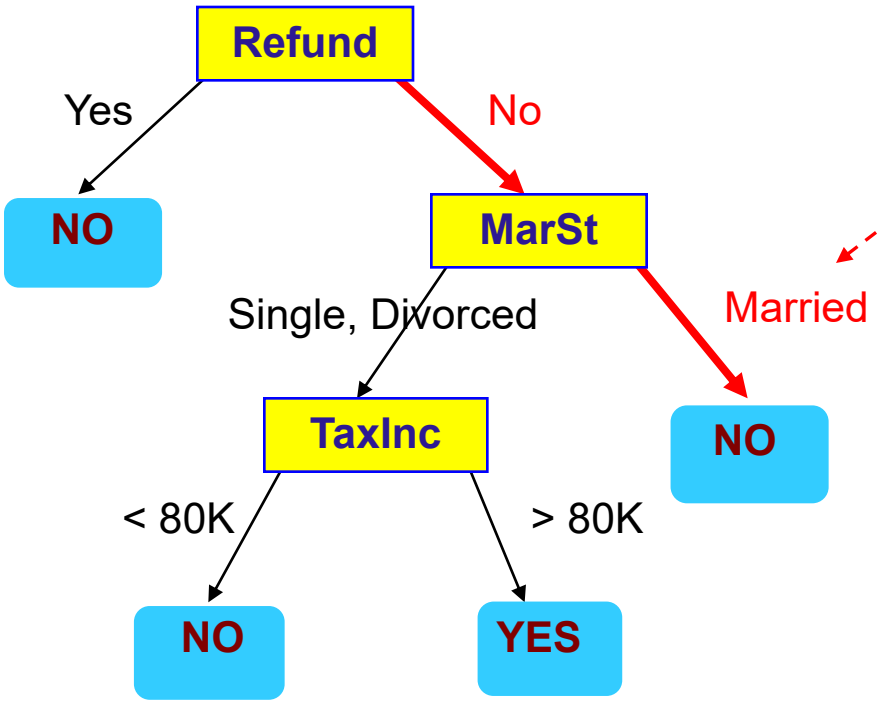


Retirado de: Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Aplicação do Modelo aos Dados

Dado de Teste

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

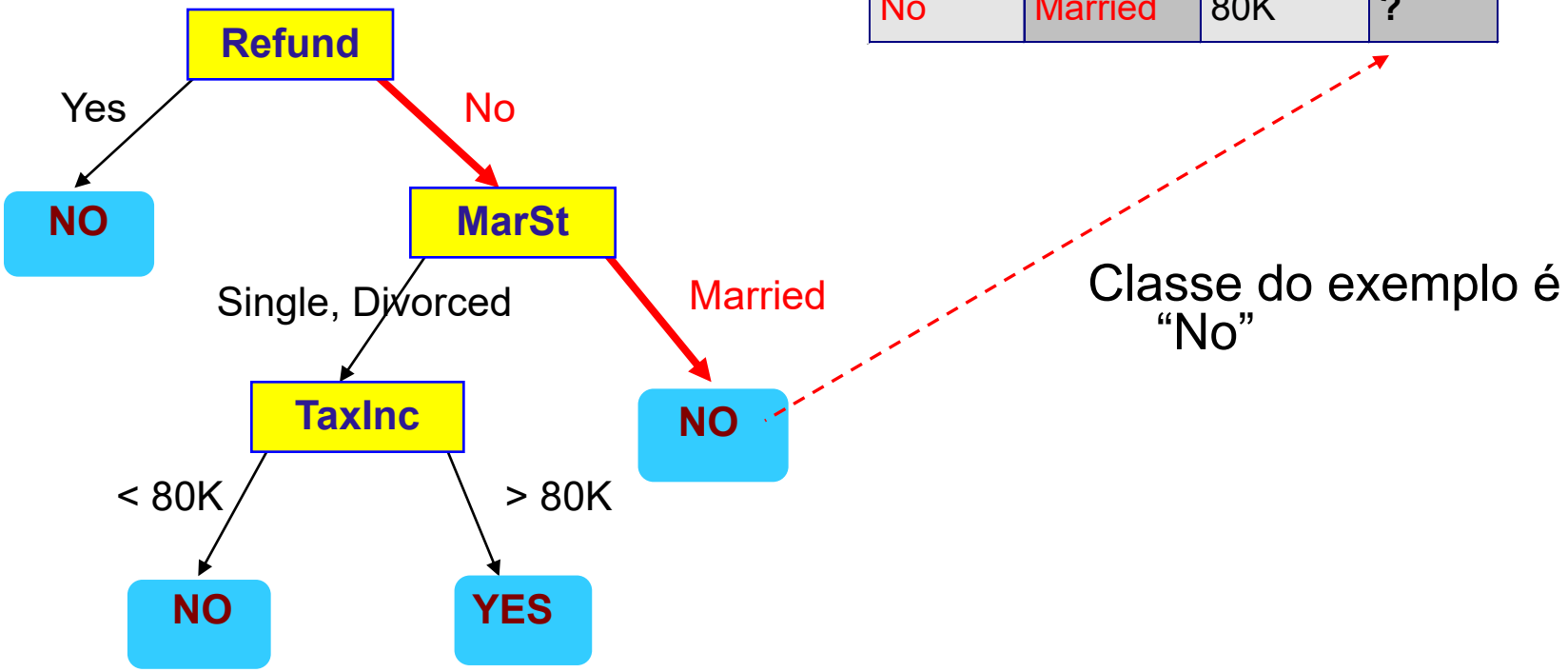


Retirado de: Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Aplicação do Modelo aos Dados

Dado de Teste

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

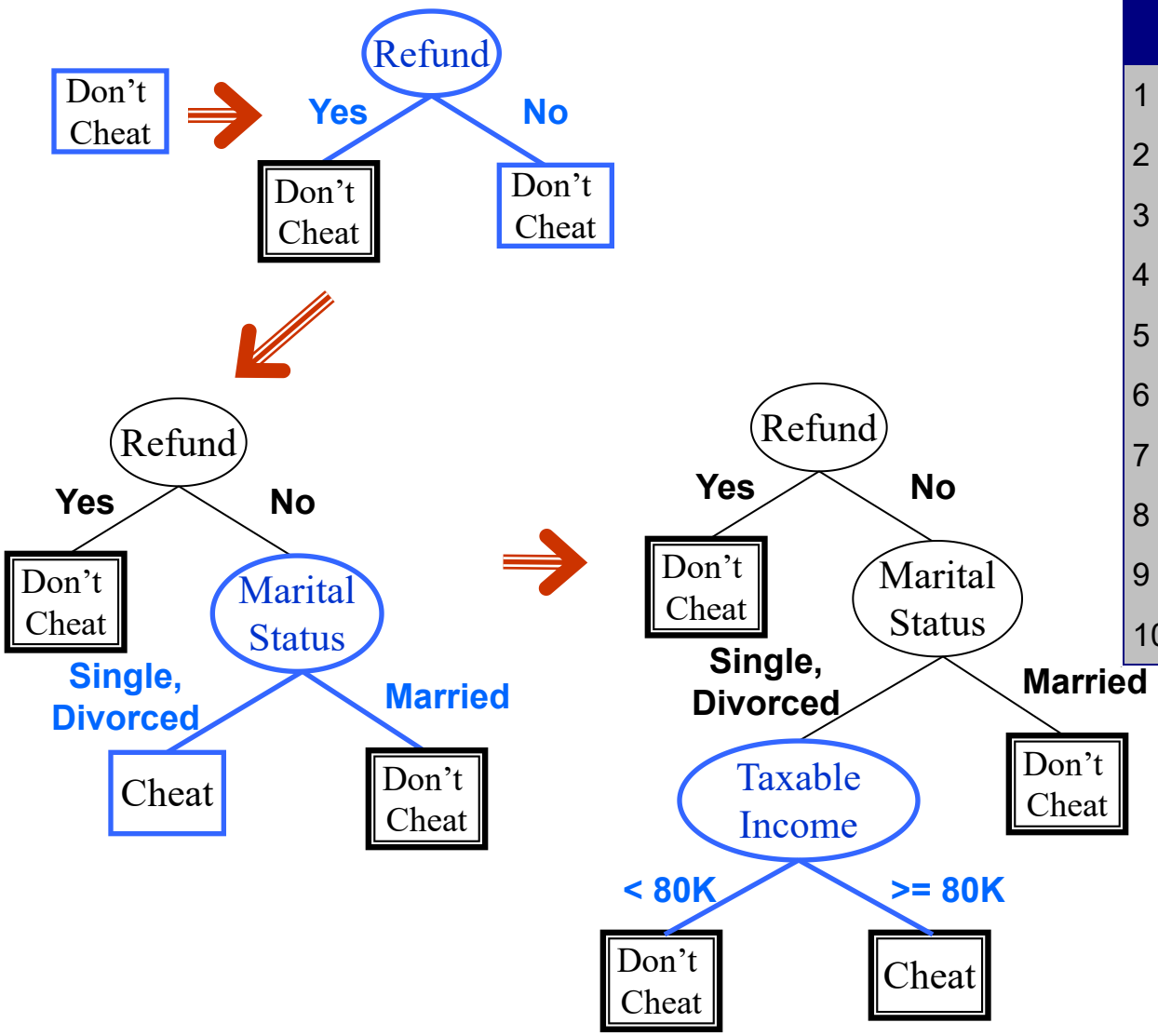


Retirado de: Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Árvore de Decisão

- Idéia base para construção:
  1. Escolher um atributo
  2. Estender a árvore adicionando um ramo para cada valor do atributo
  3. Passar os exemplos para as folhas (tendo em conta o valor do atributo escolhido)
  4. Para cada folha
     1. Se todos os exemplos são da mesma classe, associar essa classe à folha
     2. Senão repetir os passos 1 a 4

Hunt's Algorithm



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Retirado de: Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Algoritmo: Árvore de Decisão

Entrada: Um conjunto de treinamento D

Saída: Árvore de Decisão

/* **Função GeraÁrvore(D)** */

se *criterio_parada*(**D**) = *Verdadeiro* **então**

    Retorna: um nó folha rotulado com a constante que minimiza a função perda;

fim

Escolha o atributo que maximiza o criterio de divisão em **D;**

para cada *partição dos exemplos* $\mathbf{D_i}$ baseado nos valores do atributo escolhido **faça**

    Induz uma subárvore $Árvore_i$ = *GeraÁrvore*($\mathbf{D_i}$)

fim

Retorna: Árvore contendo um nó de decisão baseado no atributo escolhido e descendentes $Árvore_i$

Retirado de: Facelli, Kati et al. Inteligência Artificial: Uma abordagem de Aprendizado de Máquina, LTC, 2011

Indução da árvore

- Estratégia Gulosa
  - Dividir os registros baseando-se em um atributo teste que otimiza um certo critério
- Questões
  - Determinar como dividir os registros
    - Como escolher o atributo de teste
    - Como determinar a melhor divisão?
  - Determinar quando parar a divisão

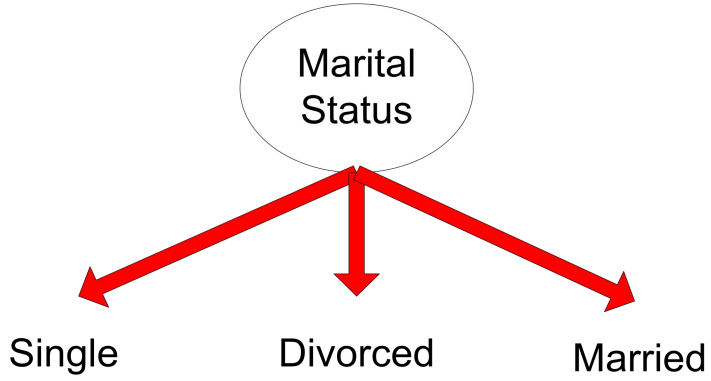
Especificando a condição de teste

- Depende do tipo do atributo
  - Nominal
  - Ordinal
  - Contínuo

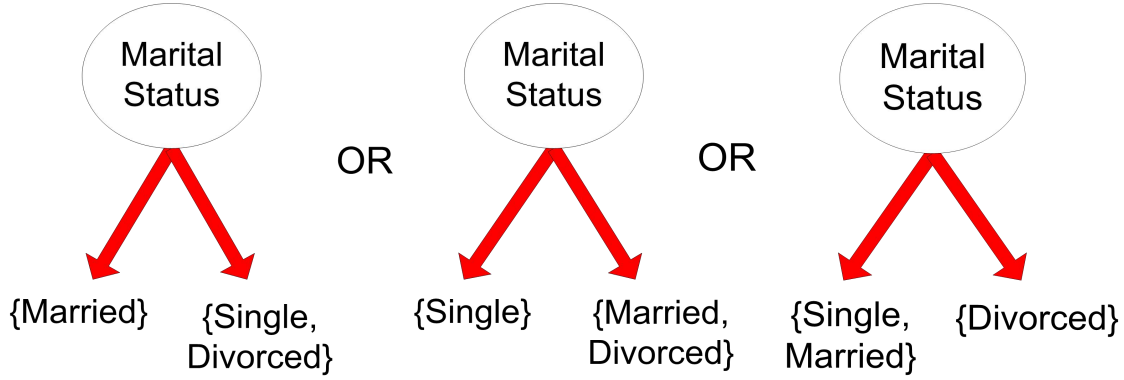
- Depende do número de modos de dividir
  - 2 divisões
  - Mais que 2 divisões

Especificando a condição de teste

- Divisão com múltiplas opções

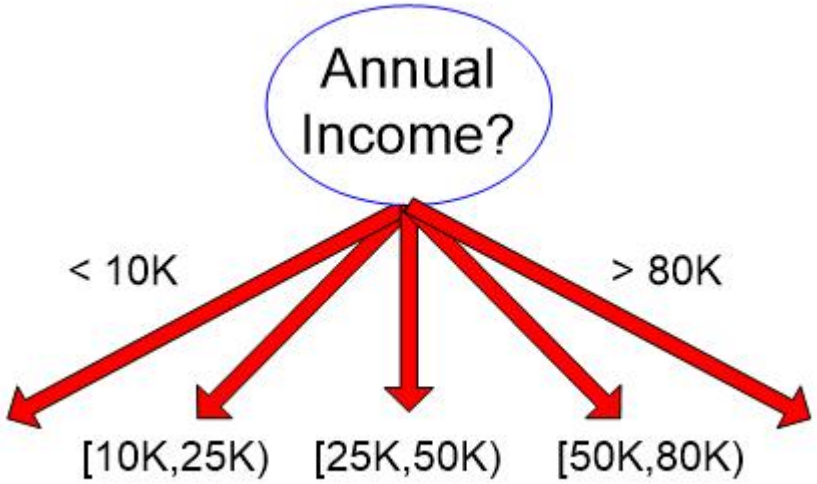


- Divisão binária



Retirado de: Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Especificando a condição de teste



Retirado de: Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Determinando a melhor divisão

- Guiada por uma medida de "*goodness of split*"
  - Indica quão bem um atributo discrimina as classes
  - Usada para selecionar o atributo que maximiza essa medida
  - Para cada teste possível, o sistema considera o subconjunto dos dados obtidos

Determinando a melhor divisão

- Proposta gulosa:
  - Nós com distribuição de classe homogênea são prefereidos
- Medida de impureza de um nó:

C0: 5 C1: 5	C0: 9 C1: 1
Não-homogêneo, **Alto nível de impureza**	**Homogêneo,** **Baixo nível de impureza**

Retirado de: Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

Medidas de Impureza de um nó

- Índice de Gini

$$Gini\ Index = 1 - \sum_{i=0}^{c-1} p_i(t)^2$$

Where $\boldsymbol{p_i(t)}$ is the frequency of class $\boldsymbol{i}$ at node t, and $\boldsymbol{c}$ is the total number of classes

- Entropia

$$Entropy = -\sum_{i=0}^{c-1} p_i(t) log_2 p_i(t)$$

- Erro de classificação

$$Classification\ error = 1 - \max[p_i(t)]$$

Retirado de: Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006

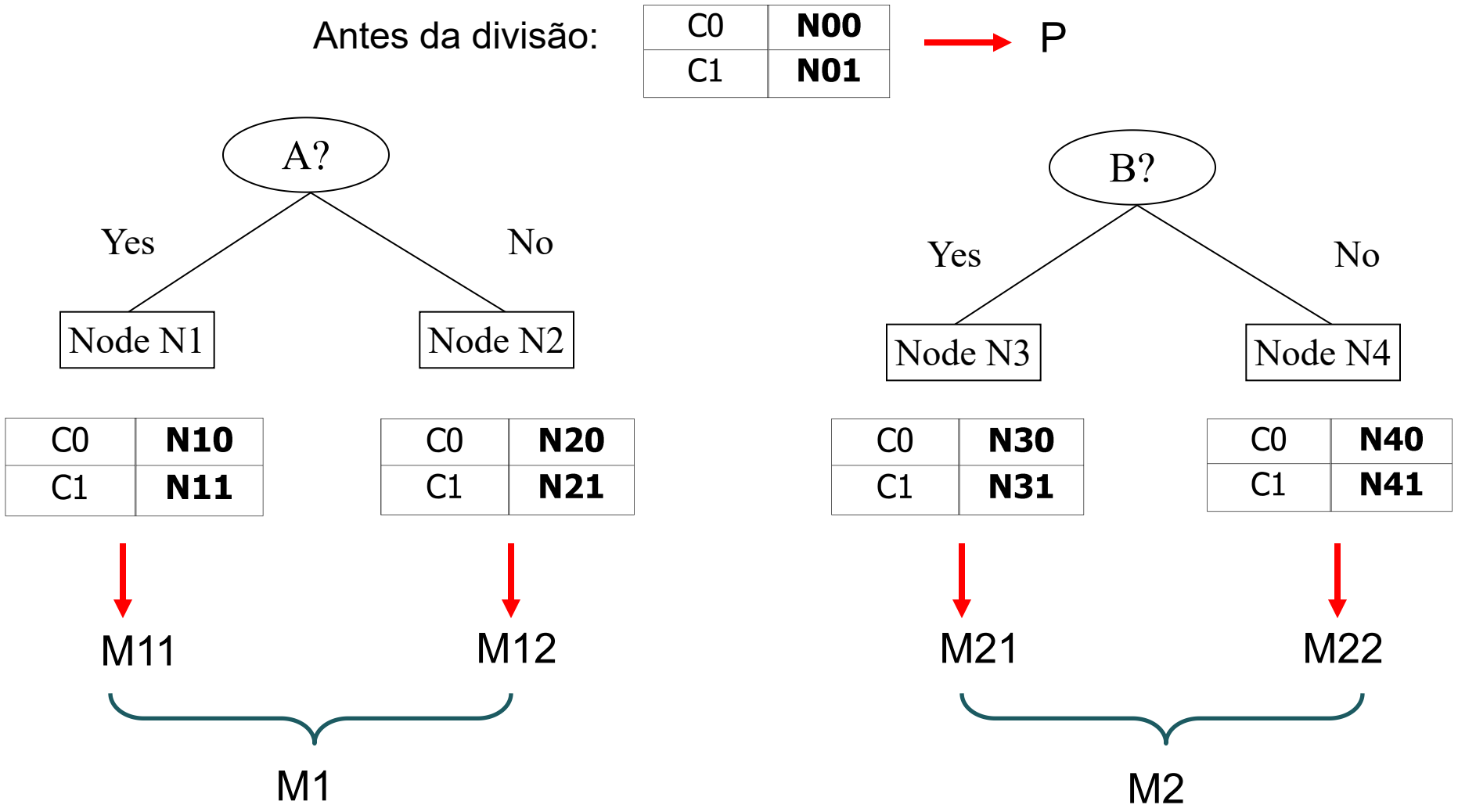
Determinando a melhor divisão

1. Calcule a medida de impureza (P) antes da divisão
2. Calcule a medida de impureza (M) depois da divisão
   - Calcule a medida de impureza de cada nó filho
   - M é a impureza ponderada dos nós filhos
3. Escolha o atributo que produz o maior ganho

Ganho = P - M

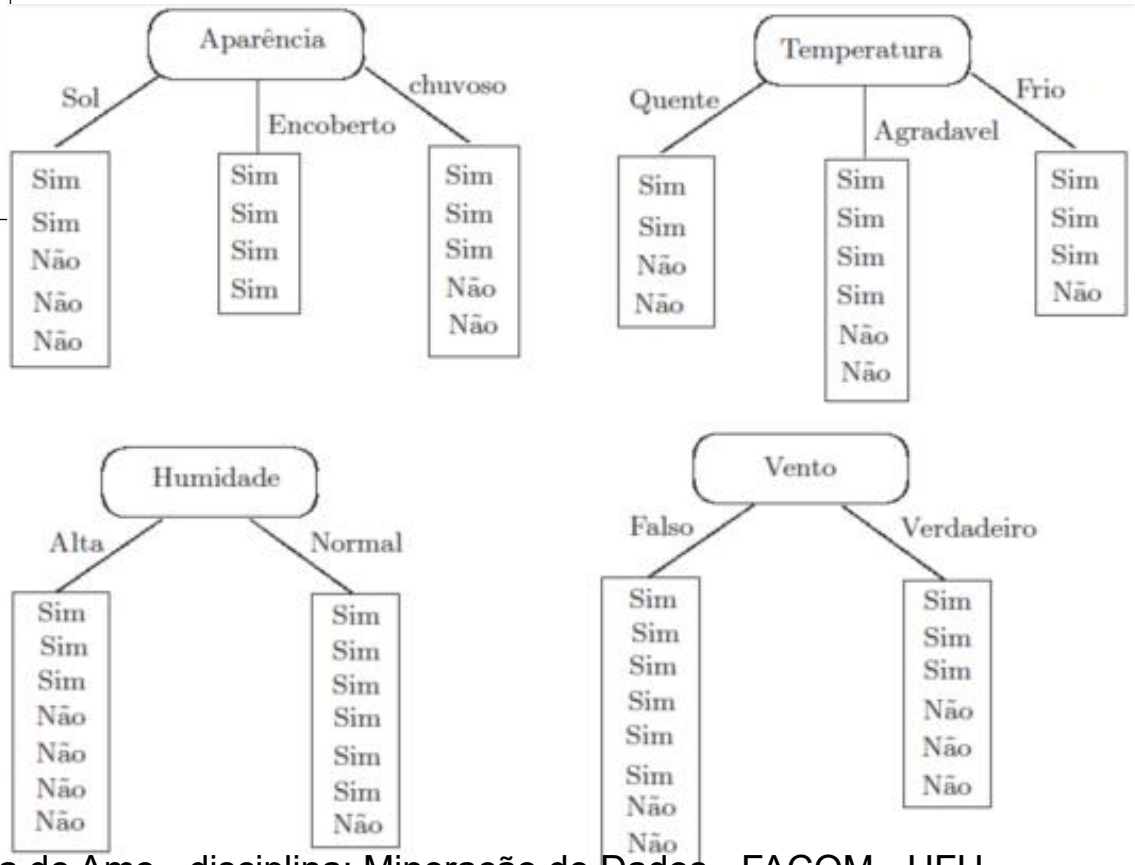
ou equivalentemente, a menor medida de impureza após a divisão (M)

Árvores de Decisão



Árvore de Decisão - Exemplo

Aparência	Temperatura	Humidade	Vento	Jogo
Sol	Quente	Alta	Falso	Não
Sol	Quente	Alta	Verdade	Não
Encoberto	Quente	Alta	Falso	Sim
Chuvoso	Agradável	Alta	Falso	Sim
Chuvoso	Frio	Normal	Falso	Sim
Chuvoso	Frio	Normal	Verdade	Não
Encoberto	Frio	Normal	Verdade	Sim
Sol	Agradável	Alta	Falso	Não
Sol	Frio	Normal	Falso	Sim
Chuvoso	Agradável	Normal	Falso	Sim
Sol	Agradável	Normal	Verdade	Sim
Encoberto	Agradável	Alta	Verdade	Sim
Encoberto	Quente	Normal	Falso	Sim
Chuvoso	Agradável	Alta	Verdade	Não



Árvore de Decisão - Exemplo

Se escolhemos o atributo Aparência:

$$\text{Info(Nó)} = \frac{5}{14}\,\text{entropia(Folha 1)} + \frac{4}{14}\,\text{entropia(Folha 2)} + \frac{5}{14}\,\text{entropia(Folha3)}$$

$$\text{entropia(Folha 1)} = \frac{2}{5}\,log_2\frac{2}{5} + \frac{3}{5}\,log_2\frac{3}{5} = 0.971$$

$$\text{entropia(Folha 2)} = \frac{4}{4}\,log_2\frac{5}{5} + \frac{0}{4}\,log_2\frac{0}{4} = 0$$

$$\text{entropia(Folha 3)} = \frac{3}{5}\,log_2\frac{3}{5} + \frac{2}{5}\,log_2\frac{2}{5} = 0.971$$

$$\text{Logo, Info(Nó)} = \frac{5}{14}\,0.971 + \frac{4}{14}\,0 + \frac{5}{14}\,0.971 = 0.693$$

Árvore de Decisão - Exemplo

- Se escolhemos o atributo Temperatura:

  $\text{Info(Nó)} = \frac{4}{14}\ \text{entropia(Folha 1)} + \frac{6}{14}\ \text{entropia(Folha 2)} + \frac{4}{14}\ \text{entropia(Folha3)} = 0.911$
- Se escolhemos o atributo Humidade:

  $\text{Info(Nó)} = \frac{7}{14}\ \text{entropia(Folha 1)} + \frac{7}{14}\ \text{entropia(Folha 2)} = 0.788$

  item Se escolhemos o atributo Humidade:

  $\text{Info(Nó)} = \frac{8}{14}\ \text{entropia(Folha 1)} + \frac{6}{14}\ \text{entropia(Folha 2)} = 0.892$

Árvore de Decisão - Exemplo

$$\text{Info-pré} = \frac{9}{14} log_2 \frac{9}{14} + \frac{5}{14} log_2 \frac{5}{14} = 0.940.$$

$$\text{ganho(Aparência)} = 0.940 - 0.693 = 0.247$$
$$\text{ganho(Temperatura)} = 0.940 - 0.911 = 0.029$$
$$\text{ganho(Humidade)} = 0.940 - 0.788 = 0.152$$
$$\text{ganho(Vento)} = 0.940 - 0.892 = 0.020$$

Aparência deve ser o atributo a ser usado neste caso!

Determinando a melhor divisão

- Como fazer a divisão para atributos contínuos?
  - Ex: Temperatura - nro real
- A estratégia usual pesquisa por uma partição binária do conjunto de treinamento
  - Conjunto dos exemplos em que o atributo é <= valor
  - Conjunto dos exemplos em que o atributo é >= valor
  - Ex: Temperatura = 70,5

Como obter o valor (ponto de corte)?

Determinando a melhor divisão

- Como fazer a divisão para atributos contínuos?
  - Ordenar os valores do atributo contínuo
  - O ponto médio entre dois valores consecutivos é um possível ponto de corte e é avaliado pela função mérito
  - O possível ponto de corte que maximiza a função mérito é escolhido

Poda de árvores

- É um importante passo da construção de árvores
  - Principalmente em domínios ruidosos
- Dados ruidosos levam a dois problemas
  - Estatísticas calculadas nos nós mais profundos da árvore têm baixos níveis de importância devido ao pequeno nro de exemplos que chegam nesse nós
    - Superajustamento ao treinamento
  - A árvore induzida tente a ser grande e difícil de compreender
- Podar uma árvore
  - Trocar nós profundos por folhas

Poda de árvores

- Podar quase sempre causa classificação incorreta de alguns exemplos do conjunto de treinamento
- A vantagem da poda aparece ao se classificar novos exemplos, não usados no treinamento
  - Erros de generalização menores
- Métodos de poda
  - Pré-poda: param a construção quando algum critério é satisfeito
  - Pós-poda: constrõem a árvore e podam posteriormente

Árvores de Decisão - Problemas

- O conjunto de treinamento pode não possuir valores para alguns atributos
  - Árvores de decisão podem ser usadas mesmo na presença de valores desconhecidos
- Estratégias
  - Trocar o valor não conhecido pelo mais comum para o atributo
  - Considerar o valor desconhecido com um outro valor possível do atributo (um ramo da árvore para o desconhecido)
  - Estratégias mais complexas usadas no C4.5 e CART

Árvores de Decisão

- Vantagens
  - Não assumem nenhuma distribuição para os dados
  - Relativamente baratas de construir
    - Complexidade linear no nro de instâncias
  - Extremamente rápido na classificação de registros desconhecidos
  - Fácil de interpretar para árvores de pequeno porte
  - Robusto para ruído
  - Pode lidar facilmente com atributos redundantes
  - Pode lidar facilmente com atributos irrelevantes

Árvores de Decisão

- Desvantagens
  - Cada limite de decisão envolve apenas um único atributo
  - Na presença de valores ausentes, os algoritmos devem empregar algum mecanismo para lidar com eles
  - A presença de atributos contínuos é uma dificuldade
  - Instabilidade
    - Pequenas variações no conjunto de treinamento podem produzir grandes variações na árvore.

Referências

- Katti, F.; Lorena, A. C.; Gama, J.; Carvalho, A. C. P. L. F. Inteligência Artificial: Uma abordagem de Aprendizado de Máquina, LTC, 2011
- Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006
- Material do curso de Mineração de Dados da profa. Sandra de Amo - FACOM-UFU