

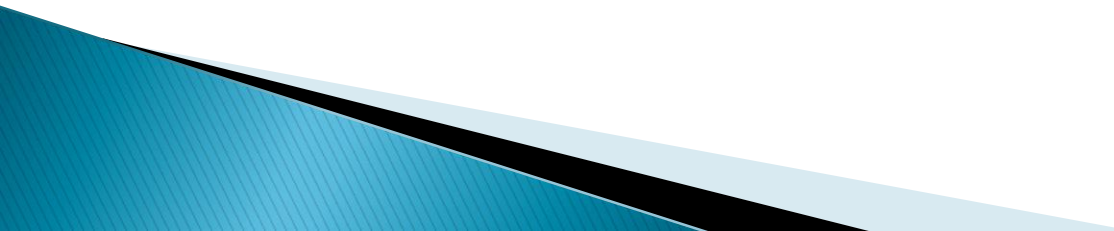
Árvores de Decisão

Aurora Trinidad Ramirez Pozo
Universidade Federal do Paraná

Árvores de Decisão

- ▶ Um dos métodos práticos mais usados
- ▶ Induz funções discretas robustas a ruído
- ▶ Capaz de aprender expressões disjuntivas
Se país = Inglaterra Ou
Se país = França e idade < 25 Então
Comprar = sim

Árvores de Decisão

- ▶ Classificação; baseado num conjunto de atributos
 - ▶ Cada nó interno corresponde a um teste sobre os valores dos atributos;
 - ▶ Os arcos são rotulados com os valores possíveis do teste;
 - ▶ Cada folha na árvore especifica a classificação.
- 

Árvores de Decisão – ADs

- ▶ Forma mais simples:
 - Lista de perguntas ⇒ respostas “sim” ou “não”
 - Hierarquicamente arranjadas
 - Levam a uma decisão
- ▶ Estrutura da árvore determinada por meio de aprendizado

ADs – treinamento

- ▶ Treinamento
 - AD encontra regras que recursivamente bifurcam (baseadas nos valores dos atributos) o conjunto de dados
 - Sub-conjuntos homogêneos intra sub-conjuntos e
 - Sub-conjuntos heterogêneos inter sub-conjuntos
- ▶ Conteúdo dos sub-conjuntos pode ser descrito por um conjunto de regras

Algoritmo básico

- ▶ ID3 (Quinlan)
- ▶ Busca top-down através do espaço de árvores de decisão possíveis
 - Que atributo deve ser testado na raiz da árvore
 - Cada atributo é testado, o melhor selecionado

Indução Top-Down

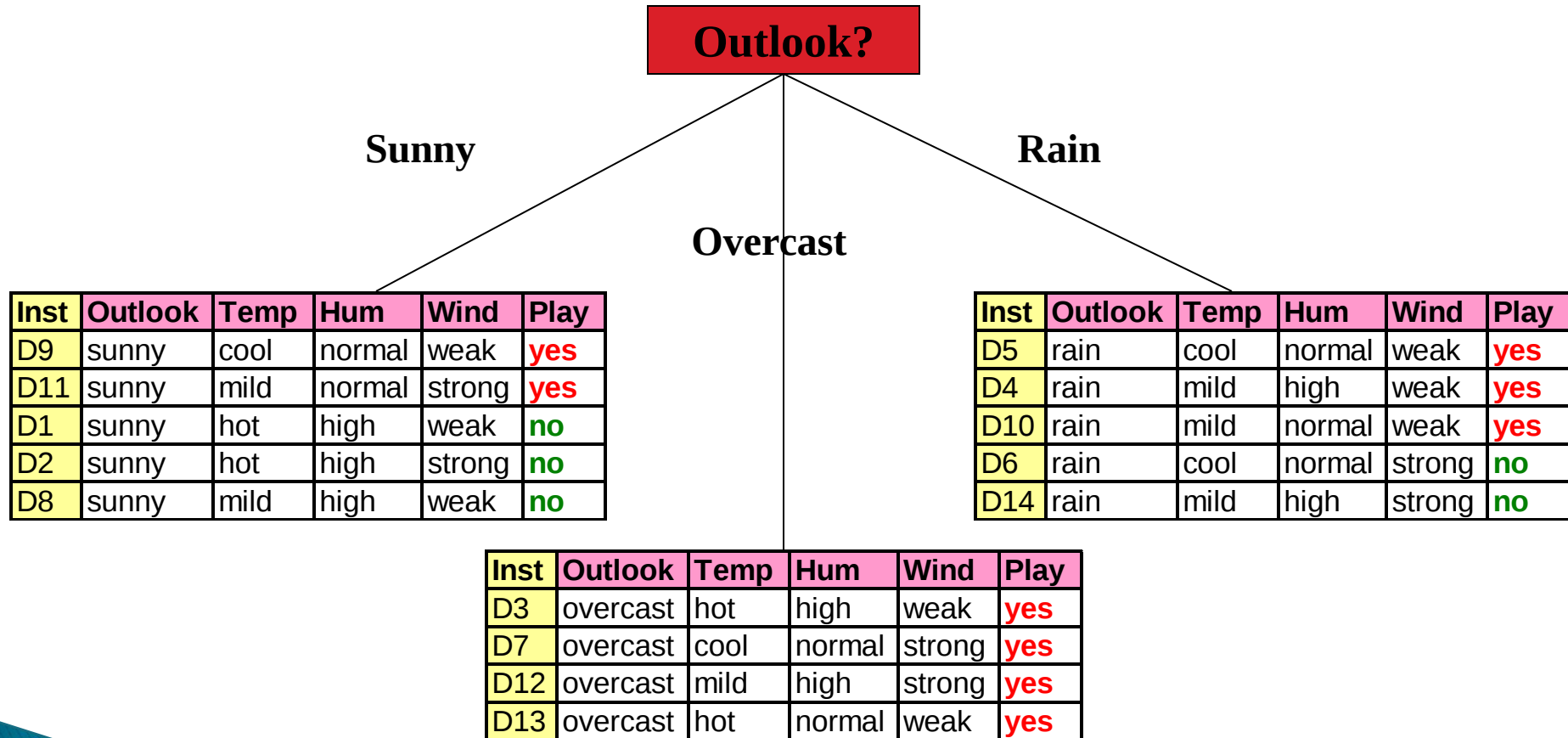
- ▶ Laço principal
 - $A \leftarrow$ o melhor atributo para o nó
 - Para cada valor de A , crie um novo descendente
 - Classifique os exemplos de treinamento segundo os valores de A
 - Se os exemplos de treinamento estão perfeitamente classificados, fim, senão volte a laço.

ADs – treinamento (2/10)

Base de Dados “Tempo”

Instância	Outlook	Temperature	Humidity	Wind	Play
D1	sunny	hot	high	weak	no
D2	sunny	hot	high	strong	no
D3	overcast	hot	high	weak	yes
D4	rain	mild	high	weak	yes
D5	rain	cool	normal	weak	yes
D6	rain	cool	normal	strong	no
D7	overcast	cool	normal	strong	yes
D8	sunny	mild	high	weak	no
D9	sunny	cool	normal	weak	yes
D10	rain	mild	normal	weak	yes
D11	sunny	mild	normal	strong	yes
D12	overcast	mild	high	strong	yes
D13	overcast	hot	normal	weak	yes
D14	rain	mild	high	strong	no

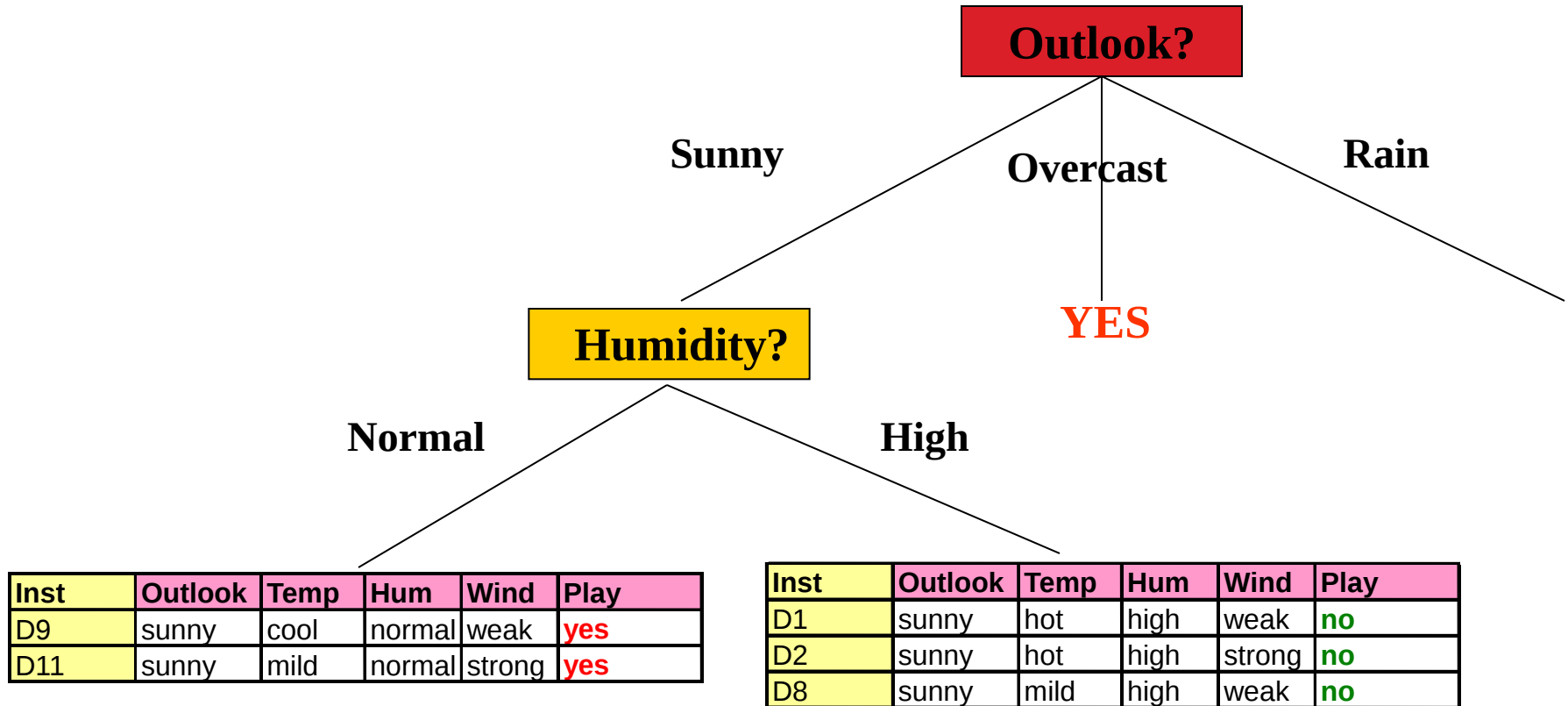
ADs – treinamento (3 / 10)



ADs – treinamento (4/10)

Teste	Exemplo	Outlook	Temperature	Humidity	Wind	Play?
If outlook=sunny	D1	Sunny	Hot	High	Weak	No
	D2	Sunny	Hot	High	Strong	No
	D8	Sunny	Mild	High	Weak	No
	D9	Sunny	Cool	Normal	Weak	Yes
	D11	Sunny	Mild	Normal	Strong	Yes
If outlook=overcast	D3	Overcast	Hot	High	Weak	Yes
	D7	Overcast	Cold	Normal	Strong	Yes
	D12	Overcast	Mild	High	Strong	Yes
	D13	Overcast	Hot	Normal	Weak	Yes
If outlook=rain	D4	Rain	Mild	High	Weak	Yes
	D5	Rain	Cool	Normal	Weak	Yes
	D6	Rain	Cool	Normal	Strong	No
	D10	Rain	Mild	Normal	Weak	Yes
	D14	Rain	Mild	High	Strong	No

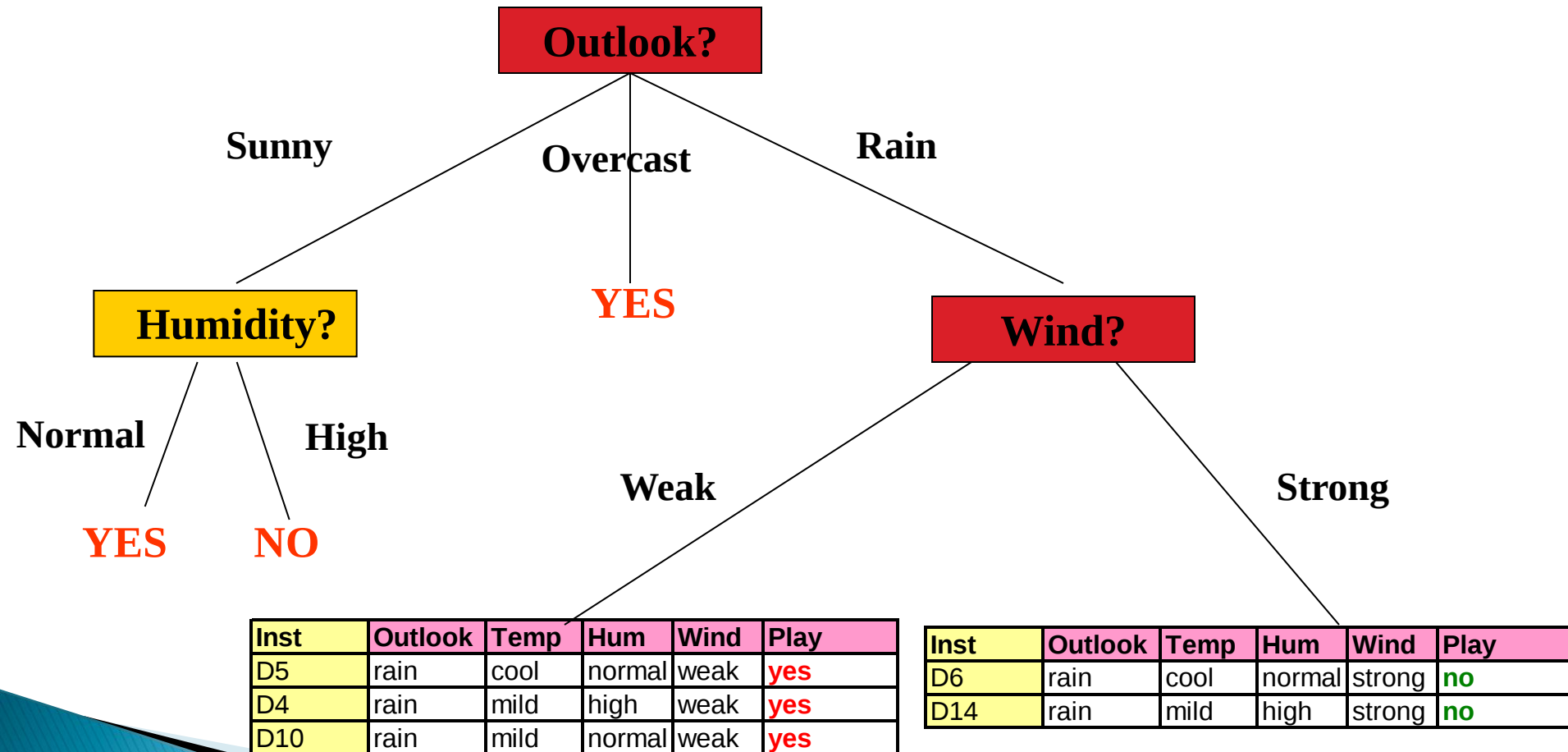
ADs – treinamento (5 / 10)



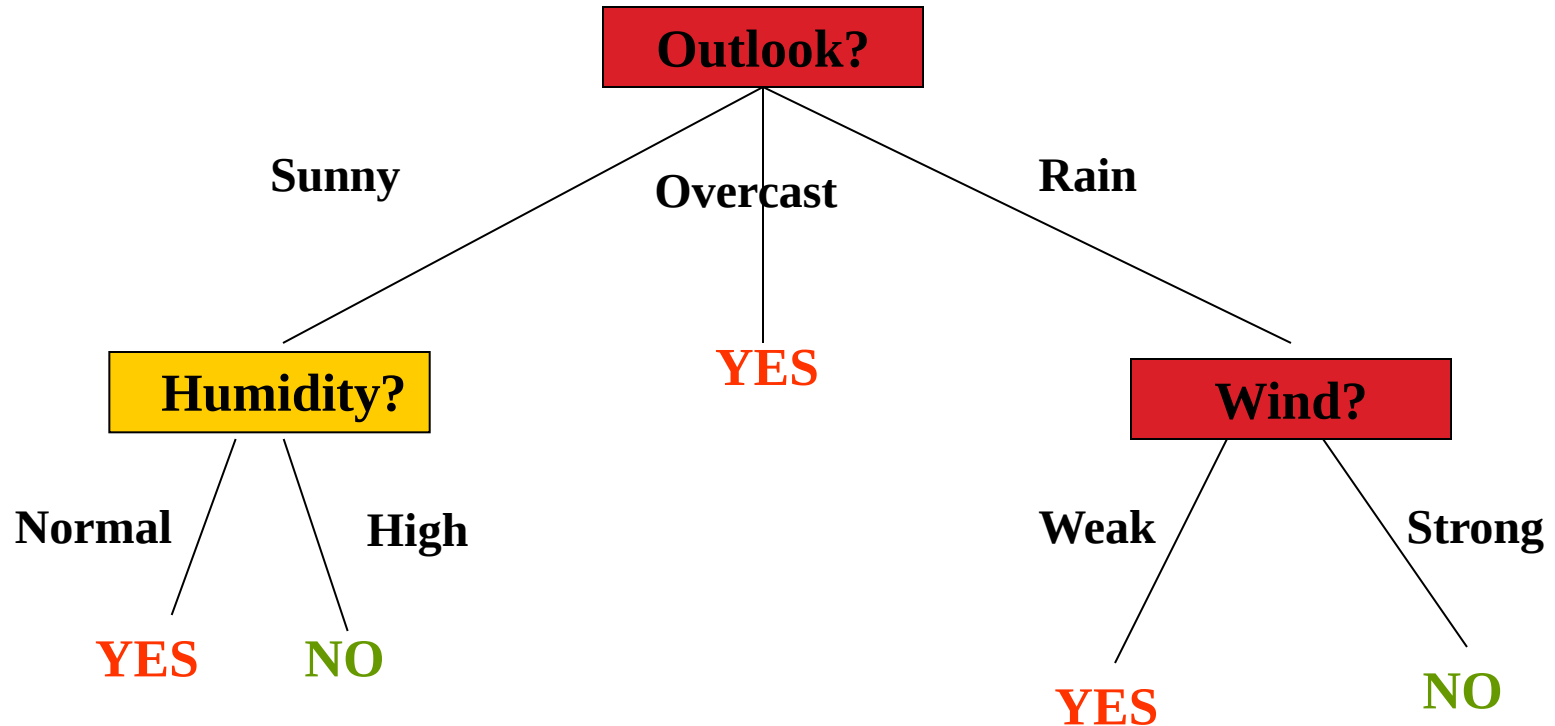
ADs – treinamento (6/10)

Teste	Exemplo	Outlook	Temperature	Humidity	Wind	Play?
If outlook=sunny and humidity=high	D1	Sunny	Hot	High	Weak	No
	D2	Sunny	Hot	High	Strong	No
	D8	Sunny	Mild	High	Weak	No
If outlook=sunny and humidity=normal	D9	Sunny	Cool	Normal	Weak	Yes
	D11	Sunny	Mild	Normal	Strong	Yes
If outlook=overcast	D3	Overcast	Hot	High	Weak	Yes
	D7	Overcast	Cold	Normal	Strong	Yes
	D12	Overcast	Mild	High	Strong	Yes
	D13	Overcast	Hot	Normal	Weak	Yes
If outlook=rain and wind=strong	D6	Rain	Cool	Normal	Strong	No
	D14	Rain	Mild	High	Strong	No
If outlook=rain and wind=weak	D4	Rain	Mild	High	Weak	Yes
	D5	Rain	Cool	Normal	Weak	Yes
	D10	Rain	Mild	Normal	Weak	Yes

ADs – treinamento (7/10)



ADs – treinamento (8/10)

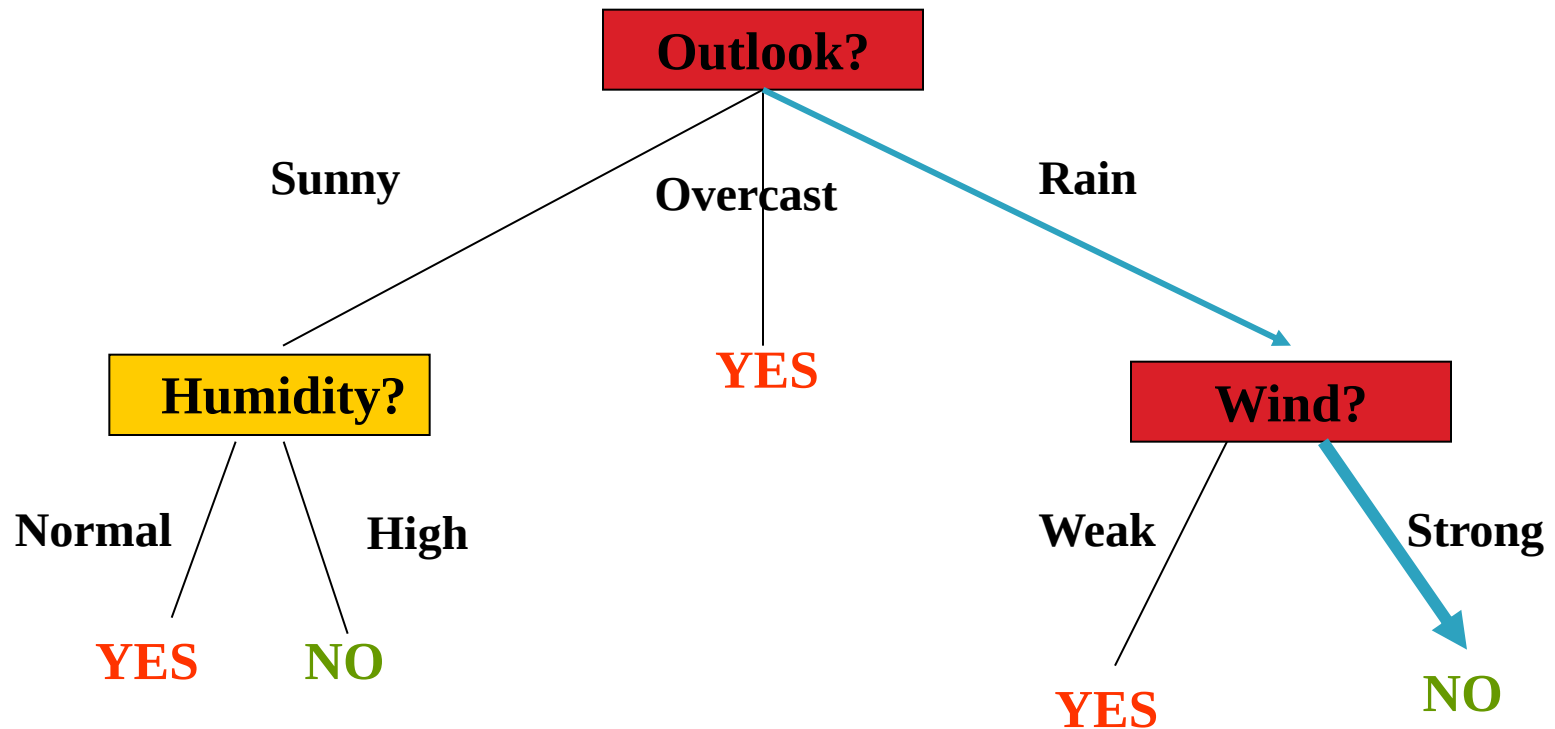


ADs – treinamento (9/10)

Teste	Exemplo	Outlook	Temperature	Humidity	Wind	Play?
If outlook=sunny and humidity=high	D1	Sunny	Hot	High	Weak	No
	D2	Sunny	Hot	High	Strong	No
	D8	Sunny	Mild	High	Weak	No
If outlook=sunny and humidity=normal	D9	Sunny	Cool	Normal	Weak	Yes
	D11	Sunny	Mild	Normal	Strong	Yes
If outlook=overcast	D3	Overcast	Hot	High	Weak	Yes
	D7	Overcast	Cold	Normal	Strong	Yes
	D12	Overcast	Mild	High	Strong	Yes
	D13	Overcast	Hot	Normal	Weak	Yes
If outlook=rain and wind=strong	D6	Rain	Cool	Normal	Strong	No
	D14	Rain	Mild	High	Strong	No
If outlook=rain and wind=weak	D4	Rain	Mild	High	Weak	Yes
	D5	Rain	Cool	Normal	Weak	Yes
	D10	Rain	Mild	Normal	Weak	Yes

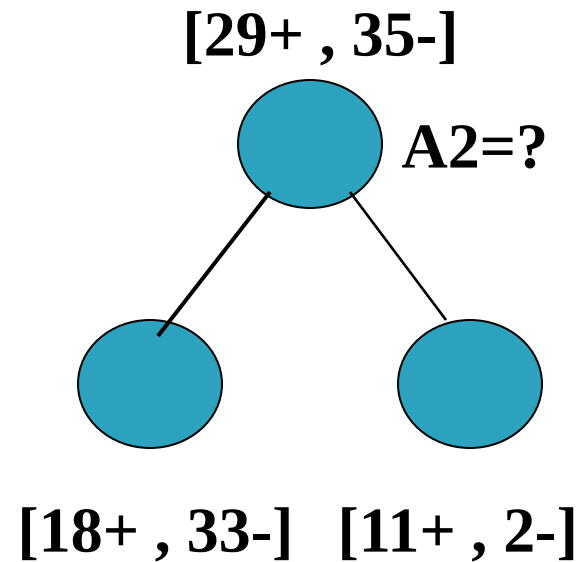
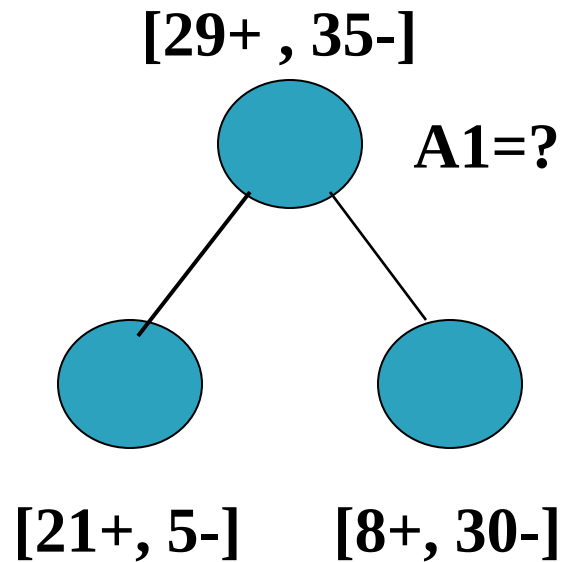
ADs – treinamento (10/10)

Como classificar <rain,hot,high,normal,strong>?

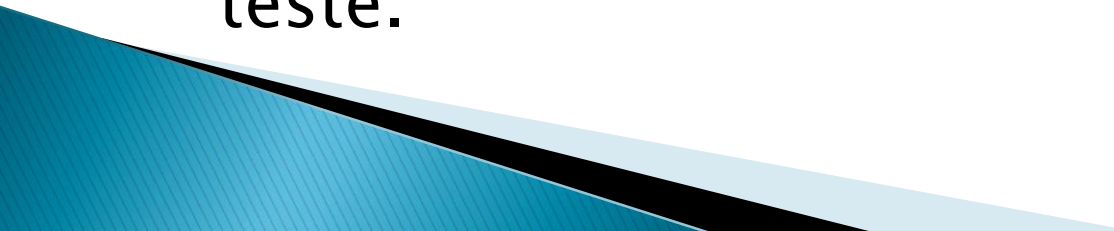


Indução *top-down* de árvores de decisão

- ▶ Qual é o melhor atributo?



Indução de Árvores

- ▶ Encontrar a árvore de decisão menor é um problema intratável;
 - ▶ Solução: Heurísticas simples, boas árvores
 - ▶ Idéia básica
 - ▶ Testar o atributo mais importante primeiro
 - ▶ Separar o maior número de casos, a cada vez.
 - ▶ Classificação correta com o menor número de teste.
- 

Indução de Árvores

- ▶ Uma árvore de decisão é construída de forma "top-down", usando o princípio de dividir-para-conquistar.
 - ▶ Inicialmente, todas as tuplas são alocadas à raiz da árvore.
 - ▶ Selecione um atributo e divida o conjunto.
 - ▶ Objetivo– separar as classes
 - ▶ Repita esse processo, recursivamente.
- 

Entropia (1 / 3)

- ▶ S é uma amostra dos exemplos de treinamento
- ▶ $p_{\oplus}$ é a proporção de exemplos positivos em S
- ▶ $p_{\ominus}$ é a proporção de exemplos negativos em S
- ▶ Entropia mede a “impureza” de S :
 - $Entropia(S) = -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$
 - Para mais de duas (n) classes

$$Entropy(S) = \sum_{i=1}^n -p(i) \log_2 p(i)$$

Entropia (2 / 3)

- ▶ Entropia(S)=especifica o nr. mínimo de bits de informação necessário para codificar uma classificação de um membro arbitrário de S.

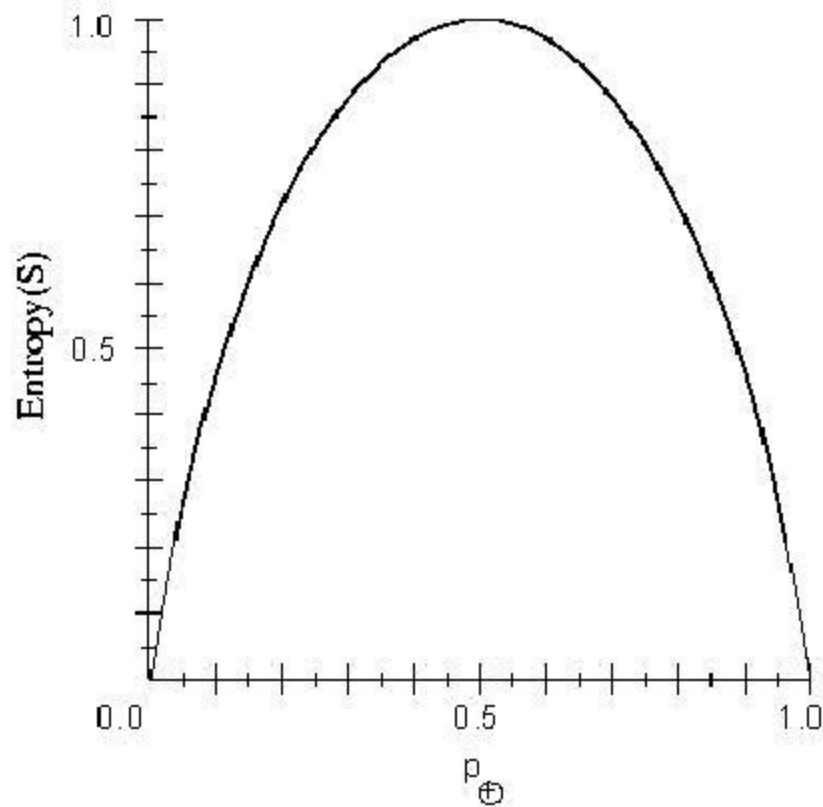
Entropia (3 / 3)

- ▶ Portanto, o nr. de bits esperados para codificar $\oplus$ ou $\ominus$ de elementos aleatórios de S :
 - $p_{\oplus}(-\log_2 p_{\oplus}) + p_{\ominus}(-\log_2 p_{\ominus})$
 - $Entropia(S) = -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$

Entropia – Exemplo I

- ▶ Se $p_{\oplus}$ é 1, o destinatário sabe que o exemplo selecionado será positivo
 - Nenhuma mensagem precisa ser enviada
 - Entropia é 0 (mínima)
- ▶ Se $p_{\oplus}$ é 0.5, um bit é necessário para indicar se o exemplo selecionado é $\oplus$ ou $\ominus$
 - Entropia é 1 (máxima)
- ▶ Se $p_{\oplus}$ é 0.8, então uma coleção de mensagens podem ser codificadas usando-se – em média menos de um bit – códigos mais curtos para $\oplus$ e mais longos para $\ominus$

Entropia – Gráfico



Entropia – Exemplo I

- ▶ Suponha que S é uma coleção de 14 exemplos, incluindo 9 positivos e 5 negativos (o exemplo da base de dados “Tempo”)
 - Notação: $[9+, 5-]$
- ▶ A entropia de S em relação a esta classificação booleana é dada por:

$$\begin{aligned} \text{Entropy}([9+, 5-]) &= -(9/14) \log_2(9/14) - (5/14) \log_2(5/14) \\ &= 0.940 \end{aligned}$$

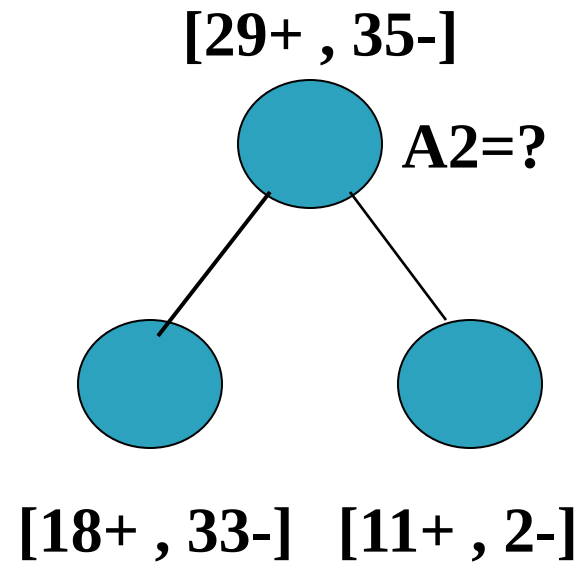
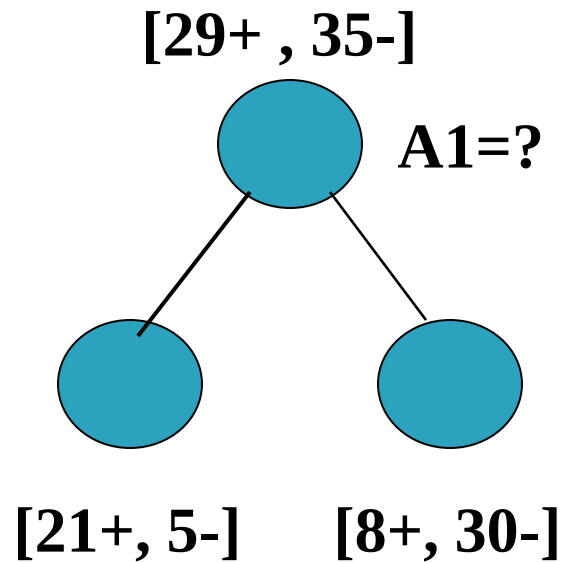
Critério de ganho (1 / 2)

$Gain(S, A)$ = redução esperada da entropia devido a “classificação” de acordo com A

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

Critério de ganho (2 / 2)

- ▶ Usar o critério de ganho para decidir!



Critério de ganho – Exemplo (1 / 2)

- Suponha que S é uma coleção de exemplos de treinamento ($[9+, 5-]$) descritos por atributos incluindo *Wind*, que pode ter os valores *Weak* and *Strong* (base d dados “Tempo”).

$$\text{Values}(\text{Wind}) = \text{Weak}, \text{Strong}$$

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

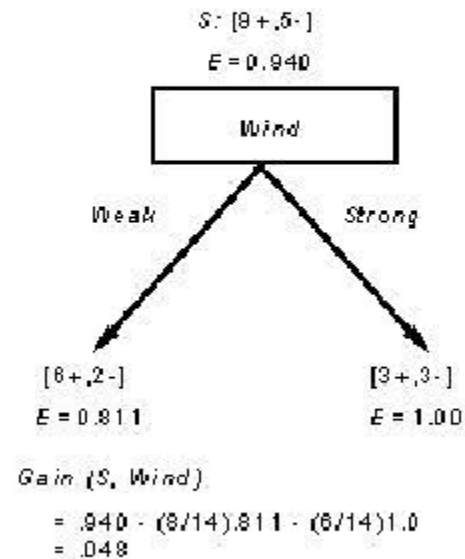
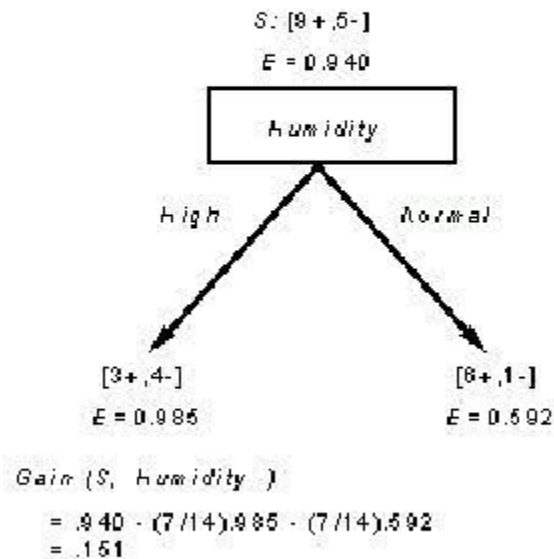
$$S_{\text{Weak}} = [6+, 2-]$$

$$S_{\text{Strong}} = [3+, 3-]$$

$$\begin{aligned} \text{Gain}(S, \text{Wind}) &= \text{Entropy}(S) - \sum_{v \in \{\text{Weak}, \text{Strong}\}} \frac{|S_v|}{|S|} \text{Entropy}(S_v) \\ &= \text{Entropy}(S) - (8/14) \text{Entropy}(S_{\text{Weak}}) - (6/14) \text{Entropy}(S_{\text{Strong}}) \\ &= 0.940 - (8/14) * 0.811 - (6/14) * 1.00 \\ &= 0.048 \end{aligned}$$

Critério de ganho – Exemplo (2/2)

Which attribute is the best classifier?

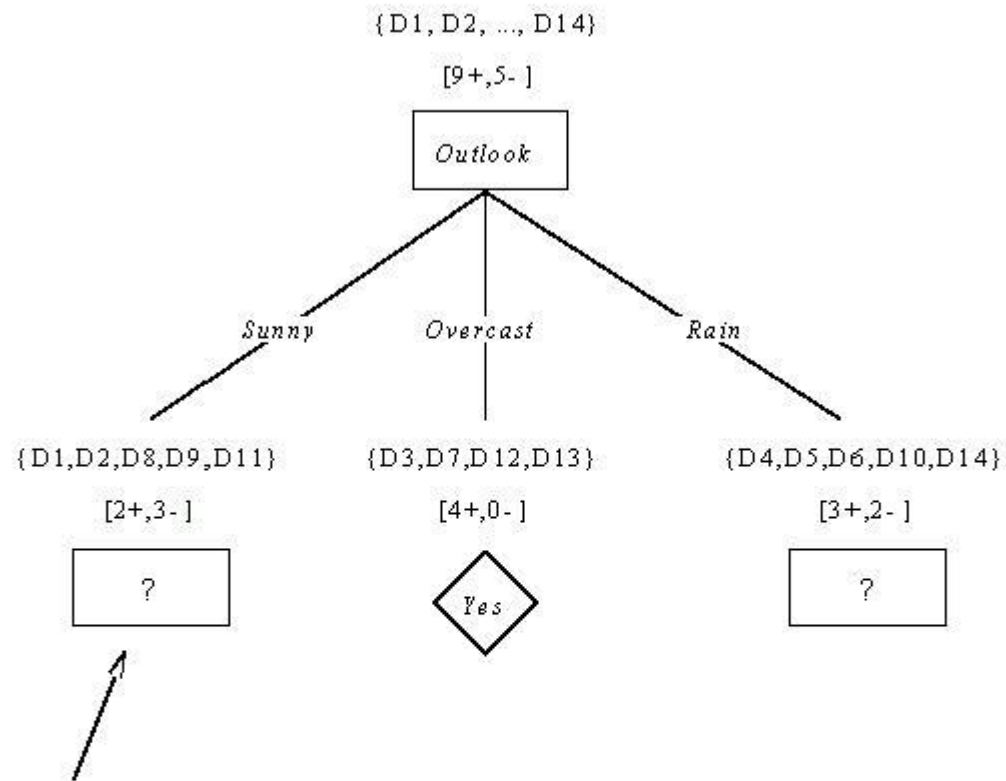


Exemplos de treinamento (1 / 3)

Instância	Outlook	Temperature	Humidity	Wind	Play
D1	sunny	hot	high	weak	no
D2	sunny	hot	high	strong	no
D3	overcast	hot	high	weak	yes
D4	rain	mild	high	weak	yes
D5	rain	cool	normal	weak	yes
D6	rain	cool	normal	strong	no
D7	overcast	cool	normal	strong	yes
D8	sunny	mild	high	weak	no
D9	sunny	cool	normal	weak	yes
D10	rain	mild	normal	weak	yes
D11	sunny	mild	normal	strong	yes
D12	overcast	mild	high	strong	yes
D13	overcast	hot	normal	weak	yes
D14	rain	mild	high	strong	no

Exemplos de treinamento (2 / 3)

- ▶ Que atributo deve ser selecionado para ser a raiz da árvore?
 - $Gain(S, Outlook) = 0.246$
 - $Gain(S, Humidity) = 0.151$
 - $Gain(S, Wind) = 0.048$
 - $Gain(S, Temperature) = 0.029$
- ▶ em que S denota a coleção de exemplos na tabela anterior



$$S_{\text{sunny}} = \{D1,D2,D8,D9,D11\}$$

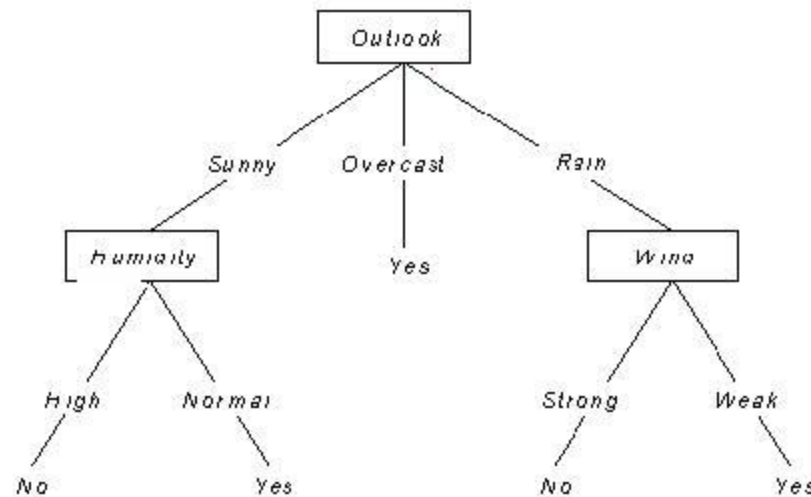
$$\text{Gain}(S_{\text{sunny}} . \text{Humidity}) = .970 - (3/5) 0.0 - (2/5) 0.0 = .970$$

$$\text{Gain}(S_{\text{sunny}} . \text{Temperature}) = .970 - (2/5) 0.0 - (2/5) 1.0 - (1/5) 0.0 = .570$$

$$\text{Gain}(S_{\text{sunny}} . \text{Wind}) = .970 - (2/5) 1.0 - (3/5) .918 = .019$$

Overfitting em árvores de decisão

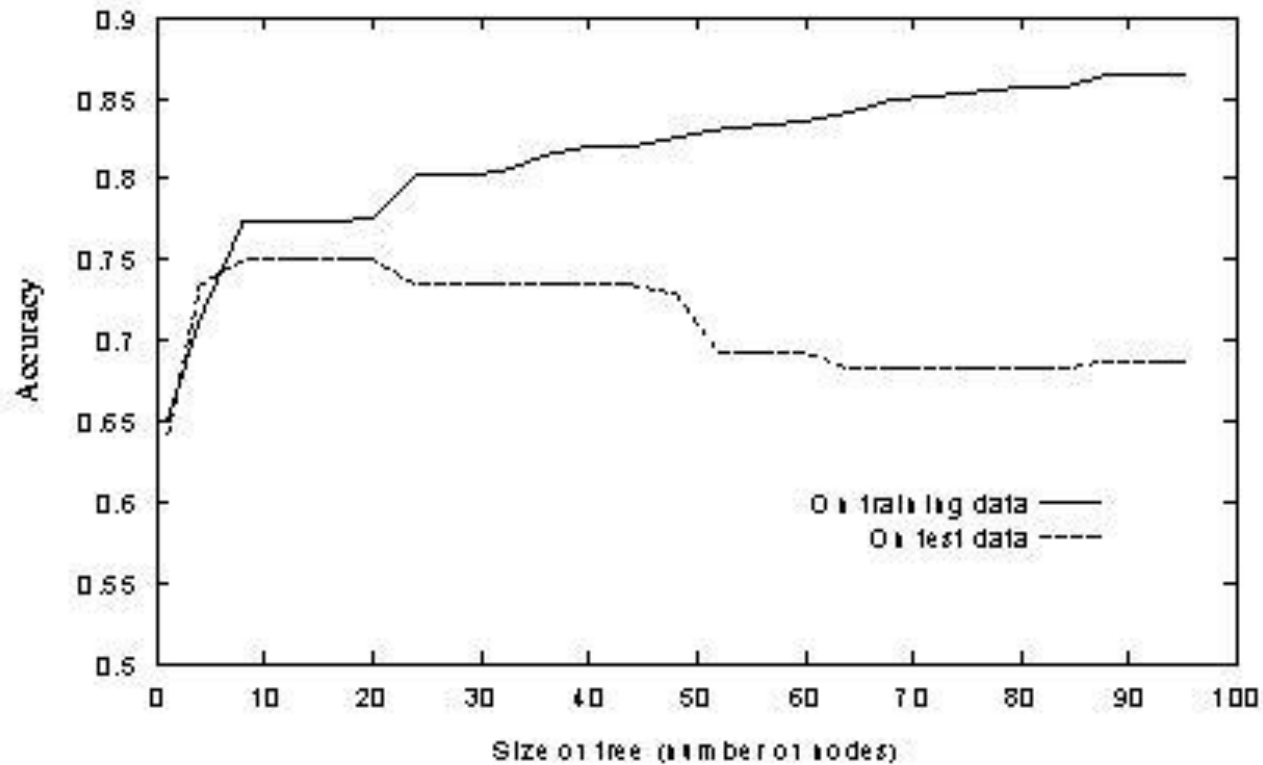
- ▶ Suponha que adicionamos um exemplo de treinamento com ruído #15:
 - *Sunny, Hot, Normal, Strong, PlayTennis=No*
 - Qual seria seu efeito na árvore anterior:



Overfitting

- ▶ Considere o erro da hipótese h sobre
 - dados do treinamento: $erro_{train}(h)$
 - distribuição completa D dos dados: $erro_D(h)$
- ▶ Uma hipótese $h \in H$ “overfits” (se super ajusta) o conjunto de dados se há uma hipótese alternativa $h' \in H$ tal que
 - $erro_{train}(h) < erro_{train}(h')$ e
 - $erro_D(h) > erro_D(h')$

Overfitting na aprendizagem de árvores de decisão (1 / 2)



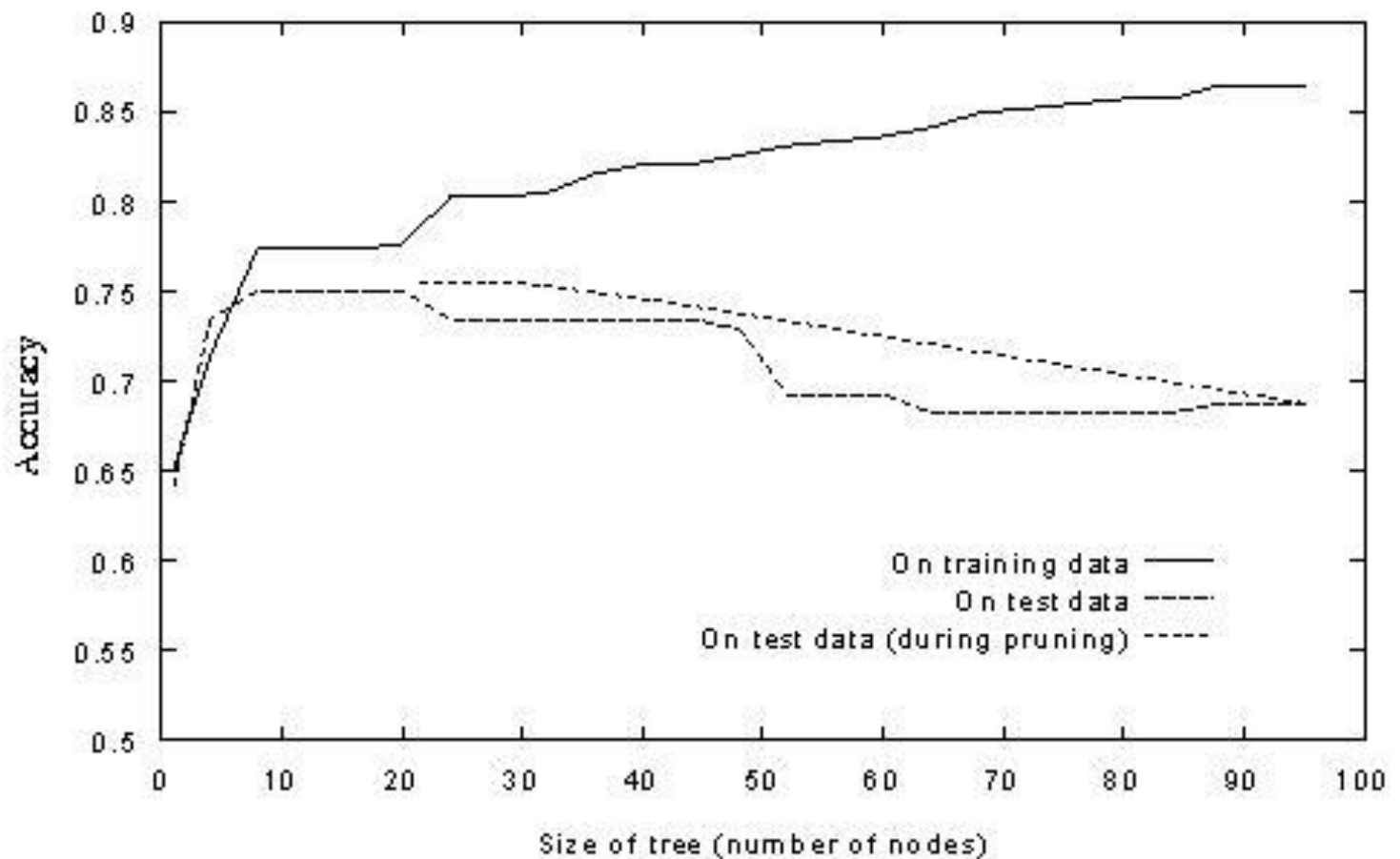
Overfitting na aprendizagem de árvores de decisão (2 / 2)

- ▶ Como evitar overfitting?
 - Parar o crescimento da árvore quando a divisão dos dados não é estatisticamente significativa
 - Deixar a árvore crescer completamente para, então, poda-la (post-prune)
- ▶ Como escolher a “melhor” árvore:
 - Medir a performance sobre o cj. de dados treinamento
 - Medir a performance sobre um cj. de dados (separado) de validação
 - MDL: minimizar –
 $size(tree) + size(misclassification(tree))$

Reduzindo o erro através da poda

- ▶ Divida os dados em *cj. de treinamento* e *validação*
- ▶ Faça até que uma poda adicional seja prejudicial:
 - 1. Avalie o impacto sobre o *cj. de validação* de podar cada nó possível (e seus descendentes)
 - 2. Gulosamente remova aquele que melhora mais a performance no *cj. de validação*

Efeito do uso da técnica de poda



Podando as regras

- ▶ 1. Converta a árvore para um conjunto equivalente de regras
- ▶ 2. Pode cada regra independentemente umas das outras
- ▶ 3. Ordene as regras finais em uma seqüência deseja para uso
 - Estes é um dos métodos mais utilizados (e.g., C4.5)

Atributos com valores contínuos

- ▶ Crie um atributo discreto para testar um que seja contínuo
 - $Temperature = 82.5$
 - $(Temperature > 72.3) = t, f$

<i>Temperature:</i>	40	48	60	72	80	90
<i>PlayTennis:</i>	NO	NO	YES	YES	YES	NO

Atributos Numéricos: Exemplo

Amostra	Sensor1	Sensor2	Classe
1	96,37	15,01	Normal
2	59,58	11,32	Normal
3	73,52	11,05	Normal
4	92,94	14,08	Normal
5	87,26	13,60	Normal
6	99,30	15,83	Normal
7	99,68	15,94	Normal
8	65,19	12,28	Normal
9	77,83	10,27	Normal
10	98,65	15,65	Normal
11	46,11	13,55	Adulterada
12	79,66	16,67	Adulterada
13	56,51	14,38	Adulterada
14	51,60	15,77	Adulterada
15	77,20	16,77	Adulterada
16	50,57	13,78	Adulterada
17	77,73	13,41	Adulterada
18	58,10	13,68	Adulterada
19	75,01	15,70	Adulterada
20	65,10	15,58	Adulterada

Atributos Numéricos: Exemplo

Amostra	sensor1	Classe
7	99,68	Normal
6	99,30	Normal
10	98,65	Normal
1	96,37	Normal
4	92,94	Normal
5	87,26	Normal
12	79,66	Adulterada
9	77,83	Normal
17	77,73	Adulterada
15	77,20	Adulterada
19	75,01	Adulterada
3	73,52	Normal
8	65,19	Normal
20	65,10	Adulterada
2	59,58	Normal
18	58,10	Adulterada
13	56,51	Adulterada
14	51,60	Adulterada
16	50,57	Adulterada
11	46,11	Adulterada

Amostra	Sensor2	Classe
15	16,77	Adulterada
12	16,67	Adulterada
7	15,94	Normal
6	15,83	Normal
14	15,77	Adulterada
19	15,70	Adulterada
10	15,65	Normal
20	15,58	Adulterada
1	15,01	Normal
13	14,38	Adulterada
4	14,08	Normal
16	13,78	Adulterada
18	13,68	Adulterada
5	13,60	Normal
11	13,55	Adulterada
17	13,41	Adulterada
8	12,28	Normal
2	11,32	Normal
3	11,05	Normal
9	10,27	Normal

Atributos Numéricos: Exemplo

Amostra	sensor1	Classe
7	99,68	Normal
6	99,30	Normal
10	98,65	Normal
1	96,37	Normal
4	92,94	Normal
5	87,26	Normal

12	79,66	Adulterada
9	77,83	Normal
17	77,73	Adulterada
15	77,20	Adulterada
19	75,01	Adulterada
3	73,52	Normal
8	65,19	Normal
20	65,10	Adulterada
2	59,58	Normal
18	58,10	Adulterada
13	56,51	Adulterada
14	51,60	Adulterada
16	50,57	Adulterada
11	46,11	Adulterada

Atributos Numéricos: Exemplo

Amostra	sensor1	Classe
7	99,68	Normal
6	99,30	Normal
10	98,65	Normal
1	96,37	Normal
4	92,94	Normal
5	87,26	Normal
12	79,66	Adulterada
9	77,83	Normal

17	77,73	Adulterada
15	77,20	Adulterada
19	75,01	Adulterada
3	73,52	Normal
8	65,19	Normal
20	65,10	Adulterada
2	59,58	Normal
18	58,10	Adulterada
13	56,51	Adulterada
14	51,60	Adulterada
16	50,57	Adulterada
11	46,11	Adulterada

Atributos Numéricos: Exemplo

Sensor_1?

> 83,46

<= 83,46

Amostra	Sensor1	Sensor2	Classe
7	99,68	15,94	Normal
6	99,30	15,83	Normal
10	98,65	15,65	Normal
1	96,37	15,01	Normal
4	92,94	14,08	Normal
5	87,26	13,60	Normal

Amostra	Sensor1	Sensor2	Classe
12	79,66	16,67	Adulterada
9	77,83	10,27	Normal
17	77,73	13,41	Adulterada
15	77,20	16,77	Adulterada
19	75,01	15,70	Adulterada
3	73,52	11,05	Normal
8	65,19	12,28	Normal
20	65,10	15,58	Adulterada
2	59,58	11,32	Normal
18	58,10	13,68	Adulterada
13	56,51	14,38	Adulterada
14	51,60	15,77	Adulterada
16	50,57	13,78	Adulterada
11	46,11	13,55	Adulterada

Atributos Numéricos: Exemplo

Amostra	Sensor1	Classe
12	79,66	Adulterada
9	77,83	Normal
17	77,73	Adulterada
15	77,20	Adulterada
19	75,01	Adulterada
3	73,52	Normal
8	65,19	Normal
20	65,10	Adulterada
2	59,58	Normal
18	58,10	Adulterada
13	56,51	Adulterada
14	51,60	Adulterada
16	50,57	Adulterada
11	46,11	Adulterada

Amostra	Sensor2	Classe
15	16,77	Adulterada
12	16,67	Adulterada
14	15,77	Adulterada
19	15,70	Adulterada
20	15,58	Adulterada
13	14,38	Adulterada
16	13,78	Adulterada
18	13,68	Adulterada
11	13,55	Adulterada
17	13,41	Adulterada
8	12,28	Normal
2	11,32	Normal
3	11,05	Normal
9	10,27	Normal

Atributos Numéricos: Exemplo

Sensor_1?

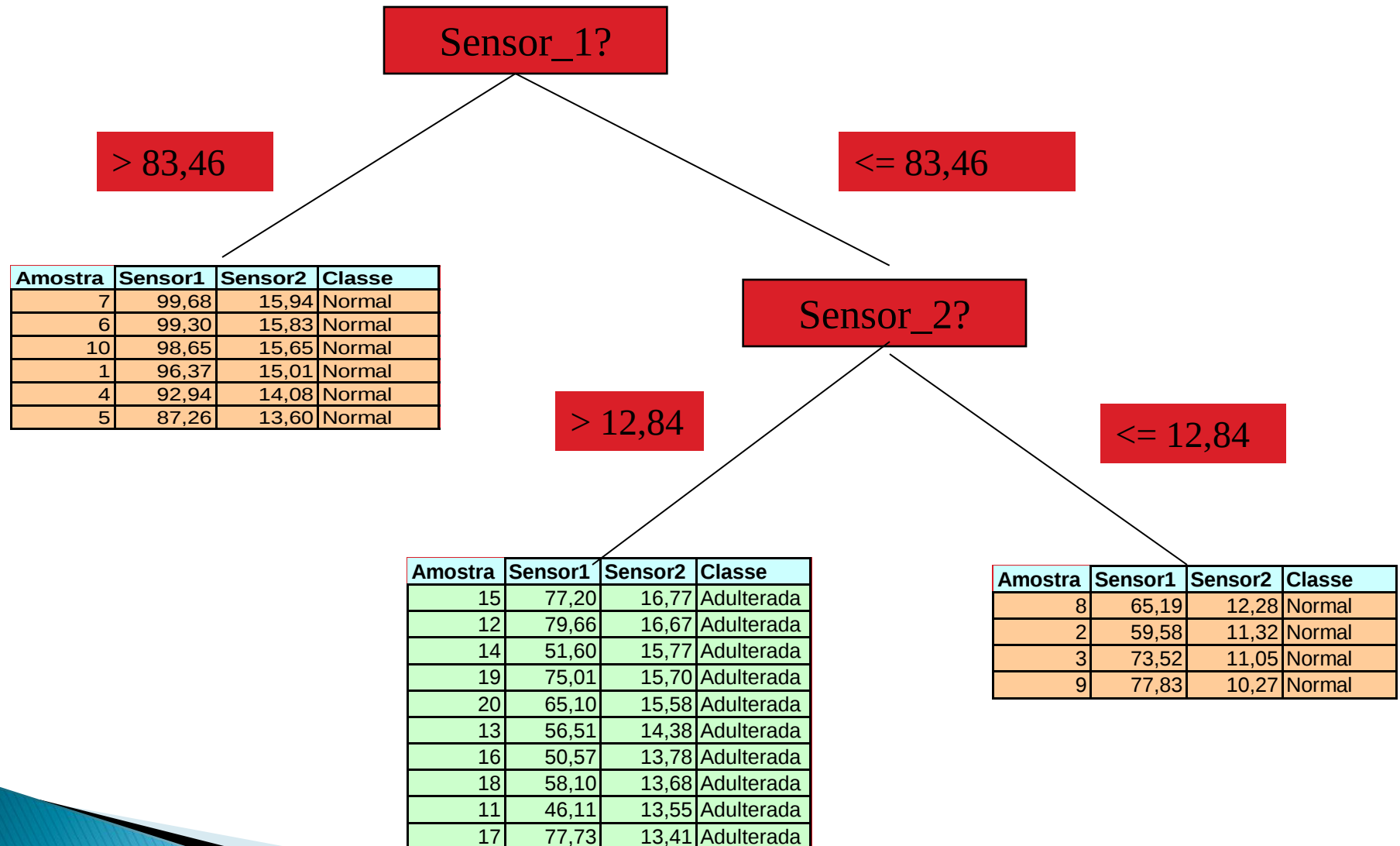
> 83,46

<= 83,46

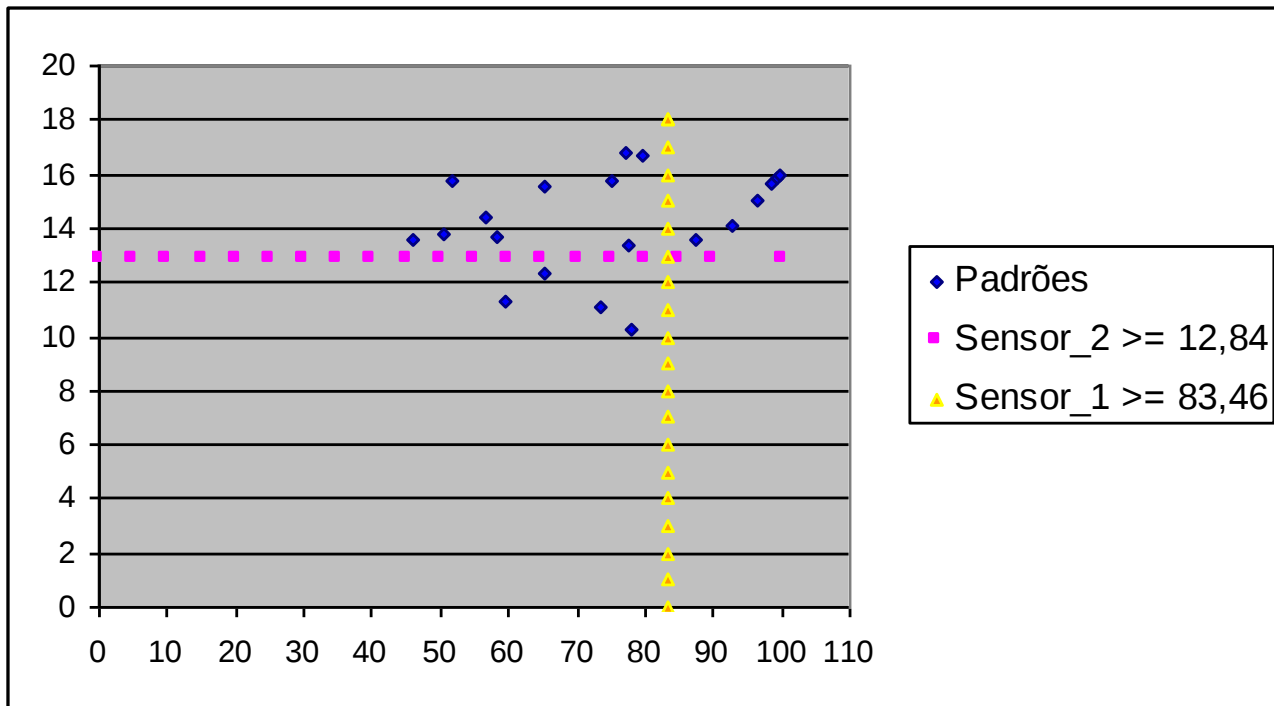
Amostra	Sensor1	Sensor2	Classe
7	99,68	15,94	Normal
6	99,30	15,83	Normal
10	98,65	15,65	Normal
1	96,37	15,01	Normal
4	92,94	14,08	Normal
5	87,26	13,60	Normal

Amostra	Sensor1	Sensor2	Classe
12	79,66	16,67	Adulterada
9	77,83	10,27	Normal
17	77,73	13,41	Adulterada
15	77,20	16,77	Adulterada
19	75,01	15,70	Adulterada
3	73,52	11,05	Normal
8	65,19	12,28	Normal
20	65,10	15,58	Adulterada
2	59,58	11,32	Normal
18	58,10	13,68	Adulterada
13	56,51	14,38	Adulterada
14	51,60	15,77	Adulterada
16	50,57	13,78	Adulterada
11	46,11	13,55	Adulterada

Atributos Numéricos: Exemplo



Atributos Numéricos: Exemplo



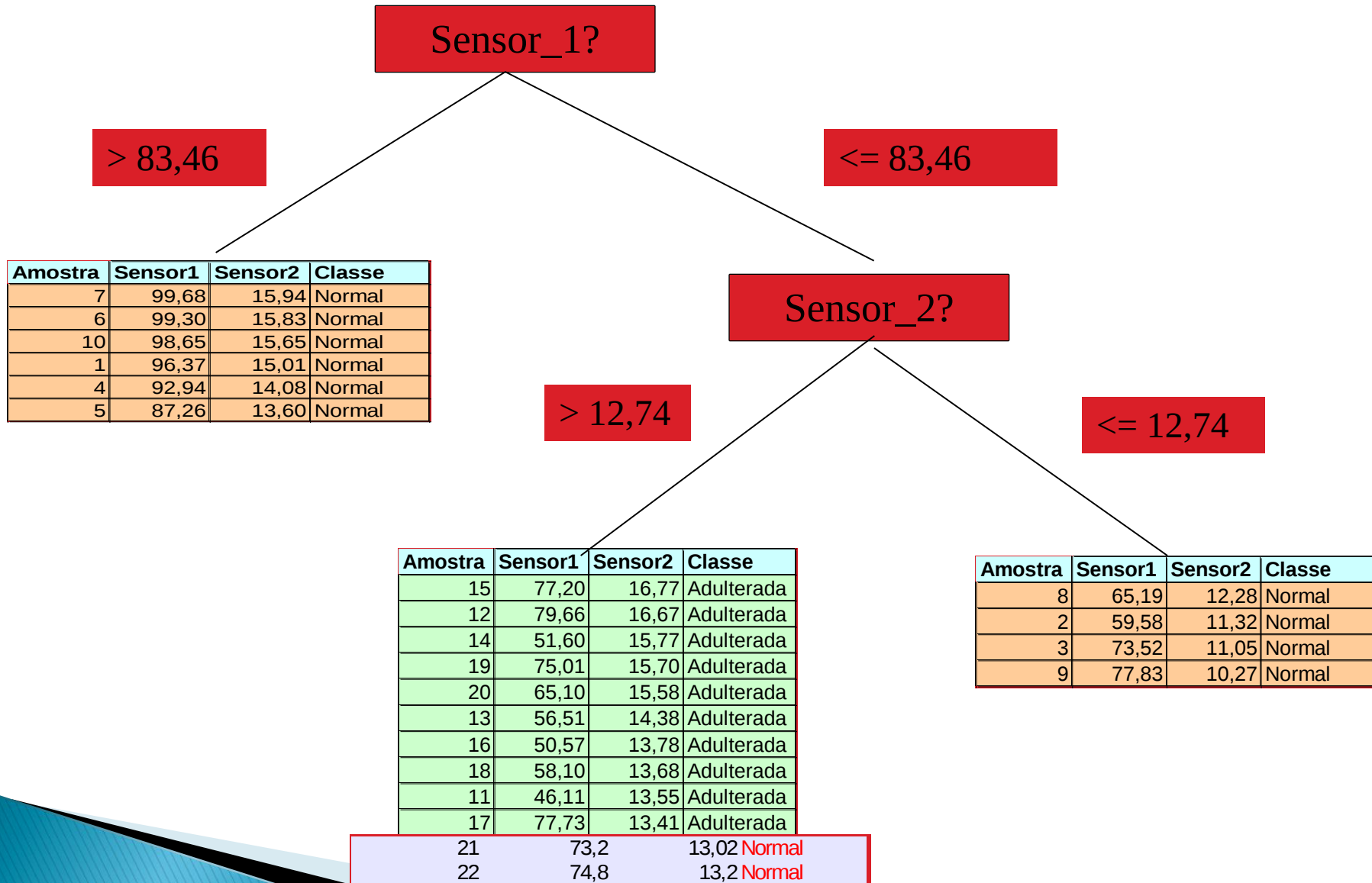
Atributos Numéricos: Exemplo (Ruído)

Amostra	Sensor1	Sensor2	Classe
1	96,37	15,01	Normal
2	59,58	11,32	Normal
3	73,52	11,05	Normal
4	92,94	14,08	Normal
5	87,26	13,60	Normal
6	99,30	15,83	Normal
7	99,68	15,94	Normal
8	65,19	12,28	Normal
9	77,83	10,27	Normal
10	98,65	15,65	Normal
11	46,11	13,55	Adulterada
12	79,66	16,67	Adulterada
13	56,51	14,38	Adulterada
14	51,60	15,77	Adulterada
15	77,20	16,77	Adulterada
16	50,57	13,78	Adulterada
17	77,73	13,41	Adulterada
18	58,10	13,68	Adulterada
19	75,01	15,70	Adulterada
20	65,10	15,58	Adulterada

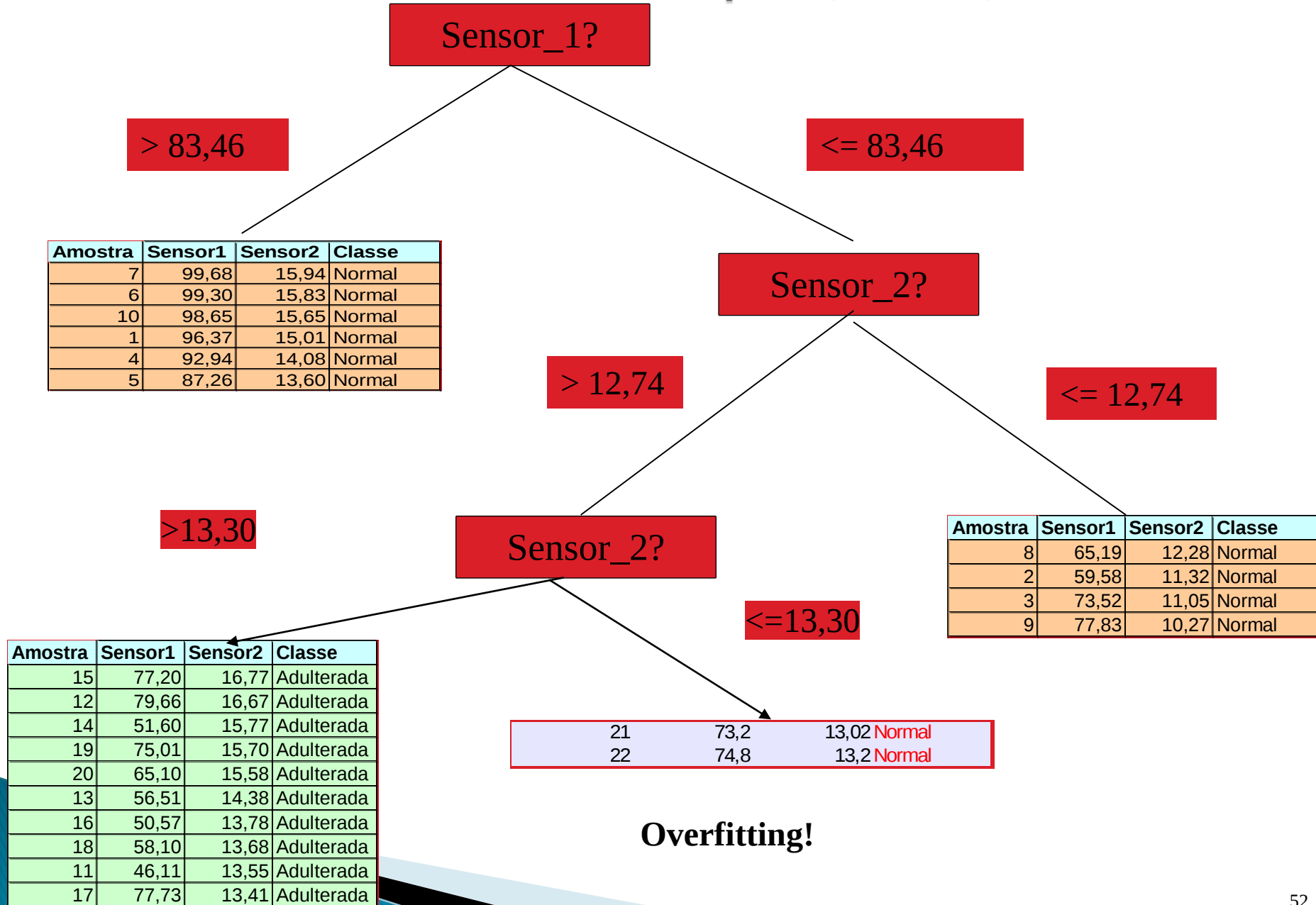
21 73,2 13,02 Normal

22 74,8 13,2 Normal

Atributos Numéricos: Exemplo (Ruído)



Atributos Numéricos: Exemplo (Ruído)



Atributos com vários valores (1 / 2)

- ▶ Problema:
 - Se o atributo tem vários valores, *Gain* o selecionará
 - Suponha o uso de *Date = 3/06/00* como atributo
- ▶ Um abordagem: use *GainRatio*

Atributos com vários valores (2 / 2)

$$GainRatio(S, A) \equiv \frac{Gain(S, A)}{SplitInformation(S, A)}$$
$$SplitInformation(S, A) \equiv - \sum_{i=1}^c \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

Onde S_i é um subconjunto de S para o qual A tem valor v_i

Atributos com custo

- ▶ Considere
 - Diagnóstico médico, *BloodTest* tem custo R\$500
 - Robótica, *Width_from_1ft* tem custo 23 sec.
- ▶ Como aprender uma árvore consistente com um valor de custo esperado baixo?
- ▶ Uma abordagem: substitua *Gain* por:

Atributos com custo (2 / 2)

Tan e Schlimmer (1990)

$$\frac{Gain^2(S, A)}{Cost(S, A)}$$

Nunez (1988)

$$\frac{2^{Gain(S, A)}}{(Cost(A) + 1)^w}$$

Onde $w \in [0, 1]$ determinar a importância do custo

Valores de atributo desconhecidos (1 / 2)

- ▶ E se valores do atributo A estão faltando para alguns exemplos?
- ▶ Mesmo assim use os exemplos de treinamento, e organize a árvore como segue:
 - Se um nó n testa A, atribua um valor para A que seja o mais comum entre os outros exemplos classificados no nó n
 - Atribua para A um valor que seja o mais comum entre os outros exemplos com o mesmo valor objetivo (*target value*)

Valores de atributo desconhecidos (2/2)

- Atribua uma probabilidade p_i para cada valor possível v_i de **A**
 - atribua uma fração p_i de exemplos para cada descendente da árvore
- Classifique exemplos novos da mesma maneira

ADs – conclusão

- ▶ Vantagens:
 - Estrutura de fácil manipulação
 - Produzem modelos que podem ser facilmente interpretados por humanos
- ▶ Desvantagens:
 - Pouca robustez a dados de grande dimensão
 - Acurácia afetada por atributos pouco relevantes
 - Dificuldade em lidar com dados contínuos

Exemplo

Sexo	País	Idade	Compra
M	França	25	Sim
M	Inglaterra	21	Sim
F	França	23	Sim
F	Inglaterra	34	Sim
F	França	30	Não
M	Alemanha	21	Não
M	Alemanha	20	Não
F	Alemanha	18	Não
F	França	34	Não
M	França	55	Não