Aprendizado de Máquina

Árvores de Decisão

Prof. Dr. André C. P. L. F. de Carvalho
Posdoutorando: Isvani Frias-Blanco
ICMC-USP

Principais tópicos

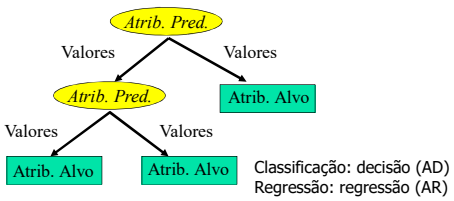
- Árvores de decisão
- Algoritmo de Hunt
- Medidas para escolha de atributos
- Ponto de referência
- Critério de parada
- Espaço de hipóteses

Introdução

- Explicação das decisões pode ser importante para algumas aplicações
  - RNAs, SVMs e RNPs são caixas pretas
- Modelos interpretáveis são gerados por algumas algoritmos de AM
  - Árvores de características (decisão)
  - Conjunto de regras
  - Redes Bayesianas

Árvores de características

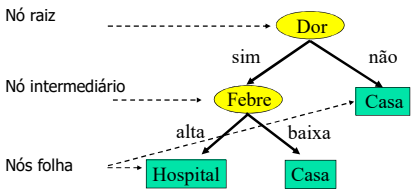
- Alguns algoritmos de AM particionam características (atributos) de forma hierárquica



Classificação: decisão (AD)
Regressão: regressão (AR)

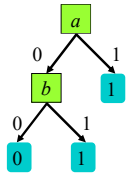
Algoritmo de indução de AD

- Induzem modelos representados por ADs



Outro exemplo simples

a	b	a v b
0	0	0
0	1	1
1	0	1
1	1	1



Nós internos e raiz: atributos preditivos
Nós externos (folhas): atributo alvo

Exercício

- Encontrar árvore de decisão para:
  - a AND b
  - a XOR b
  - (a AND b) OR (b AND $\overline{c}$)

Algoritmo de indução de AD

- Existem vários, entre eles:
  - Algoritmo de Hunt
    - Um dos primeiros
    - Base de vários algoritmos atuais
  - CART
  - ID3
  - C4.5
  - VFDT

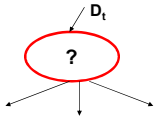
Algoritmo de Hunt

- Seja X_t o conjunto de objetos de treinamento que atingem o nó t

Se todos os objetos de $X_t \in$ a mesma classe y
Então O nó t é um nó folha rotulado pela classe y
Senão Selecionar um atributo preditivo teste para dividir X_t
Dividir X_t em subconjuntos usando valores desse atributo
Aplicar algoritmo a cada subconjunto gerado

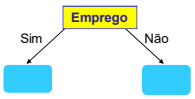
Algoritmo de Hunt

Emprego	Estado	Renda	Classe
Sim	Solteiro	9500	Não
Não	Casado	8000	Não
Não	Solteiro	7000	Não
Sim	Casado	12000	Sim
Não	Divorciado	9000	Sim
Não	Casado	6000	Não
Sim	Divorciado	4000	Não
Não	Solteiro	8500	Sim
Não	Casado	7500	Não
Não	Solteiro	9000	Sim



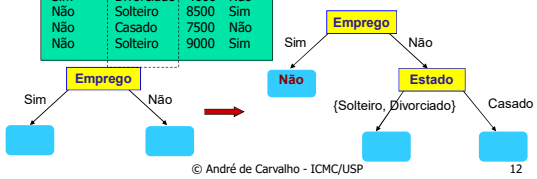
Algoritmo de Hunt

Emprego	Estado	Renda	Classe
Sim	Solteiro	9500	Não
Não	Casado	8000	Não
Não	Solteiro	7000	Não
Sim	Casado	12000	Não
Não	Divorciado	9000	Sim
Não	Casado	6000	Não
Sim	Divorciado	4000	Não
Não	Solteiro	8500	Sim
Não	Casado	7500	Não
Não	Solteiro	9000	Sim



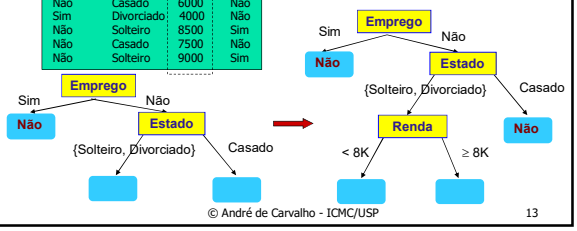
Algoritmo de Hunt

Emprego	Estado	Renda	Classe
Sim	Solteiro	9500	Não
Não	Casado	8000	Não
Não	Solteiro	7000	Não
Sim	Casado	12000	Não
Não	Divorciado	9000	Sim
Não	Casado	6000	Não
Sim	Divorciado	4000	Não
Não	Solteiro	8500	Sim
Não	Casado	7500	Não
Não	Solteiro	9000	Sim



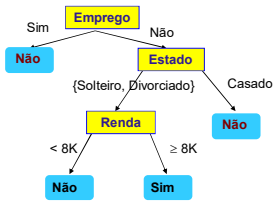
Algoritmo de Hunt

Emprego	Estado	Renda	Classe
Sim	Solteiro	9500	Não
Não	Casado	8000	Não
Não	Solteiro	7000	Não
Sim	Casado	12000	Não
Não	Divorciado	9000	Sim
Não	Casado	6000	Não
Sim	Divorciado	4000	Não
Não	Solteiro	8500	Sim
Não	Casado	7500	Não
Não	Solteiro	9000	Sim



Algoritmo de Hunt

Emprego	Estado	Renda	Classe
Sim	Solteiro	9500	Não
Não	Casado	8000	Não
Não	Solteiro	7000	Não
Sim	Casado	12000	Não
Não	Divorciado	9000	Sim
Não	Casado	6000	Não
Sim	Divorciado	4000	Não
Não	Solteiro	8500	Sim
Não	Casado	7500	Não
Não	Solteiro	9000	Sim



Indução de ADs

- Geralmente usa estratégia gulosa de divisão e conquista
  - Divide progressivamente objetos baseado em um atributo preditivo de teste
    - Escolhido para otimizar algum critério
- Decisões importantes
  - Escolha do atributo preditivo
  - Como dividir os objetos entre os ramos
  - Quando parar de dividir os objetos

Escolha do atributo preditivo

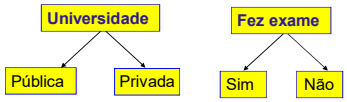
- Atributo preditivo que melhor particiona conjunto atual de objetos
  - Para um mesmo atributo preditivo, diferentes partições podem ser geradas
  - Necessário escolher:
    - Atributo preditivo mais discriminativo
    - Melhor partição para esse atributo
      - Como dividir os objetos

Como dividir os objetos

- Depende do tipo do atributo preditivo e do número de divisões a serem geradas
  - Atributo preditivo binário
    - Divisão binária
      - Árvore binária
  - Atributo preditivo n-ário
    - Divisão binária
    - Divisão n-ária (n > 2)

Atributo binário

- Teste mais simples que existe
  - Define dois possíveis resultados (filhos)



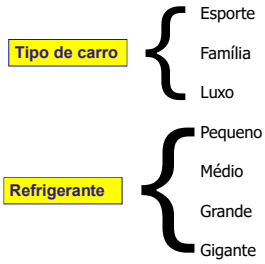
Atributo n-ário

- Divisão
  - Binária
  - N-ária
- Depende do tipo do atributo
  - Simbólico
    - Nominal
    - Ordinal
  - Numérico
    - Discreto
    - Contínuo

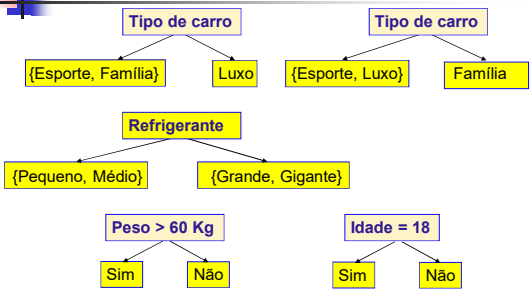
Divisão binária para atributo n-ário

- Único teste com 2 possíveis resultados (filhos)
- Condição de teste é uma comparação
  - Ex.: A < valor, A = valor, A ∈ {valores}, ...
  - Escolher valor(es) que gera(m) melhor partição
    - Ponto de referência
  - Tipo simbólico: agrupar parte dos valores em cada ramo
    - Ordinais: valores agrupados não devem violar relação de ordem
    - Nominais: grupos devem fazer sentido

Atributos simbólicos



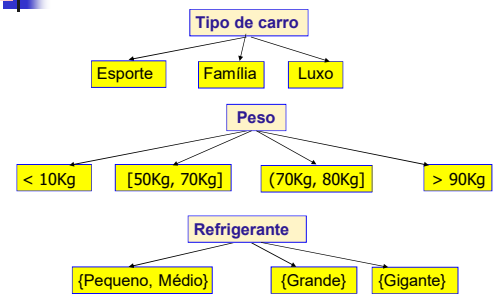
Divisão binária para atributo n-ário



Divisão n-ária para atributo n-ário

- Atributos simbólicos
  - Duas alternativas para definir número de resultados do teste
    - Fazer #ramos = #possíveis valores
    - Agrupar parte dos valores em cada ramo
      - Ordinais
      - Nominais

Divisão n-ária para atributo n-ário

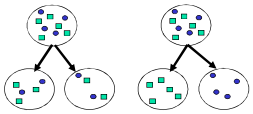


Divisão n-ária para atributo n-ário

- Condição de teste formada por 2 ou mais comparações
  - Cada comparação pode ter 1 ou mais operadores relacionais
    - Um operador
      - Ex. A < valor, A = valor
    - Mais de um operador
      - Ex.: $valor_{inf} < A < valor_{sup}$
  - Escolher valores (pontos de referência)

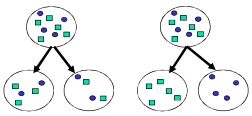
Medidas para escolha de atributo

- Selecionam atributo que melhor discrimina os objetos atuais
  - Buscam partições mais puras após divisão
    - Quanto mais homogêneas as partições, mais puras
  - Medidas de impureza



Medidas de impureza

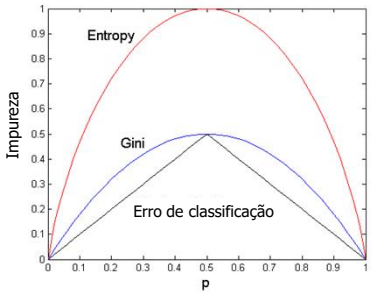
- Baseadas no grau de impureza dos nós filhos
  - Quando maior, pior
- Diferentes medidas geram diferentes partições
- Exemplos
  - Entropia
  - Gini
  - Erro de classificação
  - Qui-quadrado



Propriedade de uma medida de impureza

- Seja uma tarefa de classificação binária
  - Deve ser dependente apenas da magnitude relativa das classes
    - Não muda se a magnitude das duas classes for multiplicada por uma constante
    - Proporção dos exemplos em cada classe
  - Não deve mudar se as classes positiva e negativa são trocadas
  - Deve ser 0, se a proporção de exemplos de uma das classes for 1 (da outra será 0)

Comparação: duas classes



Medidas de impureza

$$Entropia(v) = -\sum_{i=1}^{C} p(i/v)\log_2 p(i/v)$$

$$Gini(v) = 1 - \sum_{i=1}^{C} [p(i/v)]^2$$

$$ErroClass(v) = 1 - \max_i [p(i/v)]$$

Onde:
P(i/v) = fração de dados pertencente a classe i em um nó v
C = número de classes
Considera-se que $0\log_2 0 = 0$

Exemplo

- Calcular a medida de impureza Gini para os dados abaixo:

$$Gini(v) = 1 - \sum_{i=1}^{C} [p(i/v)]^2$$

C1	0
C2	6
Gini=?	

C1	1
C2	5
Gini=?	

C1	2
C2	4
Gini=?	

C1	3
C2	3
Gini=?	

Exemplo

$$Gini(v) = 1 - \sum_{i=1}^{C} [p(i/v)]^2$$

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1
Gini = 1 – P(C1)² – P(C2)² = 1 – 0 – 1 = 0
P(C1) = 1/6 P(C2) = 5/6
Gini = 1 – (1/6)² – (5/6)² = 0.278
P(C1) = 2/6 P(C2) = 4/6
Gini = 1 – (2/6)² – (4/6)² = 0.444
P(C1) = 3/6 P(C2) = 3/6
Gini = 1 – (3/6)² – (3/6)² = 0.500

C1	0
C2	6
Gini=0.000	

C1	1
C2	5
Gini=0.278	

C1	2
C2	4
Gini=0.444	

C1	3
C2	3
Gini=0.500	

Exercício

- Fazer os mesmos cálculos para as medidas de entropia e de erro de classificação

$$Entropia(v) = -\sum_{i=1}^{C} p(i/v) \log_2 p(i/v)$$

$$ErroClass(v) = 1 - \max_i [p(i/v)]$$

C1	0
C2	6
E=?	

C1	1
C2	5
E=?	

C1	2
C2	4
E=?	

C1	3
C2	3
E=?	

C1	0
C2	6
Class=?	

C1	1
C2	5
Class=?	

C1	2
C2	4
Class=?	

C1	3
C2	3
Class=?	

Medida Gini média ponderada

- Usada pelos algoritmos CART, SLIQ, SPRINT
- Quando um nó pai possui *k* filhos, a impureza da divisão é definida por:

$$Gini_{divisão} = \sum_{f=1}^{k} \frac{N(v_f)}{N(v_p)} Gini(v_f)$$

Média ponderada

Onde:

$N(v_t)$: número de objetos no filho (v_f)
$N(v_p)$: número de objetos no nó pai (v_p)

Divisão de atributos binários

	Pai
C1	6
C2	6
Gini = 0.500	

Atrib. A? — Sim → Nó N_1, Não → Nó N_2

	Nó 1	Nó 2
C1	4	2
C2	3	3
Gini =		

$Gini_{divisão}$ =

Atrib. B? — Sim → Nó N_3, Não → Nó N_4

	Nó 3	Nó 4
C1	1	5
C2	4	2
Gini =		

$Gini_{divisão}$ =

Divisão de atributos binários

	Pai
C1	6
C2	6
Gini = 0.500	

Atrib. A? — Sim → Nó N_1, Não → Nó N_2

	Nó 1	Nó 2
C1	4	2
C2	3	3
Gini = 0.486		

$Gini_{divisão}$ = (7/12)x0.49 + (5/12)x0.48 = 0.486

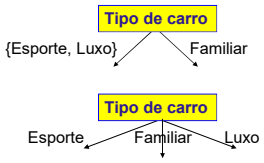
Atrib. B? — Sim → Nó N_3, Não → Nó N_4

	Nó 3	Nó 4
C1	1	5
C2	4	2
Gini = 0.375		

$Gini_{divisão}$ = tarefa de casa = 0.375

Exercício

- Divisão de atributos n-ários
  - Duas alternativas
    - Divisão binária
      - Igual a divisão de atributos binários
    - Divisão n-ária



Exercício

- Divisão n-ária
  - Índice de impureza é calculado para cada um dos n subconjuntos
    - Necessário definir valor de n
  - Resulta em subconjuntos, em geral, mais puros que a divisão binária
    - Porém, nós com menos exemplos ⇒ menor confiança

Exercício

- Definir a melhor divisão considerando divisão binária e divisão n-ária para:

	Tipo de carro		
Classe	Família	Esporte	Luxo
C1	1	2	1
C2	4	1	1
$Gini_{div}$			

	Tipo de carro	
Classe	Família	Esporte e Luxo
C1	1	3
C2	4	2
$Gini_{div}$		

	Tipo de carro	
Classe	Esporte	Família e Luxo
C1	2	2
C2	1	5
$Gini_{div}$		

Exercício

- Definir a melhor divisão considerando divisão binária e divisão n-ária para:

	Tipo de carro		
Classe	Família	Esporte	Luxo
C1	1	2	1
C2	4	1	1
$Gini_{div}$	0.393		

	Tipo de carro	
Classe	Família	Esporte e Luxo
C1	3	1
C2	2	4
$Gini_{div}$	0.400	

	Tipo de carro	
Classe	Esporte	Família e Luxo
C1	2	2
C2	1	5
$Gini_{div}$	0.419	

Atributos n-ários

- Várias possíveis posições de referência
- Cada posição tem uma matriz de contagens associada a ela
  - Contagens das proporções das classes em cada uma das partições

Critério de parada

- Diversas alternativas:
  - Os objetos do nó atual têm a mesma classe
  - Os objetos do nó atual têm valores iguais para os atributos de entrada, mas classes diferentes
  - O número de objetos do nó é menor que um dada quantidade
  - Todos os atributos preditivos já foram incluídos no caminho atual

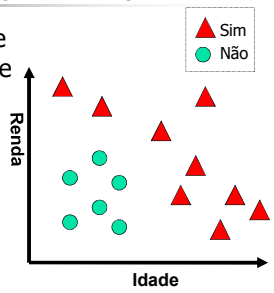
Exemplo

- Sejam os dados abaixo referentes a solicitações de crédito bancário
  - Construir uma árvore de decisão que classifica aplicação para cartão de crédito

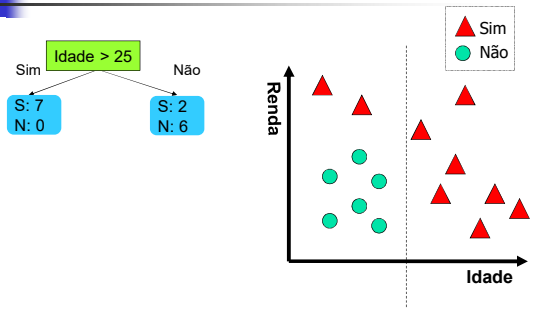
Idade	Renda	Classe
20	2000	Sim
30	5200	Não
60	5000	Sim
40	6000	Não
...		

Busca no espaço de hipóteses

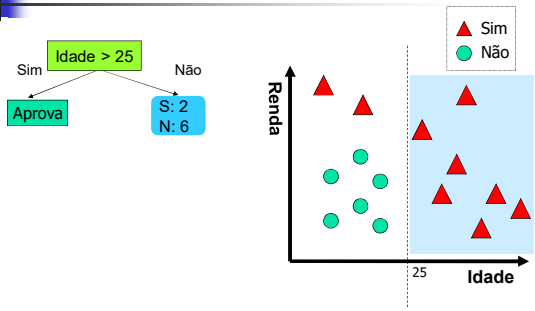
- Construir uma AD que classifica solicitante de cartão de crédito
  - Aprova (Sim)
  - Não aprova (Não)
- Atributos preditivos
  - Idade
  - Renda



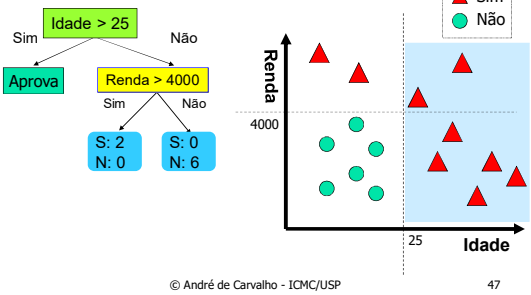
Busca no espaço de hipóteses



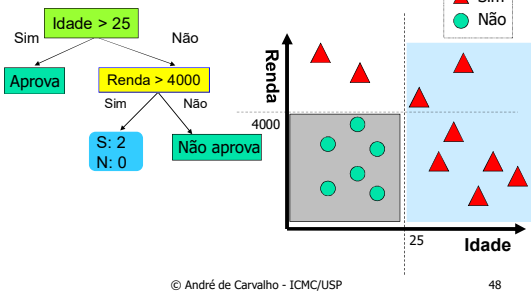
Busca no espaço de hipóteses



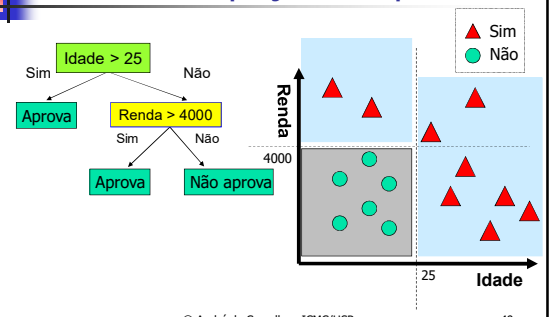
Busca no espaço de hipóteses



Busca no espaço de hipóteses

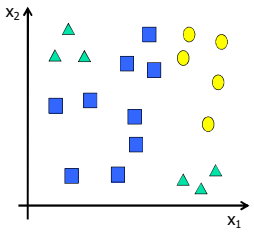


Busca no espaço de hipóteses



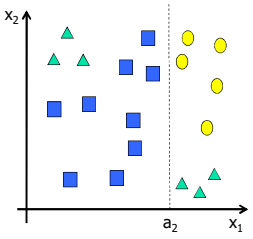
Algoritmo de indução de AD

- Valores numéricos e simbólicos
- Grau de pureza é calculado para cada atributo
  - Atributo que gera menos impureza é escolhido
- Critério de parada precisa ser definido



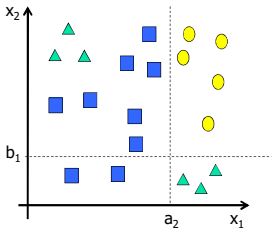
Algoritmo de indução de AD

- Valores numéricos e simbólicos
- Grau de pureza é calculado para cada atributo
  - Atributo que gera menos impureza é escolhido
- Critério de parada precisa ser definido



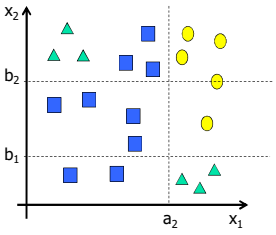
Algoritmo de indução de AD

- Valores numéricos e simbólicos
- Grau de pureza é calculado para cada atributo
  - Atributo que gera menos impureza é escolhido
- Critério de parada precisa ser definido



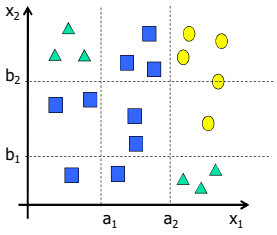
Algoritmo de indução de AD

- Valores numéricos e simbólicos
- Grau de pureza é calculado para cada atributo
  - Atributo que gera menos impureza é escolhido
- Critério de parada precisa ser definido



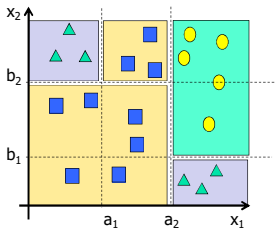
Algoritmo de indução de AD

- Valores numéricos e simbólicos
- Grau de pureza é calculado para cada atributo
  - Atributo que gera menos impureza é escolhido
- Critério de parada precisa ser definido

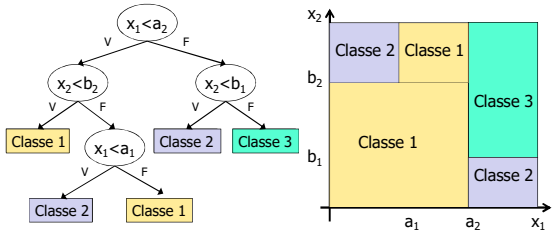


Algoritmo de indução de AD

- Valores numéricos e simbólicos
- Grau de pureza é calculado para cada atributo
  - Atributo que gera menos impureza é escolhido
- Critério de parada precisa ser definido



Árvore e partição do espaço de hipóteses



Espaço de hipóteses

- Cada percurso da raiz a um nó folha representa uma regra de classificação
- Cada folha está associada a uma classe
  - Corresponde a um hiper-retângulo no espaço de soluções
    - Cada classe é representada por um conjunto de hiper-retângulos
    - Interseção de hiper-retângulos é um conjunto vazio
    - União de hiper-retângulos cobre todo o espaço

Aspectos positivos das ADs

- Baixo custo de indução e dedução
- Fácil interpretação da hipótese induzida
  - Para árvores pequenas
- Acurácia comparável a de outros classificadores
  - Para conjuntos de dados de baixa complexidade
- Indica atributos preditivos mais relevantes
- Atributos preditivos podem ser numéricos ou simbólicos

Aspectos negativos das ADs

- Dificuldade para predição de valores contínuos
  - Árvores de regressão
- Baixo desempenho em problemas com muitas classes e poucos dados
- Abordagem gulosa
- Limitação de hipóteses a híper-retângulos

Overfitting

- Partição recursiva pode gerar árvores perfeitamente ajustadas aos dados
- Decisões são baseadas em conjuntos cada vez menores de dados
  - Níveis mais profundos podem ter muito poucos dados
    - Presença de ruído nos dados afeta bastante escolha de atributos para esses nós
    - Reduz capacidade de generalização
    - Poda

Poda de árvores

- Elimina parte da árvore
- Pode ser realizada em duas etapas
  - Durante indução (pré-poda)
    - Parar o crescimento da árvore mais cedo
  - Após indução (pós-poda)
    - Crescer a árvore completa e depois podá-la
    - Mais lento, porém mais confiável

Exercício

- Seja o seguinte cadastro de pacientes:

Nome	Febre	Enjôo	Manchas	Dores	Diagnóstico
João	sim	sim	pequenas	sim	doente
Pedro	não	não	grandes	não	saudável
Maria	sim	sim	pequenas	não	saudável
José	sim	não	grandes	sim	doente
Ana	sim	não	pequenas	sim	saudável
Leila	não	não	grandes	sim	doente

Exercício

- Usando medida de entropia,
  - Induzir uma árvore de decisão capaz de distinguir:
    - Pacientes potencialmente saudáveis
    - Pacientes potencialmente doentes
  - Testar a árvore para novos casos
    - (Luis, não, não, pequenas, sim)
    - (Laura, sim, sim, grandes, sim)

Conclusão

- Árvores de decisão
- Algoritmo de Hunt
- Medidas para escolha de atributos
- Ponto de referência
- Critério de parada
- Espaço de hipóteses

Perguntas