Curso 2 – CD, AM e DM
MBA EM IA e BIGDATA

ÁRVORES DE DECISÃO

PROFA. ROSELI AP. FRANCELIN ROMERO

SCC – ICMC - USP

ÁRVORES DE DECISÃO

- É um dos métodos mais **comuns** aplicados ao aprendizado de máquina.
- O que é árvore de decisão?
- A árvore de decisão é uma ferramenta de suporte para ajudar uma pessoa, ou um grupo de pessoas, a tomarem uma decisão ao visualizar as suas ramificações e consequências.
- Uma árvore de decisão geralmente começa com um único nó, que se divide em possíveis resultados. Cada um desses resultados leva a nós adicionais, que se ramificam em outras possibilidades. Assim, cria-se uma forma de árvore.
- Com uma árvore de decisão concluída, você está pronto para analisar a decisão diante de você.

ÁRVORES DE DECISÃO

- Exemplo: Treinamento das equipes de suporte e sucesso do cliente.

- Os atendentes precisam saber o que a empresa espera deles em diferentes situações para evitar confusão e deixar o usuário irritado.

- Por isso, é preciso dividir a equipe em times, cada um com sua finalidade definida: atendimento, cancelamento e troca de planos, por exemplo.

- Uma vez que a divisão é estabelecida, faça treinamentos separados e personalizados. Como exemplo, temos a situação do churn – métrica que indica o quanto a companhia perdeu de receita ou clientes. Costuma significar o ápice do descontentamento do usuário e deve ser evitado a todo custo.

- Se o usuário quiser cancelar o produto, o atendente deve seguir um fluxo pré-estabelecido para reverter a situação: entender os motivos que levaram o cliente a pedir o cancelamento, oferecer benefícios e seguir de acordo com a resposta do consumidor.

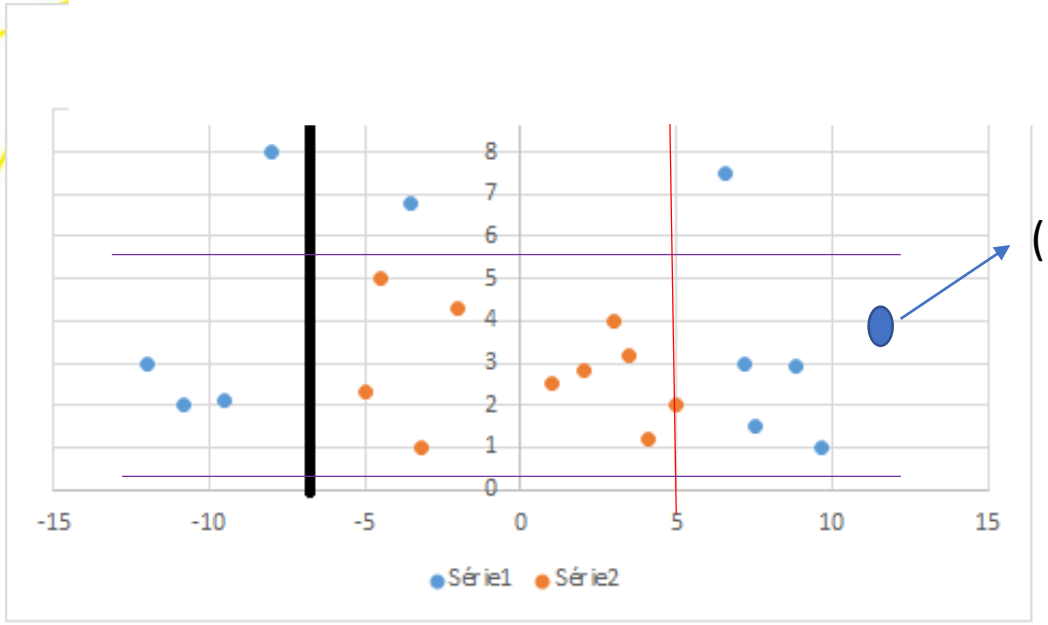
ÁRVORES DE DECISÃO

- A árvore de decisão permite visualizar as múltiplas etapas que podem seguir cada decisão.
- Dessa forma, aqueles que se baseiam na árvore para decidir têm a liberdade de supor diversos cenários e pensar quais seriam as consequências de cada um deles.
- Cada resolução deve ser pensada considerando o objetivo a longo prazo da empresa. Ele deve ser o guia, apontando quais são os caminhos a serem trilhados. Então é preciso pensar se aquela decisão vai impactar no futuro da companhia e, caso afirmativo, pondere se o impacto será positivo ou negativo.

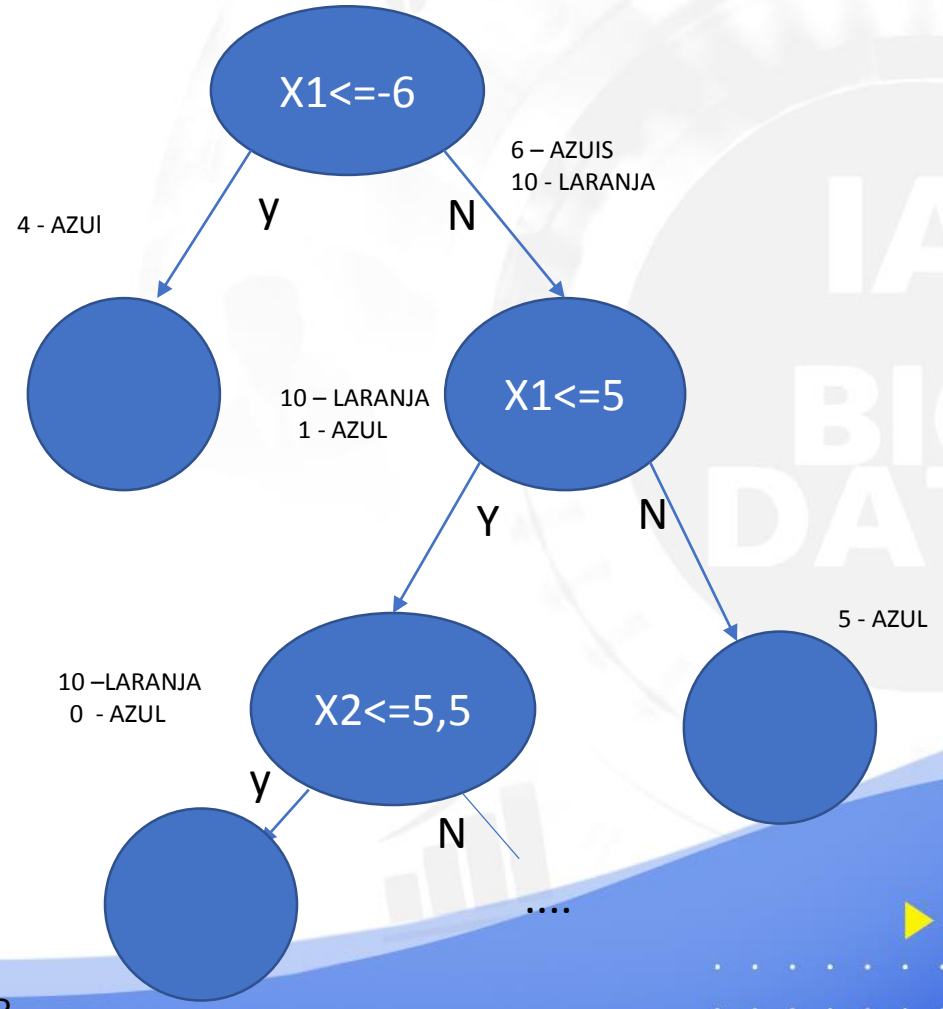
ÁRVORES DE DECISÃO

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

ÁRVORES DE DECISÃO



(12,4)



ÁRVORES DE DECISÃO

- QUAL DEVE SER O PRIMEIRO ATRIBUTO A SER ESCOLHIDO PARA QUE A ÁRVORE SEJA A MENOR POSSIVEL ?
- QUAL SERIA A MELHOR FORMA DE PARTICIONAR O ESPAÇO?

ÁRVORES DE DECISÃO

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

ÁRVORES DE DECISÃO

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

Exemplos:

dignósticos 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.

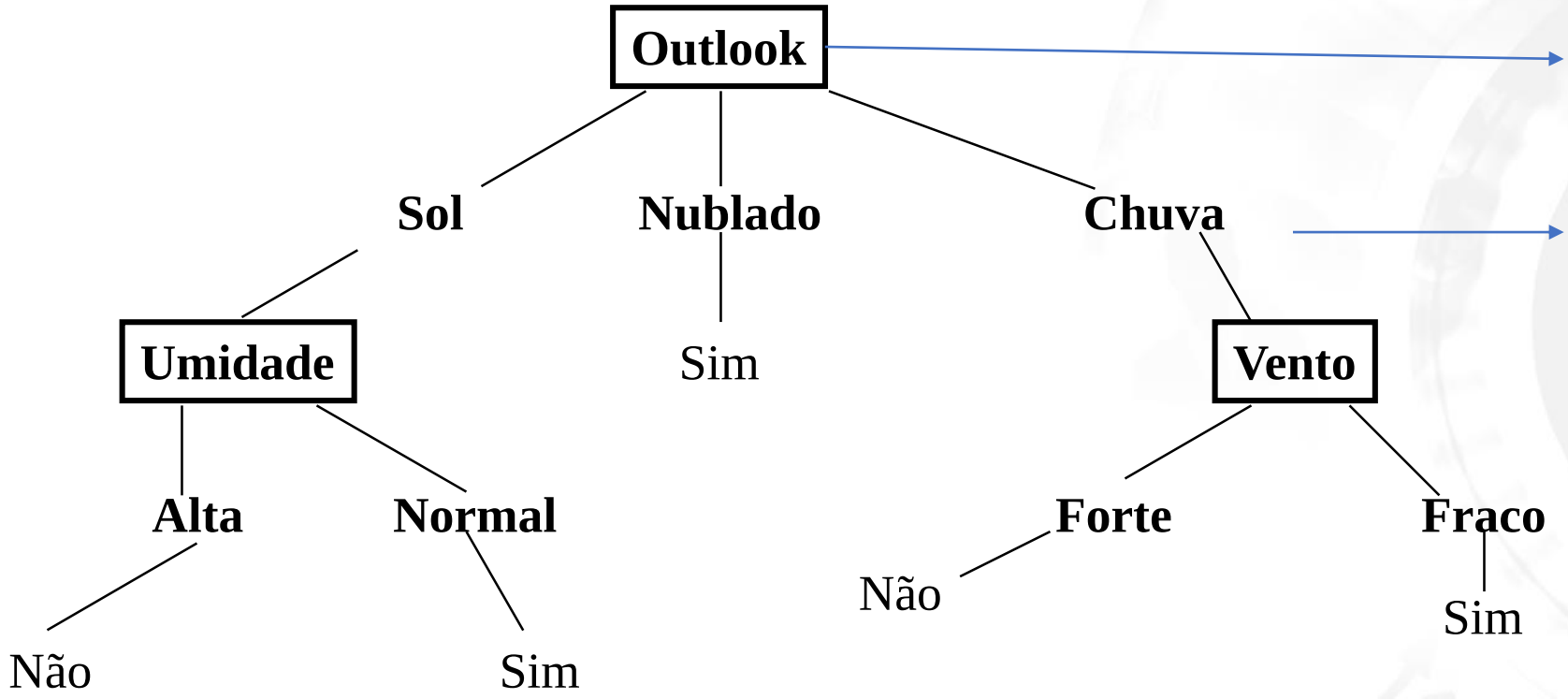
4. Escolher exemplos de treinamento para os nós folha

5. Se exemplos de treinamento forem perfeitamente classificados, então PARE, senão iterar sobre novos nós folha.

Exemplos de Treinamento

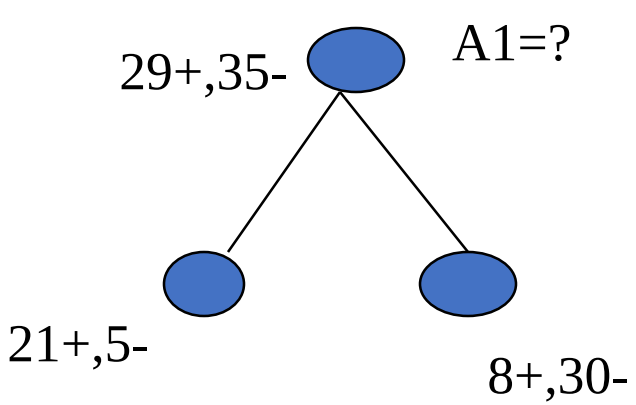
DAY	*OUTLOOK*	*TEMPERATURA*	*UMIDADE*	*VENTO*	*PLAYTENN*
D1	SOL	QUENTE	ALTA	FRACO	NÃO
D2	SOL	QUENTE	ALTA	FORTE	NÃO
D3	NUBLADO	QUENTE	ALTA	FRACO	SIM
D4	CHUVA	AMENO	ALTA	FRACO	SIM
D5	CHUVA	FRIO	NORMAL	FRACO	SIM
D6	CHUVA	FRIO	NORMAL	FORTE	NÃO
D7	NUBLADO	FRIO	NORMAL	FORTE	SIM
D8	SOL	AMENO	ALTA	FRACO	NÃO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

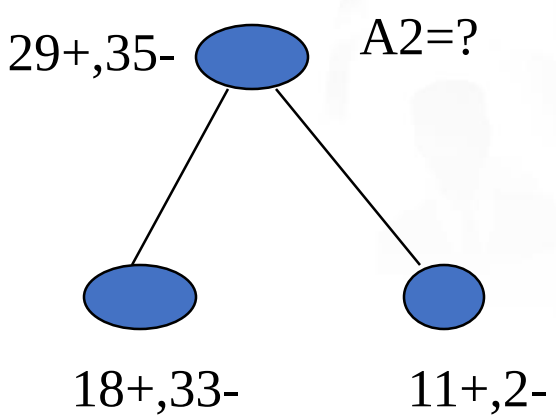
Árvore de Decisão



Árvore de Decisão

- Qual atributo é o MELHOR?

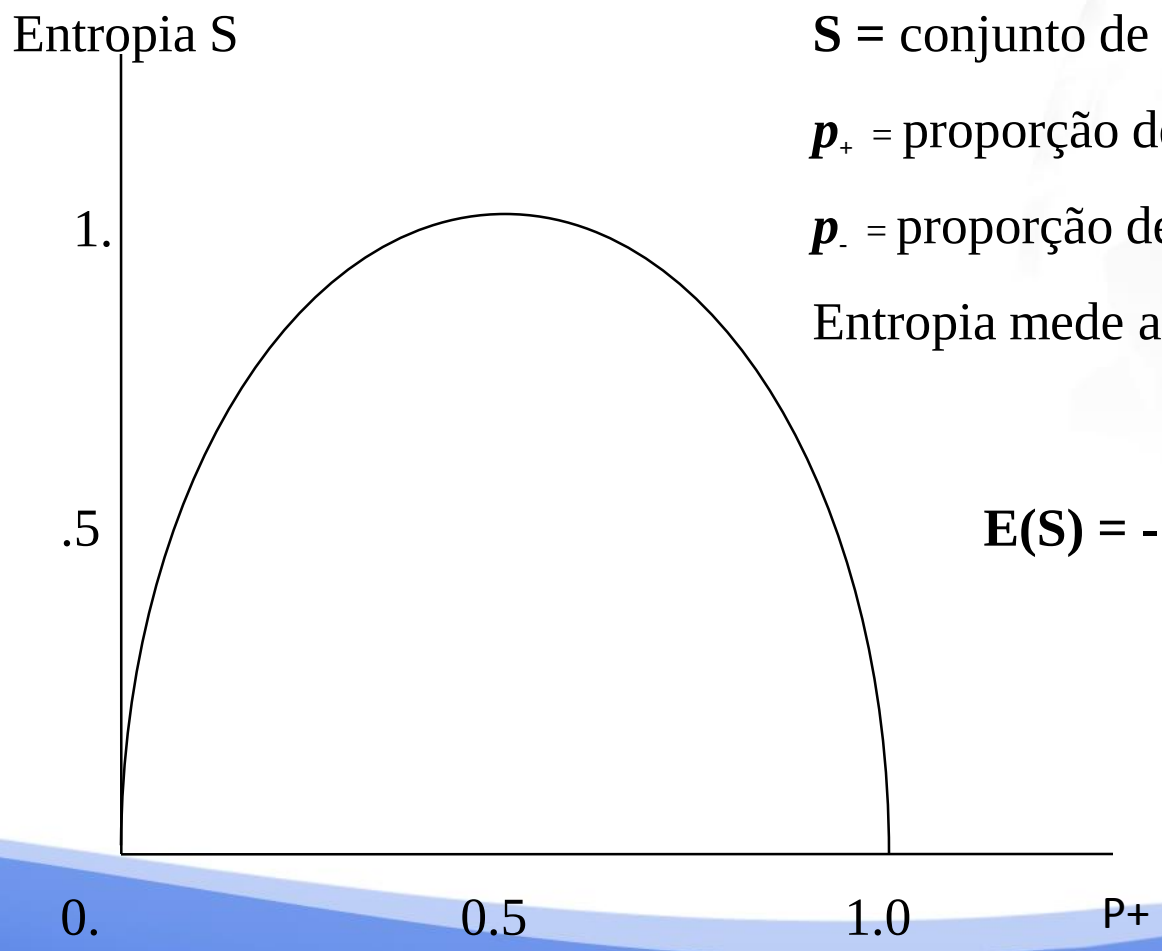




ENTROPIA

- Em Física, falamos de entropia (geralmente simbolizada pela letra S) para se referir ao grau de equilíbrio de um sistema termodinâmico, ou melhor, ao seu nível de tendência à desordem (variação de entropia).
- É comumente associada ao grau de "**desordem**" ou "**aleatoriedade**" de um sistema.

Entropia



S = conjunto de exemplos treinamento

p_+ = proporção de exemplos positivos

p_- = proporção de exemplos negativos

Entropia mede a IMPURIDADE de S

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

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).

Um código de comprimento ótimo designa $\mathbf{-\log_2 p}$ bits com probabilidade p.

Então, o número esperado de bits representar + ou - membros de ***S é***:

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

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

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

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

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

Valor (Wind) ={Fraco, Forte}

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

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

Ganho de Informação

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

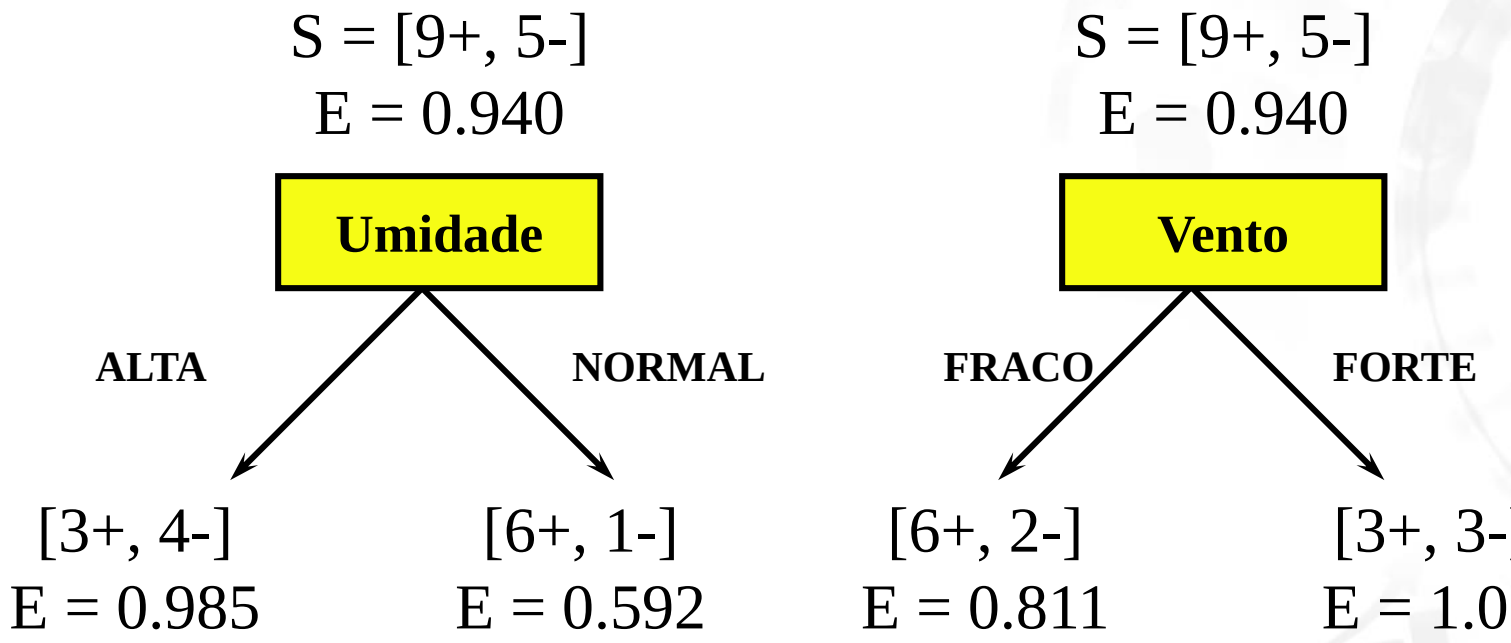
0.94 - (8/14) 0.811 - (6/14) 1.00 = 0.048

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

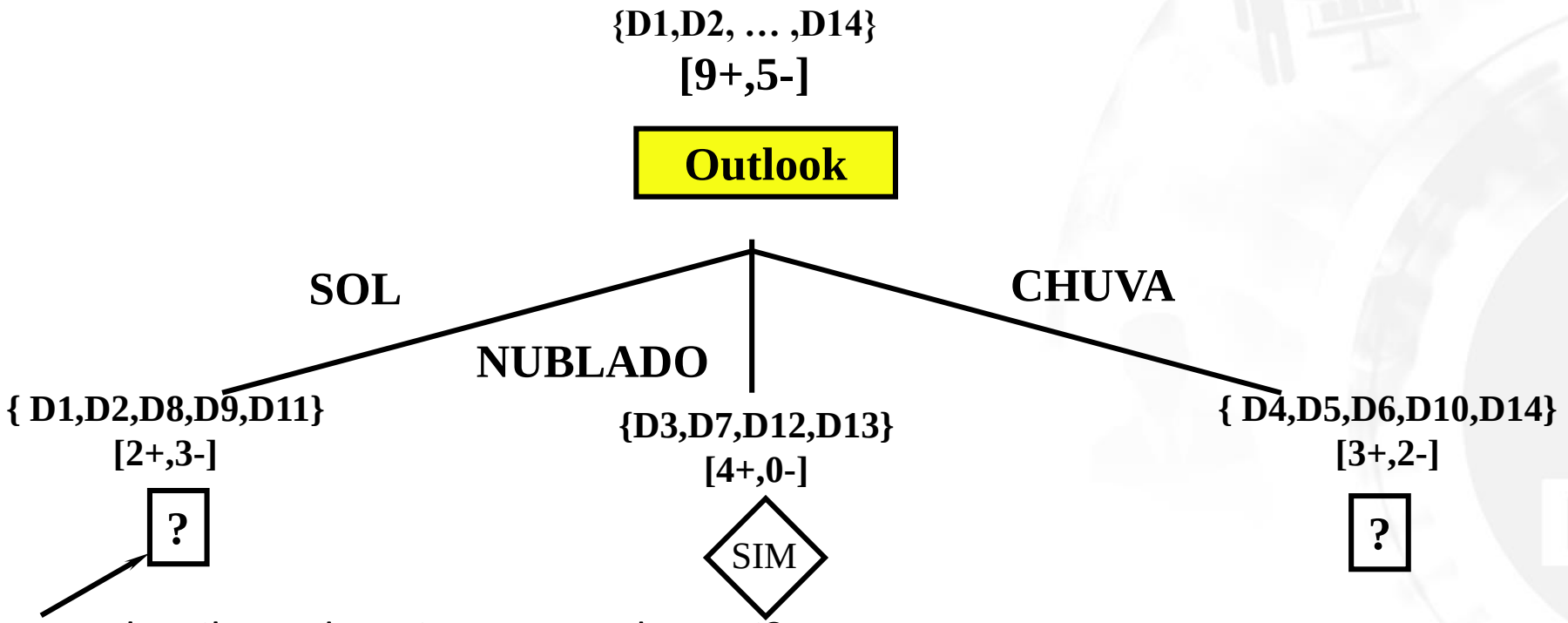
Qual atributo é o melhor classificador?



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

GANHO (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}

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

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

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

ÁRVORES DE DECISÃO

ID3 (Quinlan 1979)

- seleciona a favor de “árvores menores preferidas”

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

C4.5 – Quinlan 1993

- Usa pruning
- C5 (See5, Quinlan) - WEKA

INDICE GINI

- $Indice\ Gini = 1 - \sum_{i=1}^{C} p_i^2$

Onde p_i é a frequência relativa de cada classe em cada nó

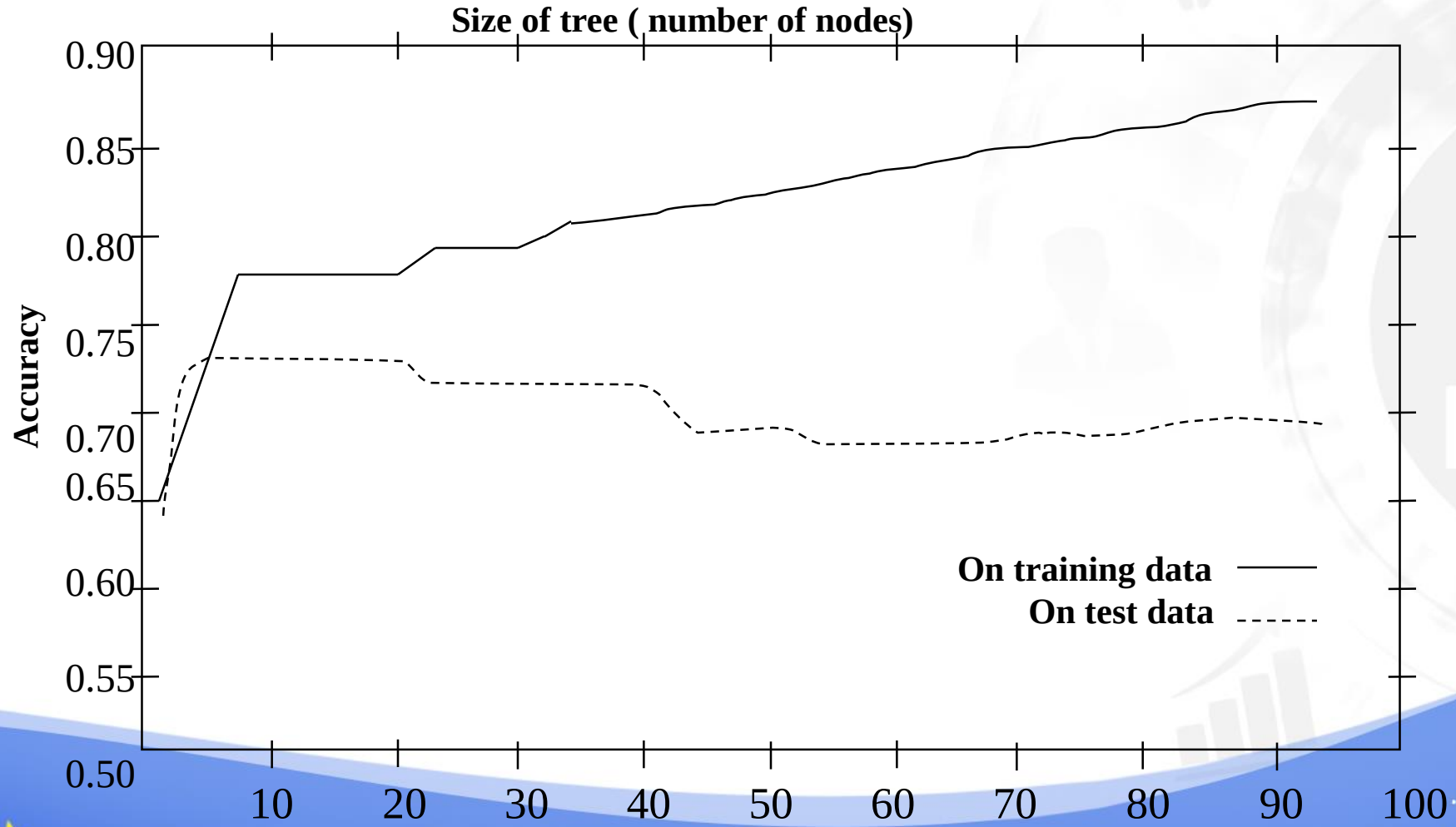
C – é o número de classes

Quando o Indice Gini é ZERO, o nó é PURO
Quando o Indice Gini é próximo de 1, o nó é IMPURO.

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 …

Overfitting in Decision Trees Learning



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)

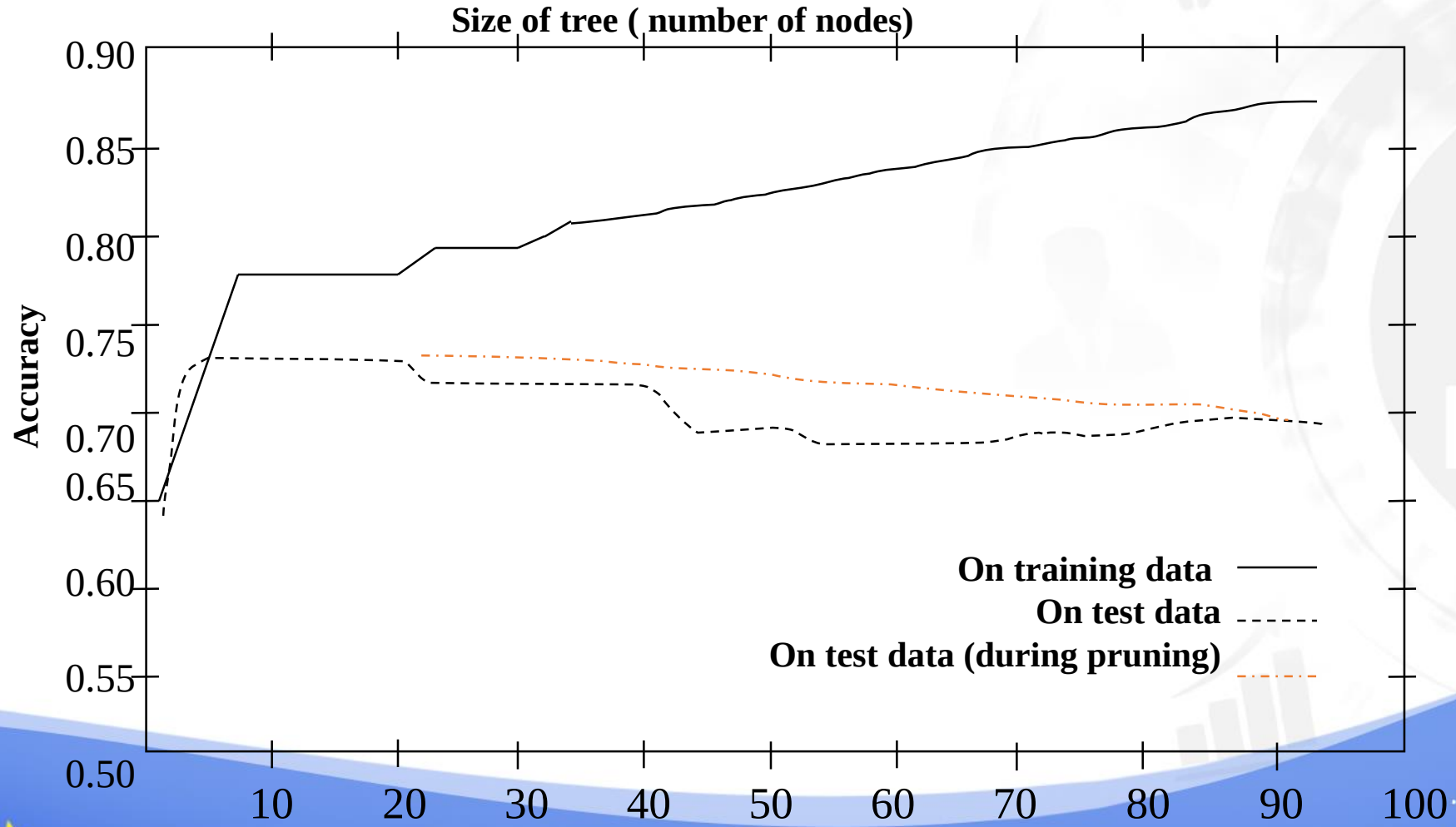
Considerar cada nó da árvore como um nó candidato a corte. Cortar um nó significa remover a sub- árvore a partir daquele nó, tornando-o um nó terminal (ou folha) e designando-o a mais comum classificação dos exemplos de treinamento afiliados com aquele nó.

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 pelo método 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 = Sol) $\wedge$ (Humidade = Alta)
  THEN PlayTennis = No

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

Atributos de valores contínuos

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

Atributos de valores contínuos

- 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?

Atributos de valores contínuos

- 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 de valores contínuos

- 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

Referencias

- **LIVRO. Mitchell, T. M., & Learning, M. (1997). Mcgraw-hill.**
-
- LIVRO. Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.
-
- LIVRO. Von Luxburg, U., & Schölkopf, B. (2011). Statistical learning theory: Models, concepts, and results. In Handbook of the History of Logic (Vol. 10, pp. 651-706). North-Holland.