ARVORES DE DECISÃO

SCC0276/SCC5871/MAI5025

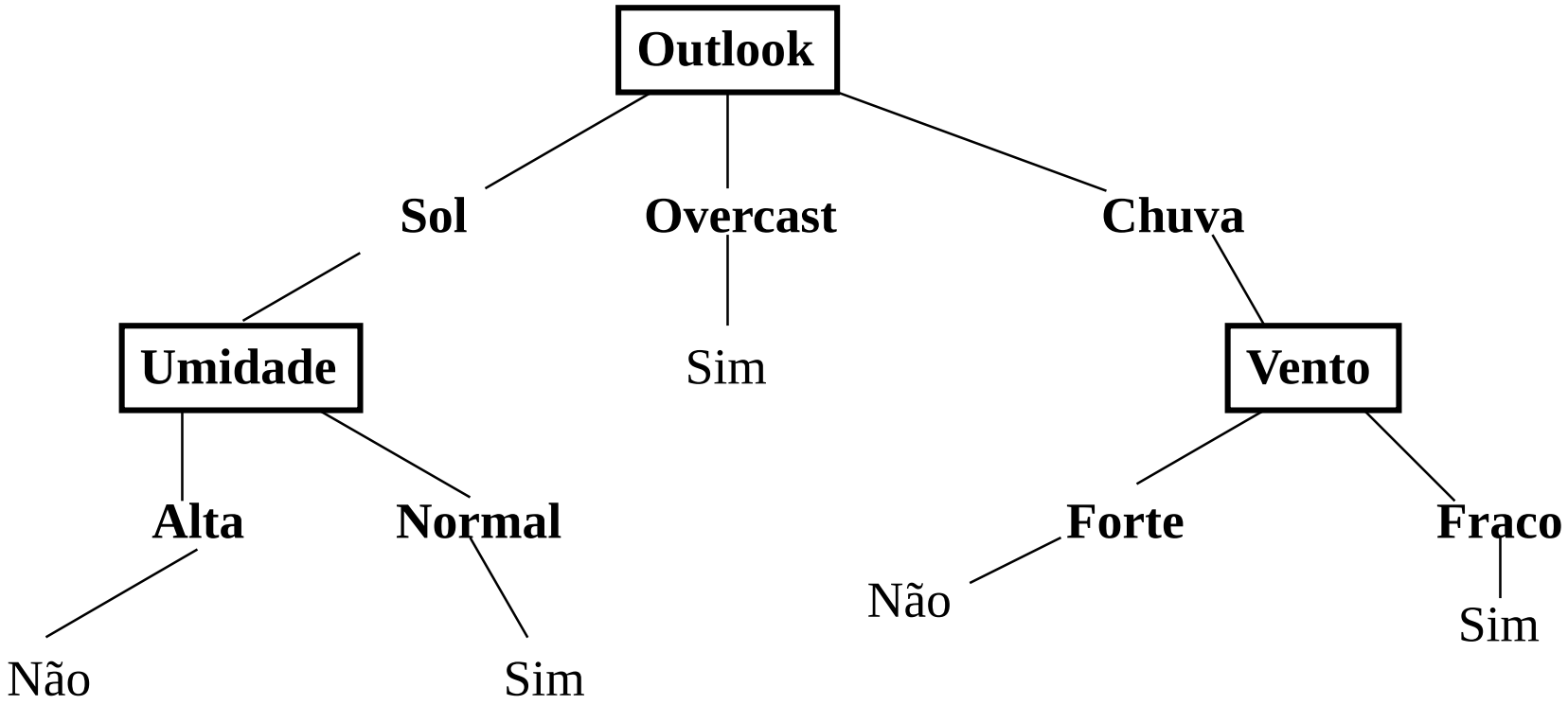
APRENDIZADO DE MÁQUINA

Profa. Roseli Ap. Francelin Romero

Árvore de Decisão

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

Árvore de Decisão

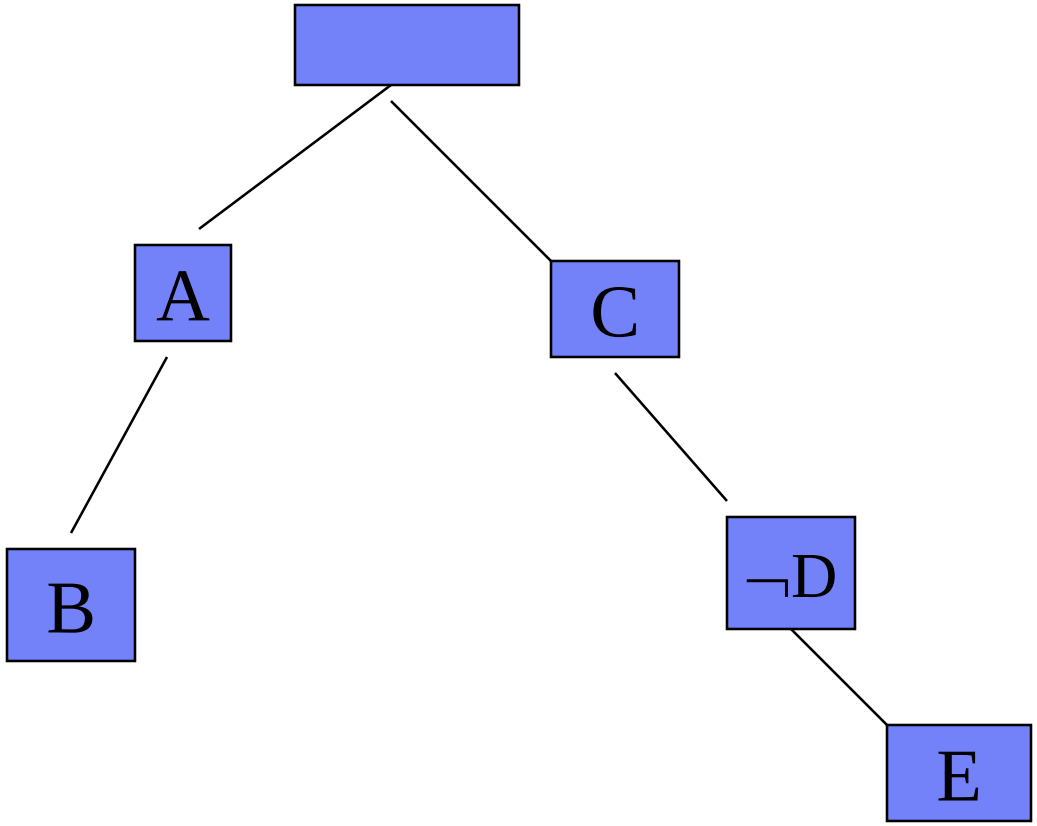


Árvore de Decisão

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

Árvore de Decisão

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



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.

Árvore de Decisão

Indução Top-Down

Main Loop

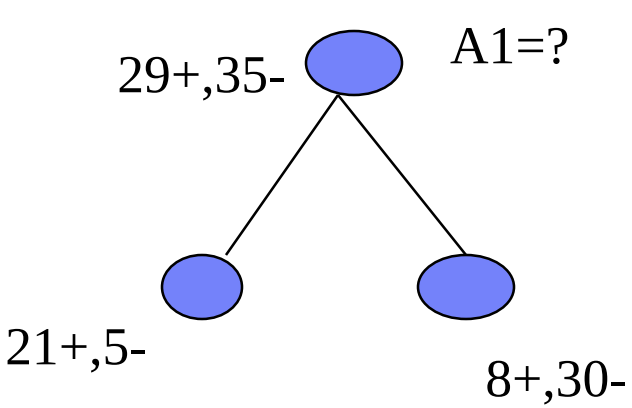
1. ***A*** o melhor atributo de decisão para o próximo nó.
2. Designar ***A*** como o atributo de decisão p/ o nó.
3. Para cada valor de ***A***, criar um novo descendente.
4. Escolher exemplos de treinamento para os nós folha
5. Se exemplos de treinamento forem perfeitamente classificados, então PARE,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

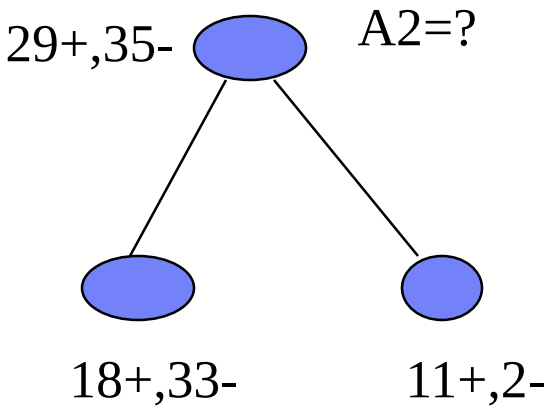
Exemplos de Treinamento

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

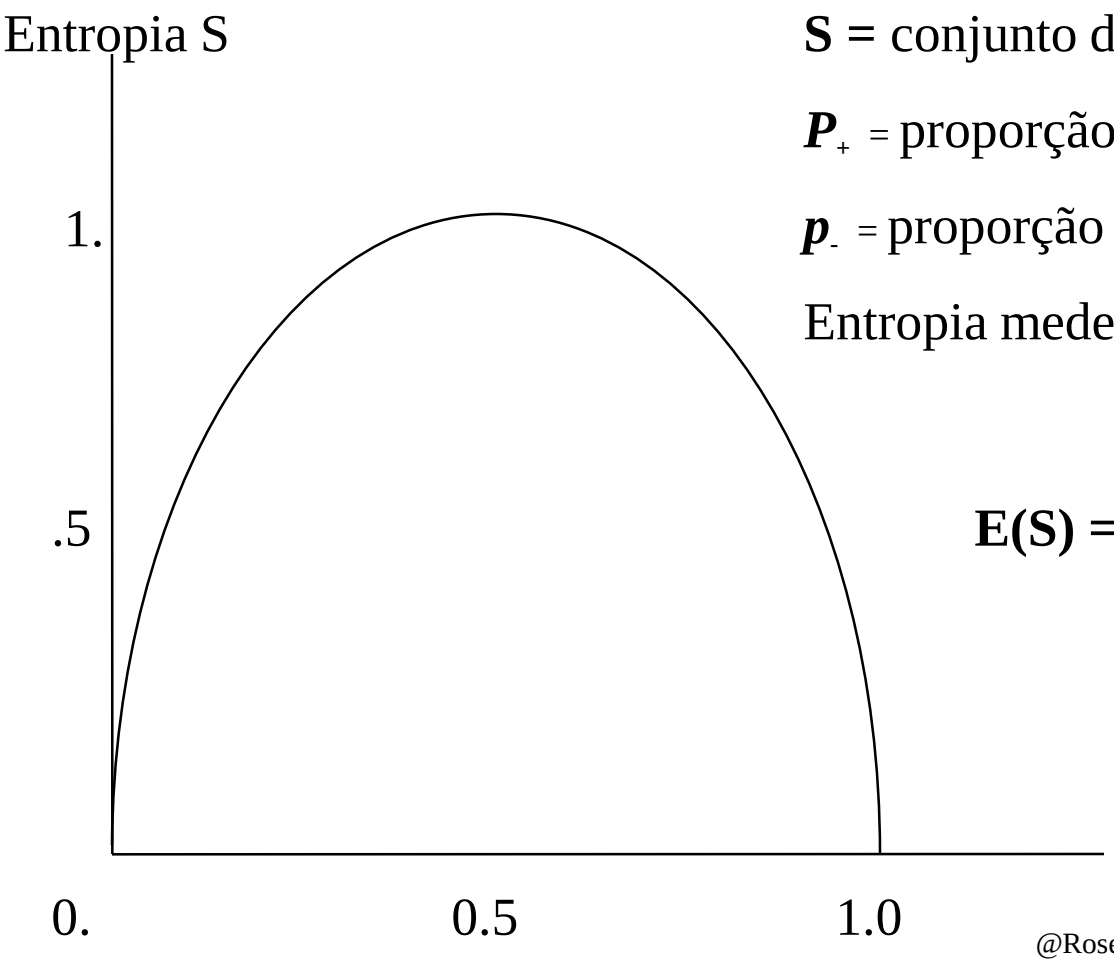
Árvore de Decisão

- Qual atributo é o MELHOR?





Entropia



S = conjunto de exemplos treinamento

P_+ = proporção de exemplos positivos

p_- = proporção de exemplos negativos

Entropia mede a IMPURIDADE de S

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

@Roseli A. F. Romero
SCC-ICMC - USP

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***:

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

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

Entropia

- EXEMPLO:

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

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

Ganho de Informação

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

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

Valor (Wind) =Fraco, Forte

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

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

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

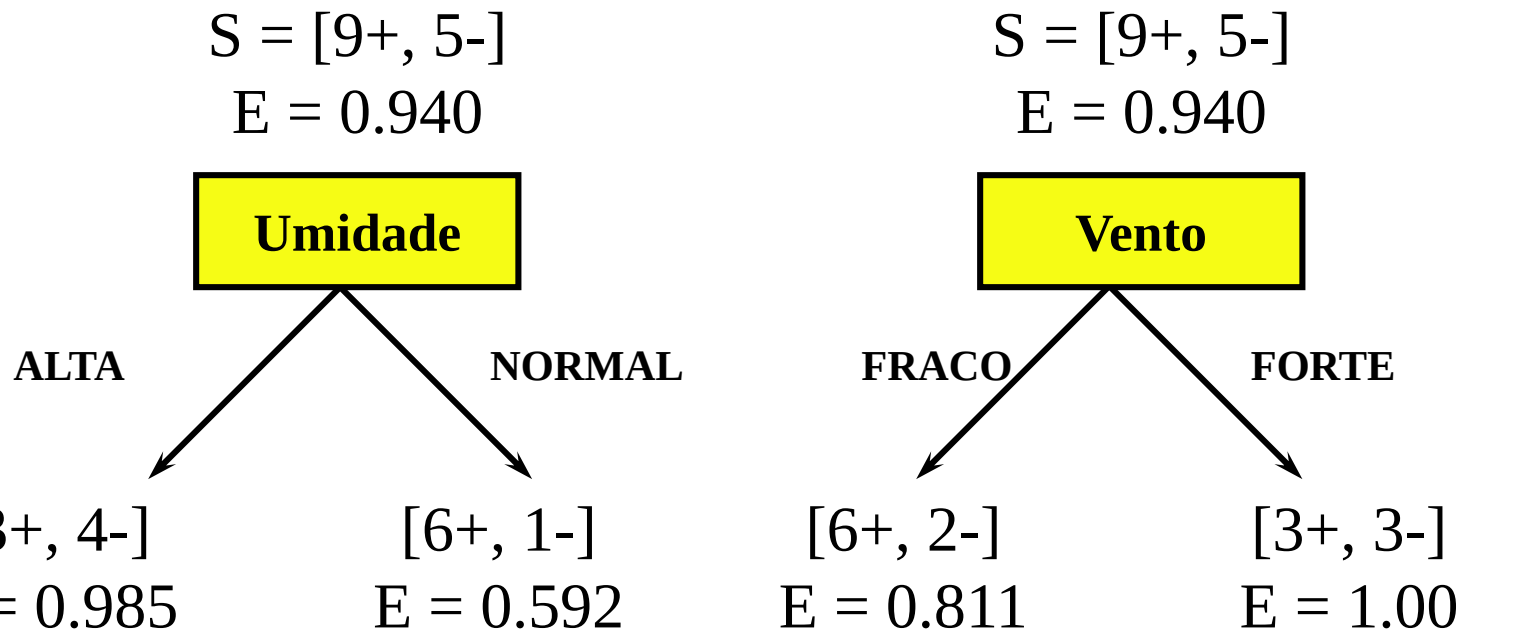
0.94 - (8/14) 0.811 - (6/14) 1.00 = 0.048

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

Selecionando o Próximo Atributo

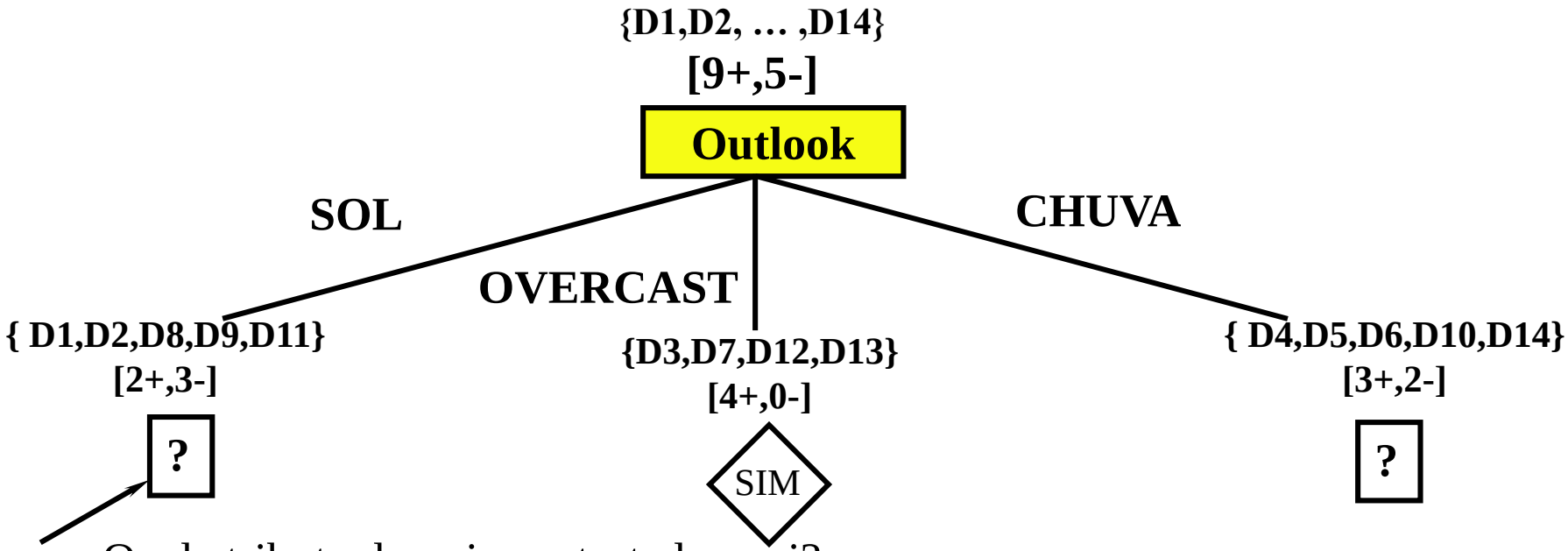
Qual atributo é o melhor classificador?



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

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

Selecionando o Próximo Atributo



- Qual atributo deveria ser testado aqui?

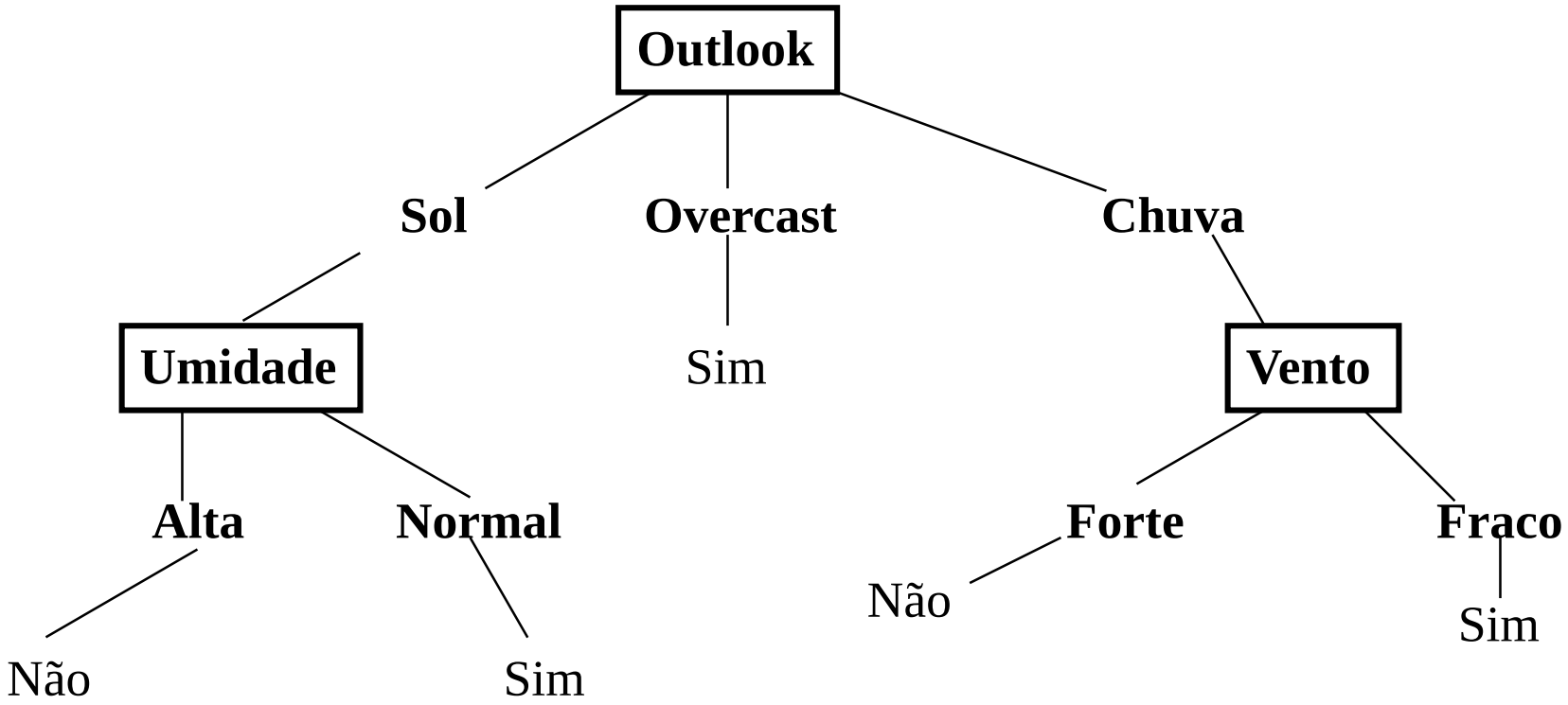
S_{sol} = { D1,D2,D8,D9,D11}

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

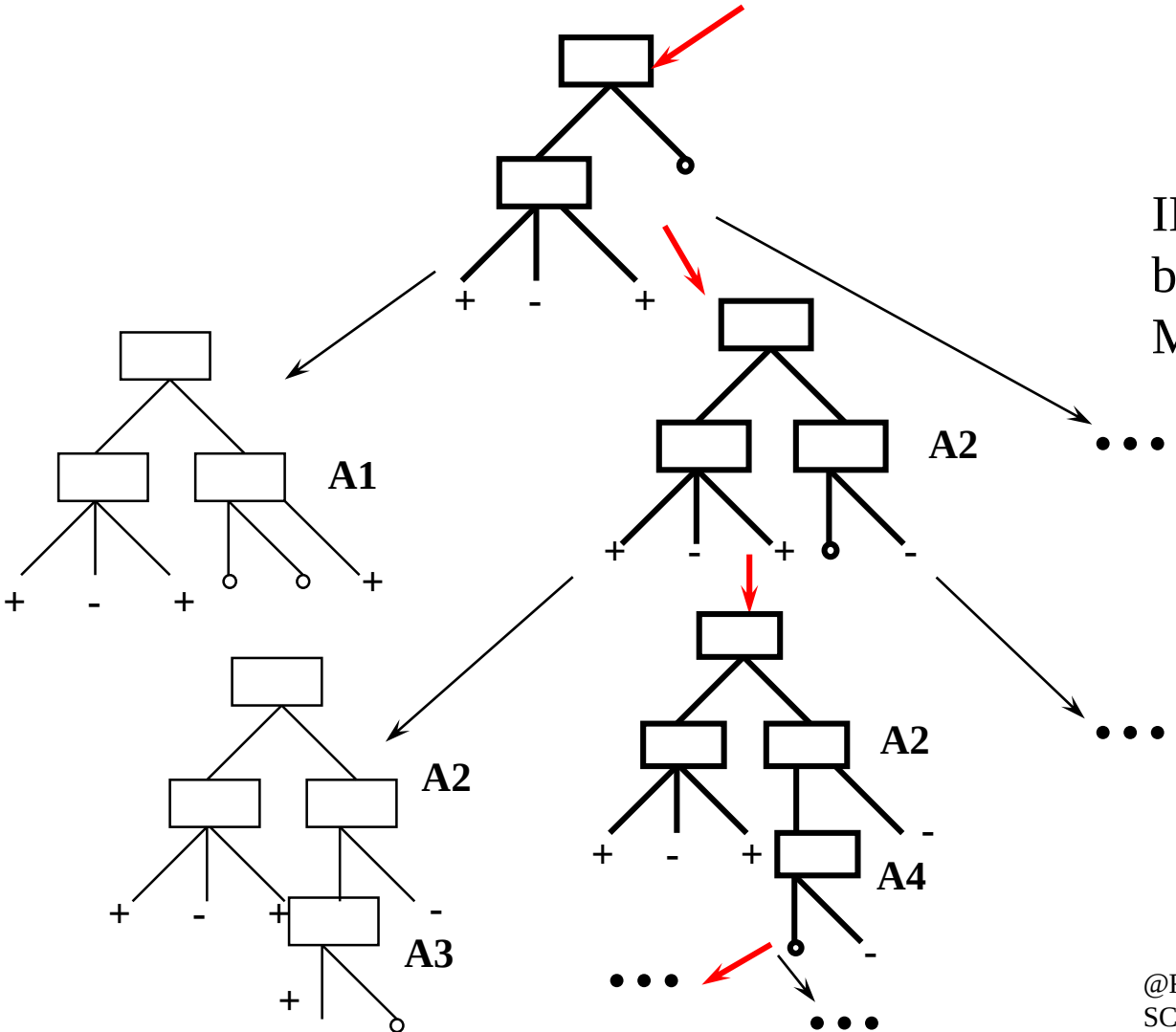
Gain (S_{sol} , Temperatura) = 0.97 - (2/5) 0.0 - (2/5) 1.0 - (1/5) 0.0 = 0.57

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

Árvore de Decisão



Espaço de Hipóteses pesquisado por ID3



ID3 trabalha sem backtracking=> MINIMO local

A tendencia indutiva ID3 favorece arvores mais simples (?)

@Roseli A. F. Romero
SCC-ICMC - USP

Espaço de Hipóteses pesquisado por ID3

- Espaço de Hipótese é completo!
  - Com certeza a função alvo se encontra no espaço de busca
- ID3 fornece uma única saída
  - Não se pode considerar quantas DT alternativas são consistente com o conjunto de treinamento
- Nenhum back-tracking
  - Mínimo Local …
- Usa propr.estatistica dos dados (ganho de inf.)
  - Robusto a dados com ruídos …

Inductive Bias

- Como ID3 consegue generalizar a partir dos exemplos aprendidos?
- Inductive bias de ID3:

  - seleciona a favor de “árvores menores preferidas”

  - seleciona árvores que colocam os atributos com maior “ganho de informação” próximo ao nó raiz.

Overfitting

- Considere o erro da hipótese ***h*** sobre
  - conjunto de treinamento: $error_{train}(h)$
  - Distribuição inteira ***D*** dos dados: $error_{D}(h)$
- Dada uma Hipótese $\boldsymbol{h} \in \boldsymbol{H}$ para a qual ocorre um **overfitting** se existir uma hipótese alternativa $\boldsymbol{h'} \in \boldsymbol{H}$ tal que

$$error_{train}(h) < error_{train}(h')$$

mas

$$error_{D}(h) > error_{D}(h')$$

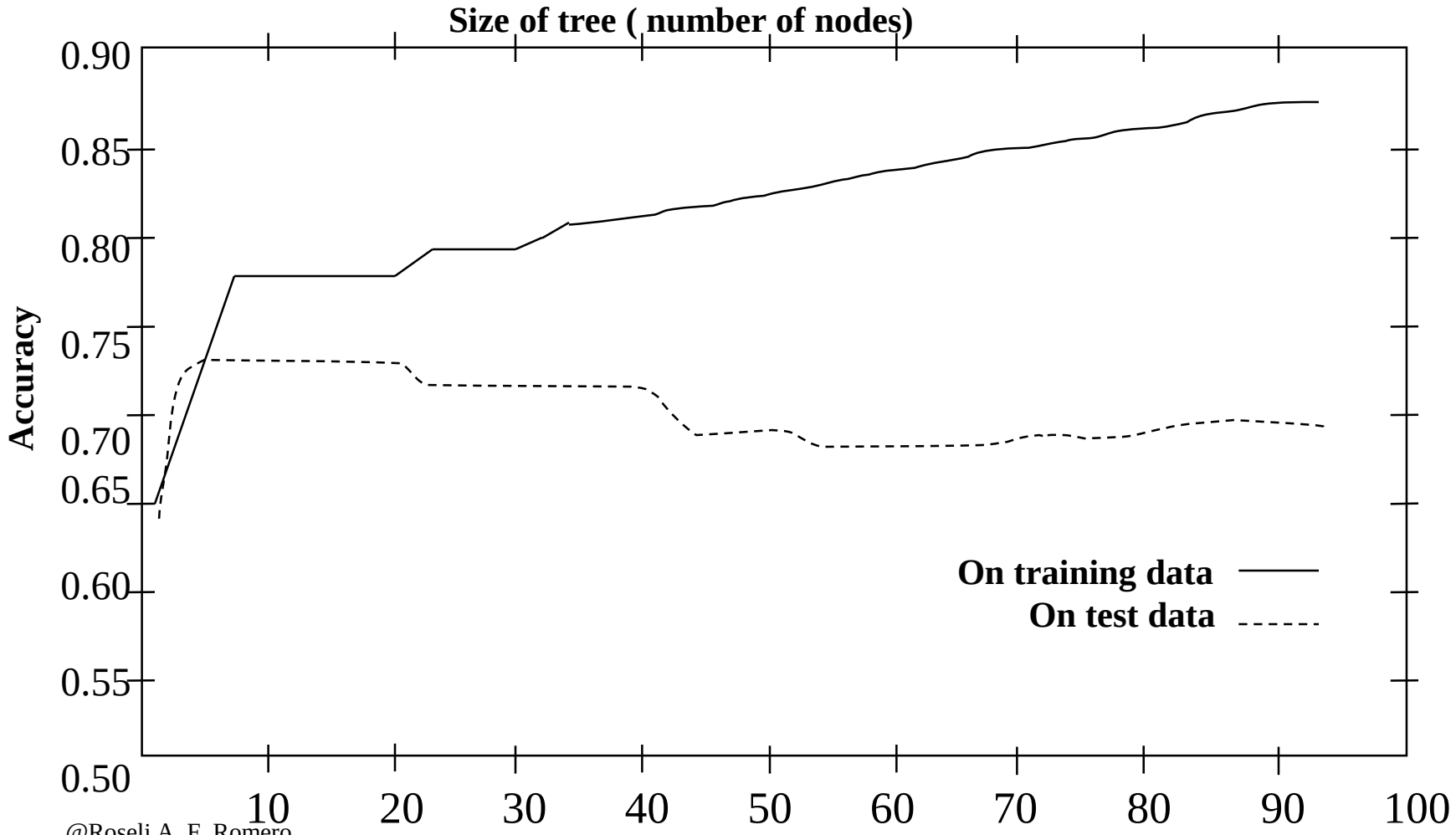
@Roseli A. F. Romero
SCC-ICMC - USP

Overfitting é possível de ocorrer quando:

- existem ruídos nos dados ou erros aleatórios.
- Existem poucos exemplos associados com os nós folhas.

Ocorre muito na prática em métodos de aprendizado. Um estudo experimental com ID3 envolvendo 5 tarefas com dados não determinísticos e ruídos, overfitting diminuiu a precisão da DT em 10 - 25% dos problemas.

Overfitting in Decision Trees Learning (Mitchell, 1998)

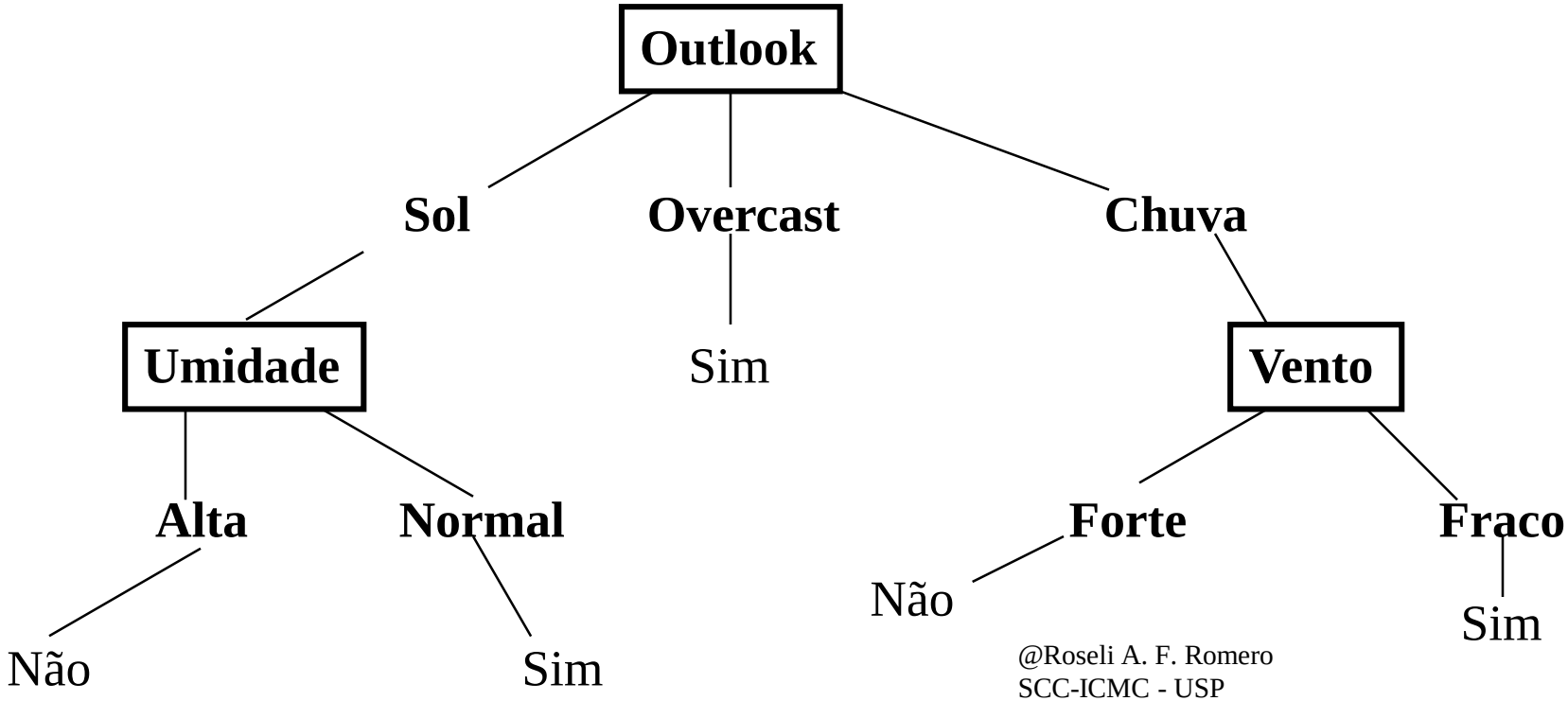


Overfitting in Decision Trees

Considere a adição de ruido no Exemplo anterior no. 15:

(Sol, Quente, Normal, Forte, Playtennis = Não)

Qual o efeito na árvore de decisão anterior?



Overfitting

Evitar Overfitting

- Abordagens que cessam o crescimento da árvore antes que ela atinja o ponto onde ela perfeitamente classifica os dados de treinamento.
- Abordagens que permitem ocorrer um overfitting e então depois através de um pruning diminuir a árvore.

Overfitting

A segunda abordagem é mais usada porque na primeira não se sabe exatamente qual é o ponto onde se deve parar.

Pruning de Erro Reduzido (Quinlan, 1987)

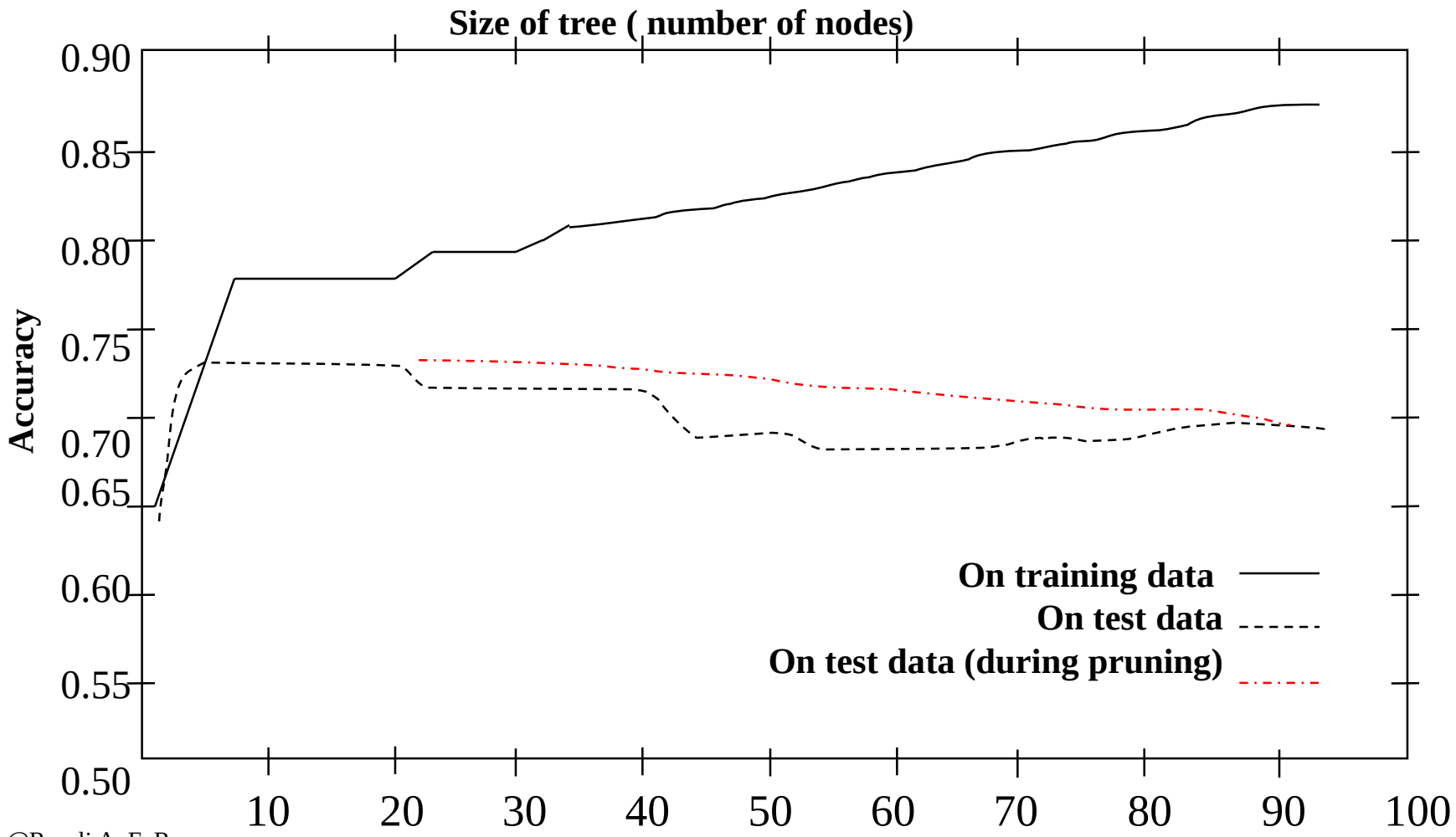
Considerar cada nó da árvore como um nó candidato a corte. Cortar um nó significa remover a sub- árvore a partir daquele nó, tornando-o um nó terminal (ou folha) e designando-o a mais comum classificação dos exemplos de treinamento afiliados com aquele nó.

Pruning de Erro Reduzido

Nós são removidos apenas se a árvore resultante não desempenha pior que aquela original.

Nós são cortados iterativamente sempre escolhendo o nó cuja remoção aumenta a precisão da árvore de decisão sobre o conjunto de Validação.

Efeito do Pruning de Erro Reduzido



Regra Post-Pruning

Usado por C4.5 (Quinlan, l993)

- Deixar ocorrer overfitting crescendo a árvore para treinar os dados.
- Converter a árvore aprendida num conjunto de regras.
- Cortar (generalizar) cada regra removendo algumas pré-condições que resultam em melhorar sua precisão estimada.
- Escolher as regras finais por sua precisão estimada e considerá-las nesta sequência quando classificando instâncias subsequentes.

Regra Post-Pruning

- IF (Outlook = Sunny) $\wedge$ (Humidity = High)
  THEN PlayTennis = No

- cada regra é podada removendo-se algum antecedente: ou (Outlook = Sunny) ou (Humidity = High)
- escolhe-se o antecedente que melhora a precisão estimada da regra. Nada é feito se a precisão piorar.

Regra Post-Pruning

- Uma forma de avaliar a precisão estimada da regra é considerar um Conjunto de Validação disjunto do conjunto de treinamento.
- Usar um conjunto de validação para guiar a poda é uma indicação boa quando se tem grande volume de dados. Do contrário, ele acaba reduzindo ainda mais o conjunto de treinamento.

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

- 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?

- 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 $_{>85}$

- 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

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

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

(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? II - Seja os exemplos de treinamento: