LABIC

SISTEMAS INTELIGENTES

Profa. Roseli Ap. Francelin Romero

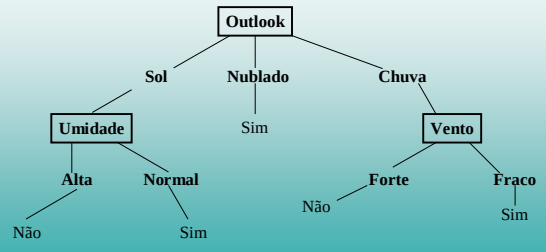
LABIC

Árvore de Decisão

- Representação de Árvores de Decisão
- Algoritmo ID3
- Conceito de Entropia e Ganho de Informação
- Overfitting

LABIC

Árvore de Decisão



LABIC

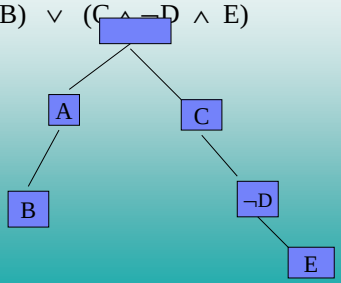
Árvore de Decisão

- PlayTennis
- Cada nó interno testa um ATRIBUTO
- Cada ramo corresponde a um valor do atributo
- Cada nó terminal designa uma classificação.

LABIC

Árvore de Decisão

- $(A \wedge B) \vee (C \wedge \neg D \wedge E)$



LABIC

Árvore de Decisão

- Quando utilizar?
  - Problemas descritos por pares de atributo/valor
  - Função objetivo é discreta
  - Hipóteses disjuntivas são requeridas
  - ruídos nos dados

Exemplos: dignóstico médicos e de equipamentos, análise de crédito.

LABIC

Árvore de Decisão

Indução Top-Down

Main Loop

1. ***A*** o melhor atributo de decisão para o próximo nó.
2. Designar ***A*** como o atributo de decisão p/ o nó.
3. Para cada valor de ***A***, criar um novo descendente.
4. Escolher exemplos de treinamento para os nós folha
5. Se exemplos de treinamento forem perfeitamente classificados, então PARE, senão iterar sobre novos nós folha.

LABIC

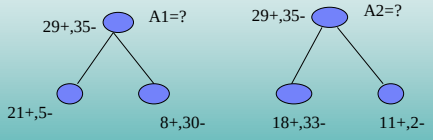
Exemplos de Treinamento

DAY	*OUTLOOK*	*TEMPERATURA*	*UMIDADE*	*VENTO*	*PLAYTENN*
D1	SOL	QUENTE	ALTA	FRACO	NÃO
D2	SOL	QUENTE	ALTA	FORTE	NÃO
D3	NUBLADO	QUENTE	ALTA	FRACO	SIM
D4	CHUVA	AMENO	ALTA	FRACO	SIM
D5	CHUVA	FRIO	NORMAL	FRACO	SIM
D6	CHUVA	FRIO	NORMAL	FORTE	NÃO
D7	NUBLADO	FRIO	NORMAL	FORTE	SIM
D8	SOL	AMENO	ALTA	FRACO	NÃO
D9	SOL	FRIO	NORMAL	FRACO	SIM
D10	CHUVA	AMENO	NORMAL	FRACO	SIM
D11	SOL	AMENO	NORMAL	FORTE	SIM
D12	NUBLADO	AMENO	ALTA	FORTE	SIM
D13	NUBLADO	QUENTE	NORMAL	FRACO	SIM
D14	CHUVA	AMENO	ALTA	FORTE	NÃO

LABIC

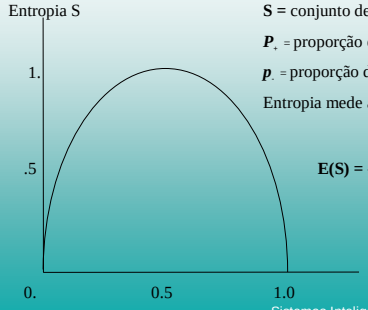
Árvore de Decisão

- Qual atributo é o MELHOR?



LABIC

Entropia



S = conjunto de exemplos treinamento

p_+ = proporção de exemplos positivos

p_- = proporção de exemplos negativos

Entropia mede a IMPURIDADE de S

$$E(S) = - p_+ \log_2 p_+ - p_- \log_2 p_-$$

LABIC

Entropia

- Da teoria de Informação:

Entropia (***S***) = número esperado de bits necessários para representar uma classe (+ or -) dos membros de ***S*** (sob código de menor comprimento e ótimo).

- PORQUE? (se $p_+ = 1$ ou $p_+ = 0.5$)

Um código de compr. ótimo designa $-\log_2 p$ bits com probabilidade p. Então, o número esperado de bits representar + ou - membros de ***S é***:

$$p_+ (-\log_2 p_+) + p_- (-\log_2 p_-)$$

$$\text{Entropia } (S) \equiv - p_+ \log_2 p_+ - p_- \log_2 p_-$$

LABIC

Entropia

- EXEMPLO:

$$S = [\, 9+ , 5- \,]$$

$$\text{ENTROPIA } ([\, 9+ , 5- \,]) = -(9/14) \log_2(9/14) - (5/14) \log_2(5/14) = 0.940$$

LABIC

Ganho de Informação

- Gain (***S, A***) = redução esperada na entropia devido a escolha do atributo A.

$$Gain(S, A) \equiv Entropia(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropia(S_v)$$

Valor (Wind) =Fraco, Forte

S= [9+, 5-] S_{fraco} = [6+, 2-] S_{forte} = [3+, 3-]

$$Gain(S, Wind) = Entropia(S) - \sum_{v \in \{Fraco, Forte\}} \frac{|S_v|}{|S|} Entropia(S_v)$$

LABIC

=Entropia(S) - (8/14)Entropia(Fraco) - (6/14)Entropia(Forte)=

0.94 - (8/14) 0.811 - (6/14) 1.00 = 0.048

LABIC

Exemplos de Treinamento

DAY	OUTLOOK	TEMPERATURA	UMIDADE	VENTO	PLAYTENN
D1	SOL	QUENTE	ALTA	FRACO	NÃO
D2	SOL	QUENTE	ALTA	FORTE	NÃO
D3	NUBLADO	QUENTE	ALTA	FRACO	SIM
D4	CHUVA	AMENO	ALTA	FRACO	SIM
D5	CHUVA	FRIO	NORMAL	FRACO	SIM
D6	CHUVA	FRIO	NORMAL	FORTE	NÃO
D7	NUBLADO	FRIO	NORMAL	FORTE	SIM
D8	SOL	AMENO	ALTA	FRACO	NÃO
D9	SOL	FRIO	NORMAL	FRACO	SIM
D10	CHUVA	AMENO	NORMAL	FRACO	SIM
D11	SOL	AMENO	NORMAL	FORTE	SIM
D12	NUBLADO	AMENO	ALTA	FORTE	SIM
D13	NUBLADO	QUENTE	NORMAL	FRACO	SIM
D14	CHUVA	AMENO	ALTA	FORTE	NÃO

LABIC

Selecionando o Próximo Atributo

Qual atributo é o melhor classificador?

S = [9+, 5-]
E = 0.940

Umidade

ALTA: [3+, 4-] E = 0.985

NORMAL: [6+, 1-] E = 0.592

GAIN (S, Umidade) = 0.94 -(7/14) 0.985 + - (7/14) 0.592 = 0.151

S = [9+, 5-]
E = 0.940

Vento

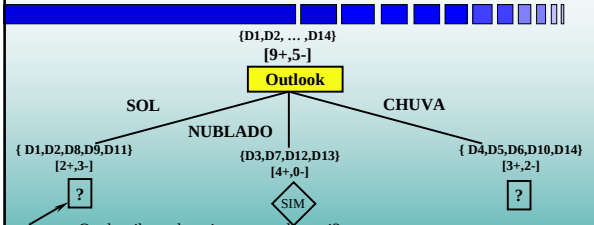
FRACO: [6+, 2-] E = 0.811

FORTE: [3+, 3-] E = 1.00

GAIN (S, Vento) = 0.94 -(8/14) 0.811 + - (6/14) 1.00 = 0.048

LABIC

Selecionando o Próximo Atributo



- Qual atributo deveria ser testado aqui?

S_{sol} = { D1,D2,D8,D9,D11}

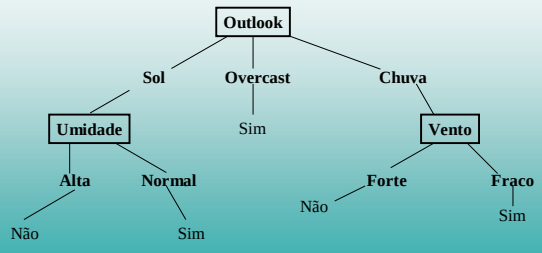
Gain (S_{sol} , Umidade) = 0.97 - (3/5) 0.0 - (2/5) 0.0 = 0.97

Gain (S_{sol} , Temperatura) = 0.97 - (2/5) 0.0 - (2/5) 1.0 - (1/5) 0.0 = 0.57

Gain (S_{sol} , Vento) = 0.97 - (2/5) 1.0 - (3/5) 0.918 = 0.19

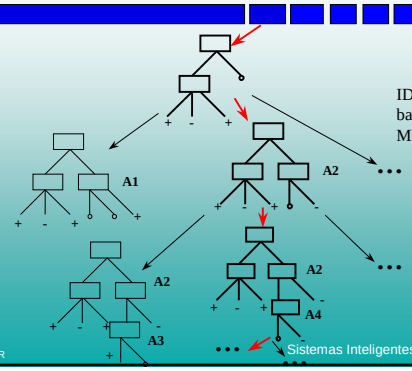
LABIC

Árvore de Decisão



LABIC

Espaço de Hipóteses pesquisado por ID3



ID3 trabalha sem backtracking=> MINIMO local

A tendencia indutiva ID3 favorece arvores mais simples (?)

LABIC

- ID3 – Weka
  C4.5

LABIC

Atributos de valores continuos

- ID3 é restrito a assumir apenas valores discretos:

- atributo alvo predito pela árvore é discreto

- os atributos testado nos nós de decisão da árvore deve também ser discretos.

Mas, a segunda restrição pode ser relaxada para valores contínuos

Para um atributo A, que é um atributo de valor contínuo, o algoritmo cria um novo

LABIC

- Um novo atributo booleano Ac que
  - se A < c então Ac = true
    caso contrário Ac = false

Exemplo:

Temperatura: 40 48 60 72 80 90

PlayTennis: No No Yes Yes Yes No

Qual **valor de c** escolher?

LABIC

- O **valor de c** deveria ser escolhido de modo a produzir o maior ganho de informação.
  - Fayyad (1991) mostrou que o **valor de c**

que maximiza o ganho de informação fica entre os limites de mudança do atributo.

Exemplo:

PlayTennys muda : (48+60)/2 --- Temp $_{>54}$

(80+90)/2 --- Temp $_{>85}$

LABIC

- Atributos candidatos: Temp $_{>54}$ Temp $_{>85}$
- Calculado o ganho de informação para cada atributo é selecionamos o melhor:

  Temp $_{>54.}$

Este atributo booleano criado pode então competir com outros atributos candidatos discretos para o crescimento da árvores

LABIC

Exercícios

I - Construa árvores de decisão para as seguintes funções booleanas:

(a) A $\wedge \neg$ B

(b) A $\vee$ [B $\wedge$ C]

(c) A XOR B

(d) [A $\wedge$ B] $\vee$ [C $\wedge$ D]

LABIC

II - Seja os exemplos de treinamento:

		a1	a2
1	+	T	T
2	+	T	T
3	-	T	F
4	+	F	F
5	-	F	T
6	-	F	T

LABIC

(a) Qual é a entropia desta coleção de exemplos de treinamento com a função objetivo de classificação.

(b) Qual é o ganho e informação de a2 relativo aos exemplos de treinamento?

LABIC

- Referencia:

Mitchell, T., “Machine Learning”, McGraw Hill, 1997