END3900 – DATA MINING PROBLEM SESSION-2

Res. Asst. Eyüp Ensar IŞIK

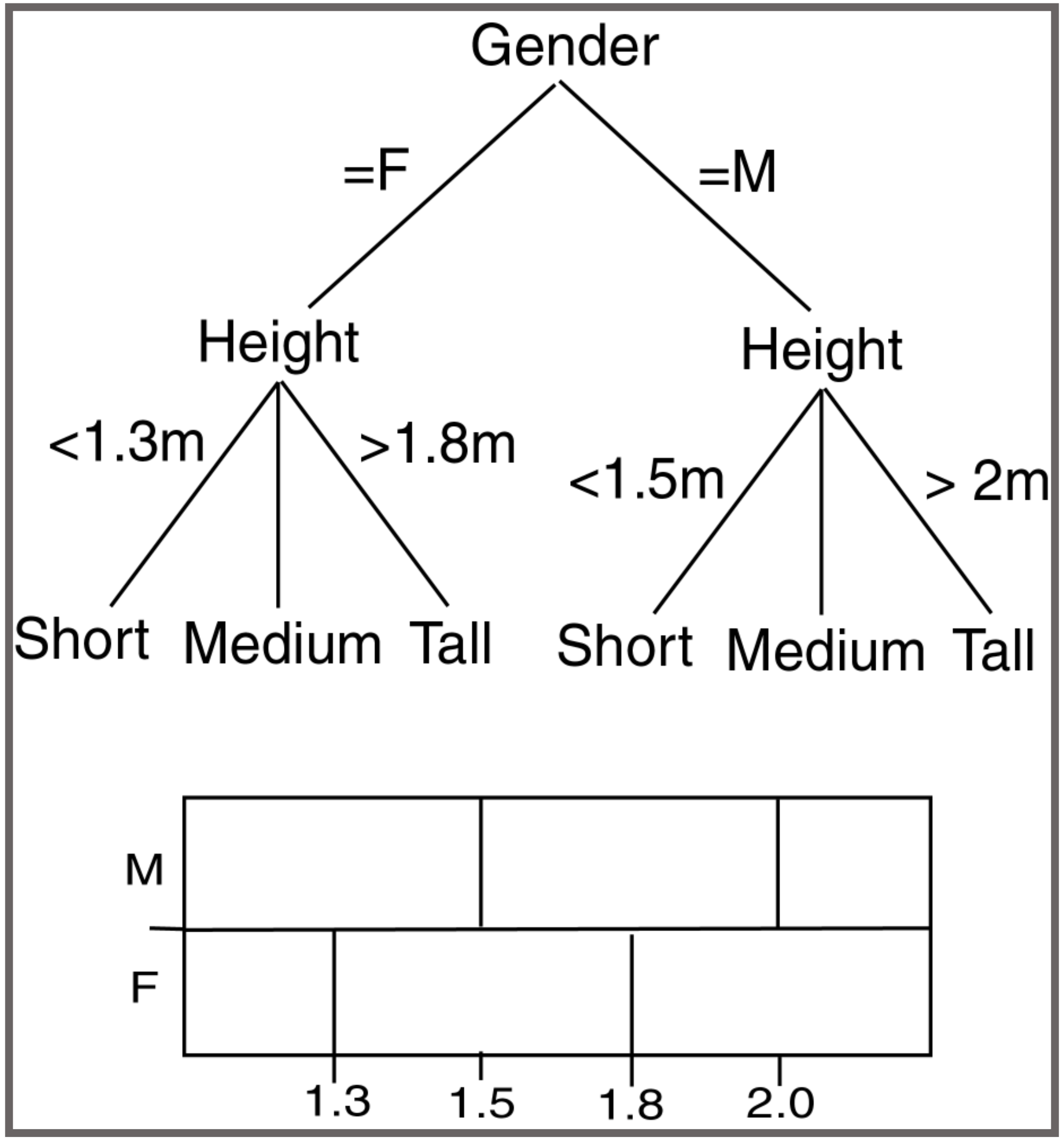
DECISION TREES

- A decision tree is a collection of ***decision nodes***, connected by ***branches***, extending downward from the ***root node*** until terminating in ***leaf nodes***.
- Beginning at the root node, which by convention is placed at the top of the decision tree diagram, attributes are tested at the decision nodes, with each possible outcome resulting in a branch. Each branch then leads either to another decision node or to a terminating leaf node.

DECISION TREES

- ***Decision Tree*** is a tree where
  - the model begins with the ***root*** for the training set,
  - ***internal nodes*** are simple decision rules tested on one or more attributes,
  - Each node makes a split into various number of ***branches***, according to the outcome of the test,
  - ***leaf nodes*** represent the prediction for the class labels.

DECISION TREES

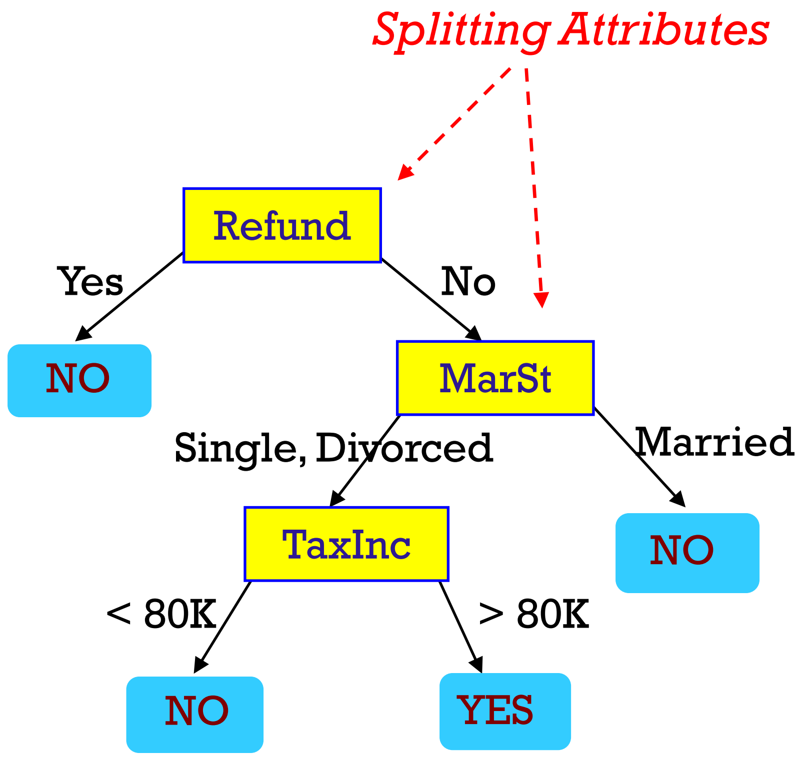


DECISION TREES

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data



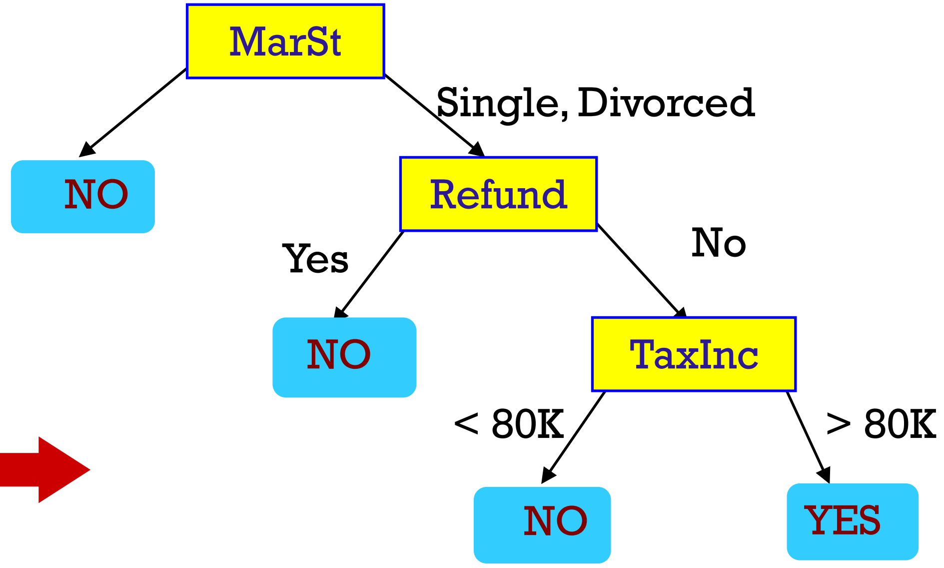
Model: Decision Tree

DECISION TREES

categorical categorical continuous class

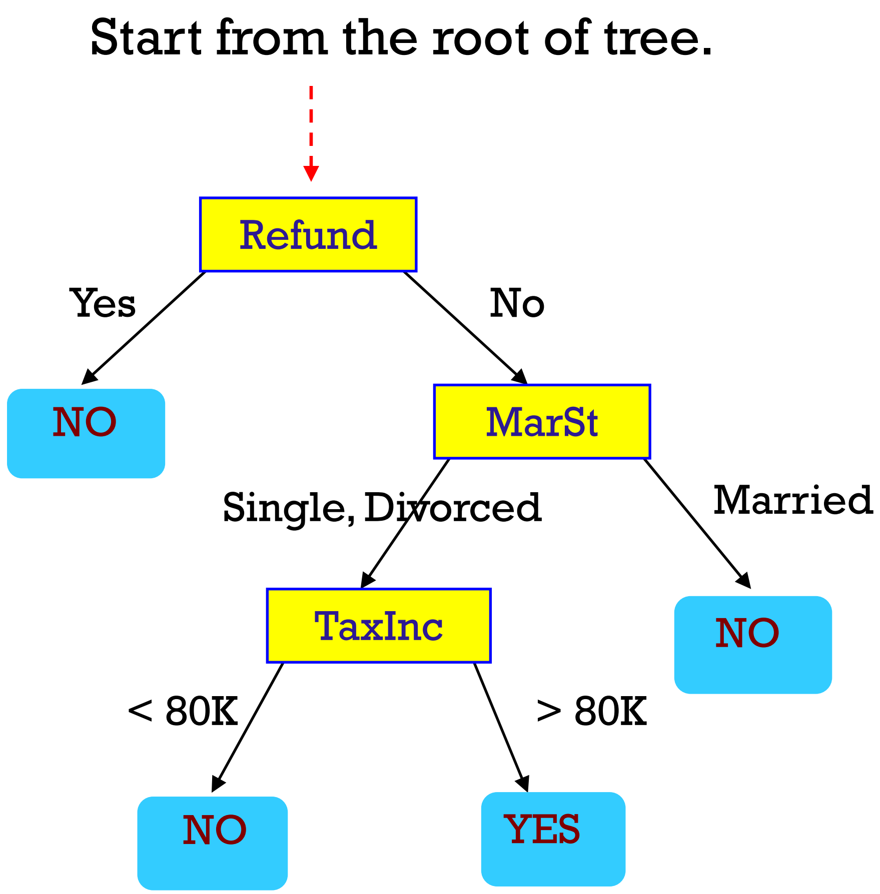
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data



There could be more than one tree that fits the same data!

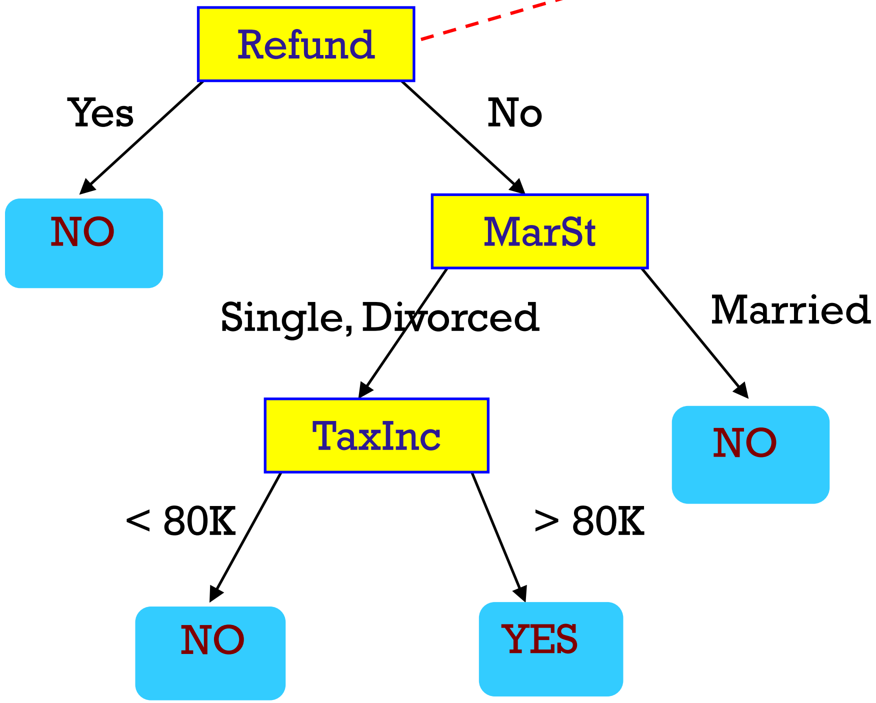
APPLY MODEL TO TEST DATA



Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

APPLY MODEL TO TEST DATA



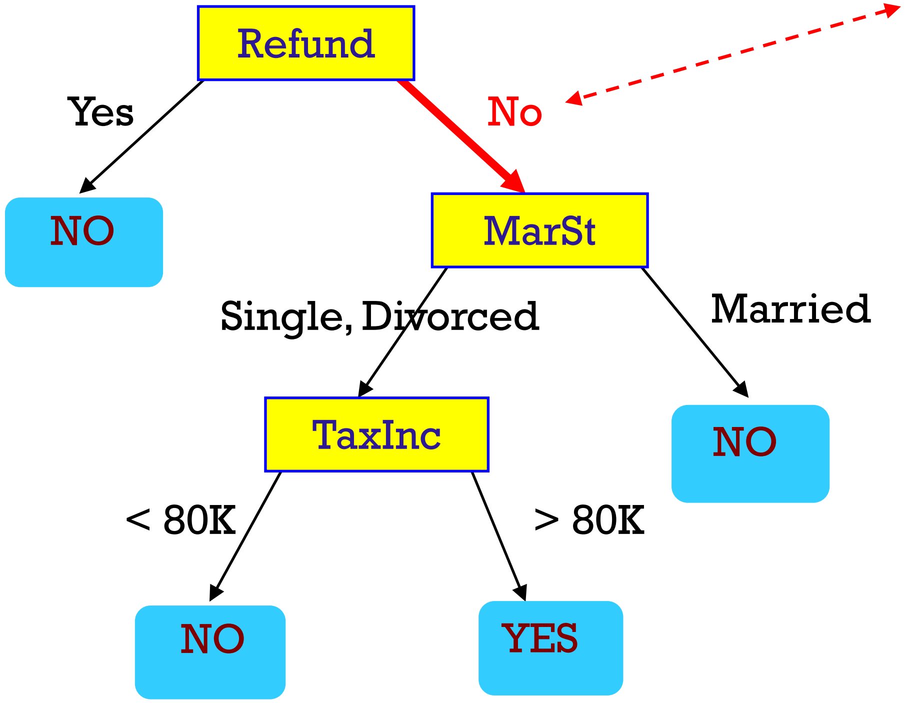
Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

APPLY MODEL TO TEST DATA

Test Data

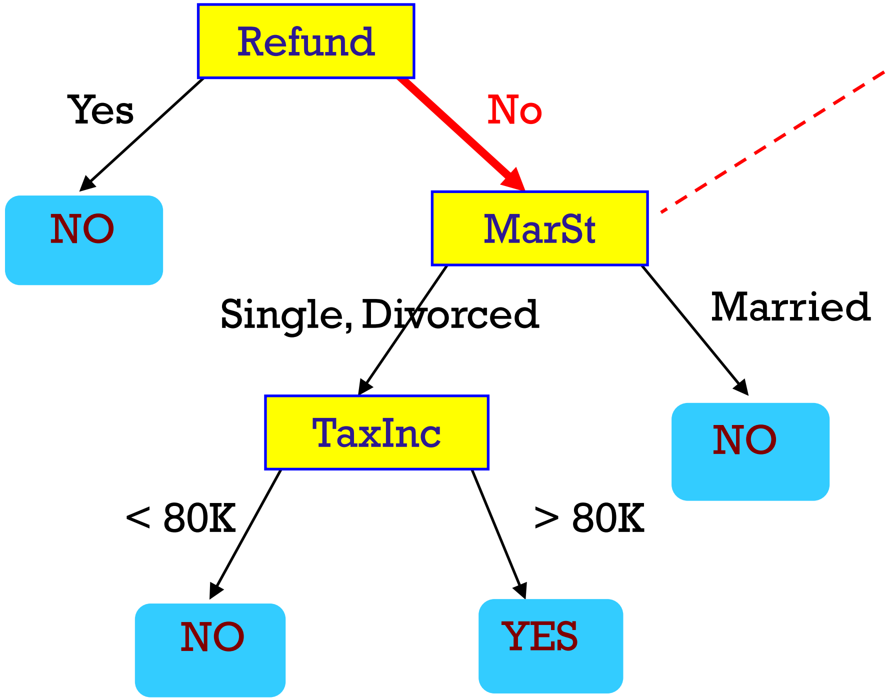
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



APPLY MODEL TO TEST DATA

Test Data

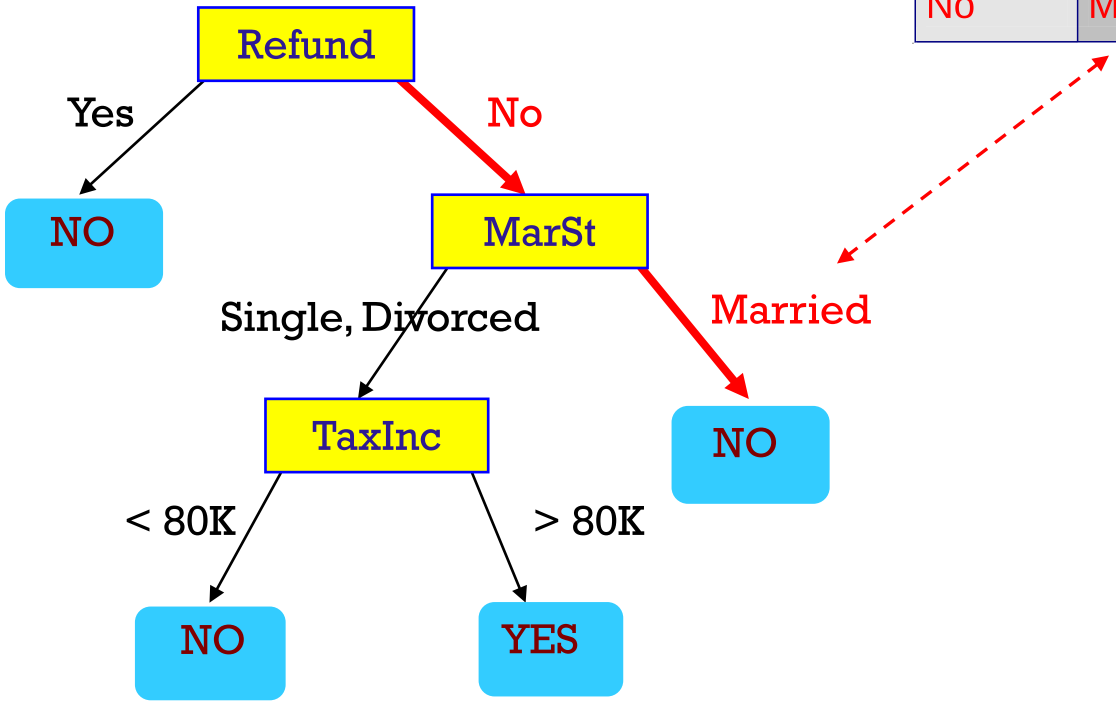
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



APPLY MODEL TO TEST DATA

Test Data

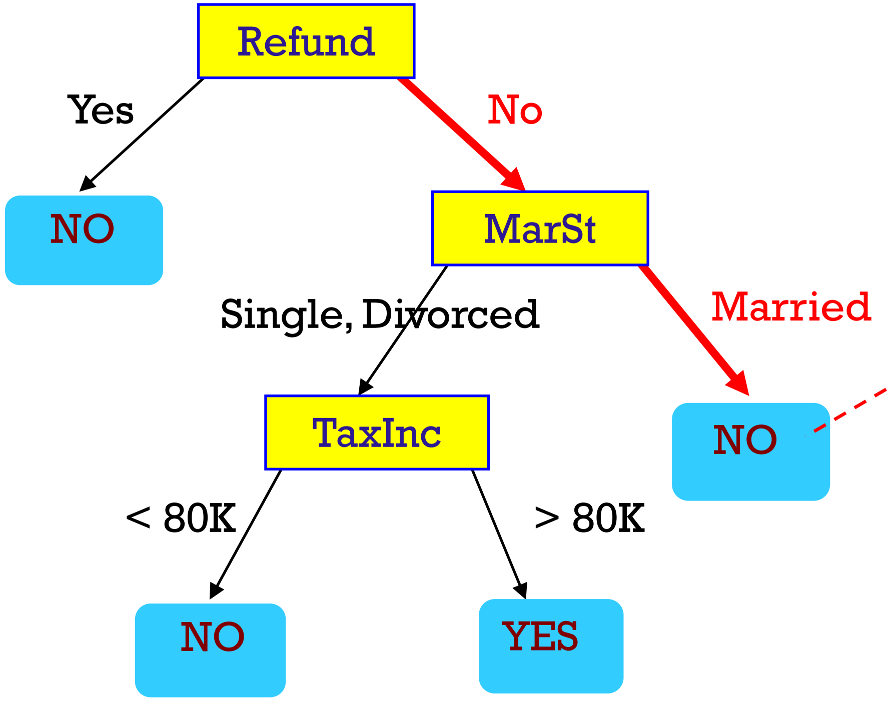
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



APPLY MODEL TO TEST DATA

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Assign Cheat to
"No"

DECISION TREE ALGORITHMS

Some requirements:

1. Pre-classified target variables.
2. A training data set – rich and varied.
3. Discrete target attribute classes.

- Issues
  - Determine how to split the records
    - How to specify the attribute test condition?
    - How to determine the best split?
  - Determine when to stop splitting

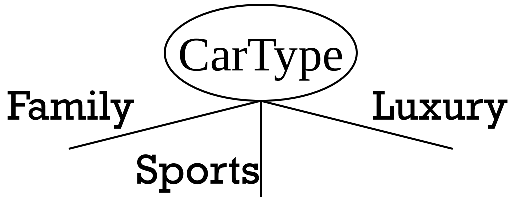
HOW TO SPECIFY TEST CONDITION?

- Depends on attribute types
  - Nominal
  - Ordinal
  - Continuous

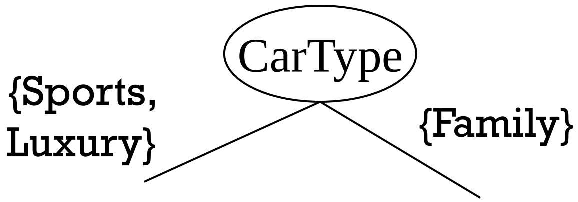
- Depends on number of ways to split
  - 2-way split
  - Multi-way split

SPLITTING BASED ON NOMINAL ATTRIBUTES

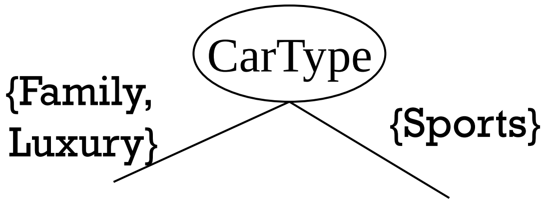
- Multi-way split: Use as many partitions as distinct values.



- Binary split: Divides values into two subsets.
  Need to find optimal partitioning.

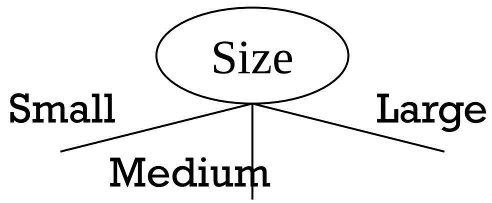


OR

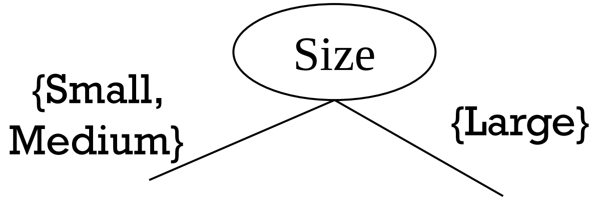


SPLITTING BASED ON ORDINAL ATTRIBUTES

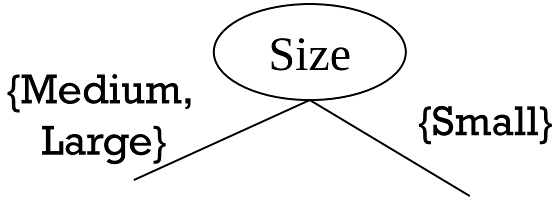
- Multi-way split: Use as many partitions as distinct values.



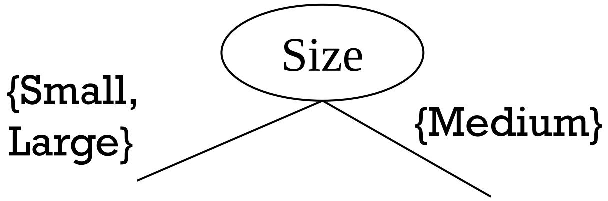
- Binary split: Divides values into two subsets.
  Need to find optimal partitioning.



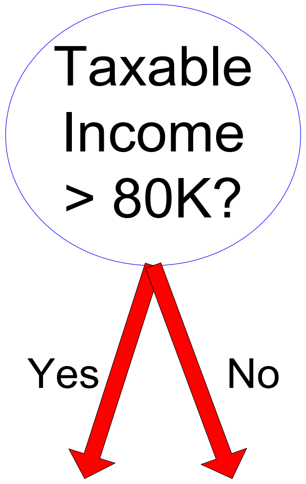
OR



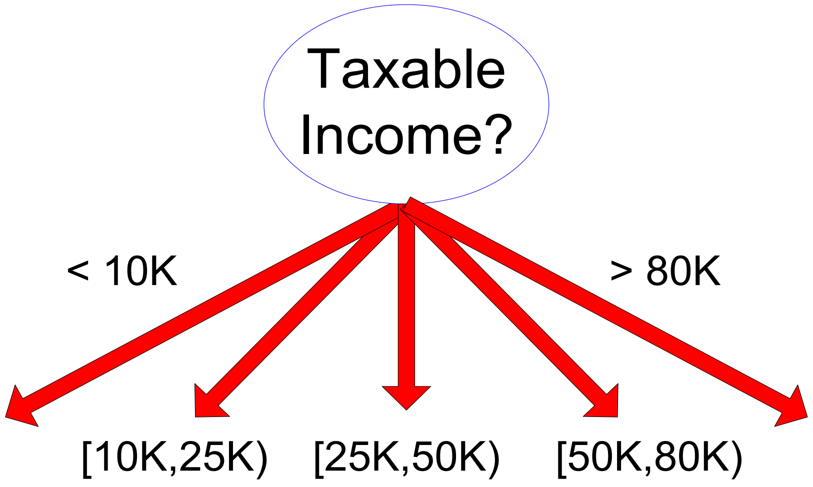
- What about this split?



SPLITTING BASED ON CONTINUOUS ATTRIBUTES



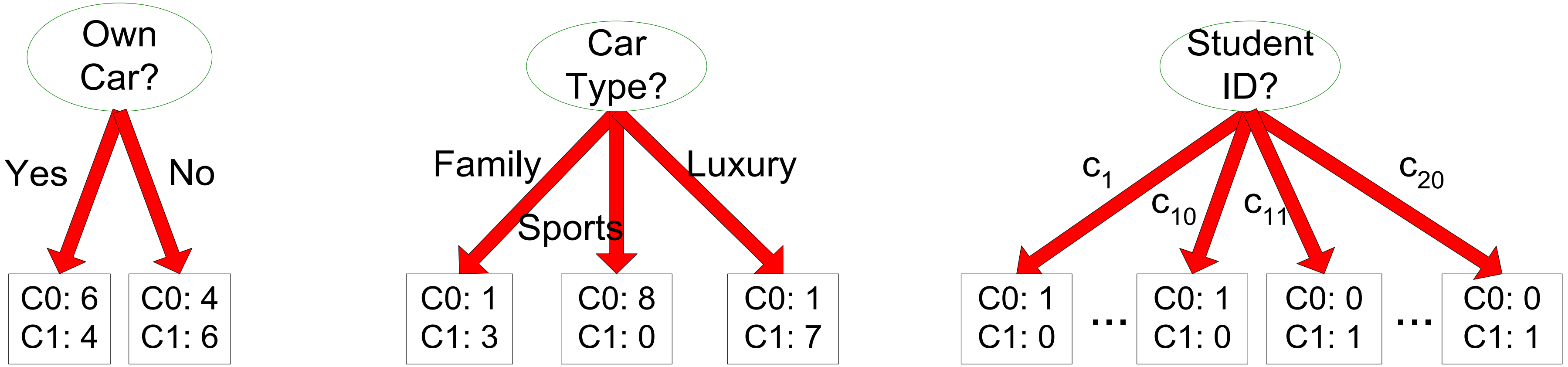
(i) Binary split



(ii) Multi-way split

HOW TO DETERMINE THE BEST SPLIT

Before Splitting: 10 records of class 0,
10 records of class 1



Which test condition is the best?

HOW TO DETERMINE THE BEST SPLIT

- Greedy approach:
  - Nodes with homogeneous class distribution are preferred
- Need a measure of node impurity:

C0: 5 C1: 5	C0: 9 C1: 1
Non-homogeneous, High degree of impurity	Homogeneous, Low degree of impurity

DECISION TREE ALGORITHMS

- ID3
  - Quinlan (1981)
  - Tries to reduce expected number of comparison
- C 4.5
  - Quinlan (1993)
  - It is an extension of ID3
  - Used in many data mining applications (C5.0)
  - Also used for rule induction
- CART
  - Breiman, Friedman, Olshen, and Stone (1984)
  - Classification and Regression Trees
- CHAID
  - Kass (1980)
  - Oldest decision tree algorithm
  - Well established in database marketing industry
- QUEST
  - Loh and Shih (1997)

C4.5 ALGORITHM

- Quinlan's extension of his own ID3 algorithm (Quinlan,1992).
- Multi-way split, (not a binary tree).
- Uses "***Information gain***" or "***entropy reduction***" to compute impurity to select the optimal split.
- Let X is an attribute with *k* possible values of probabilities *p1, p2,..., pk* .
- ***Entropy*** is the smallest number of bits, on average per symbol, needed to transmit a stream of symbols representing the values of X observed.

$$Entropy = H(x) = -\sum_{j} p_j \log_2(p_j)$$

C4.5 ALGORITHM – ENTROPY

- Entropy is a measure of randomness, a measure of the impurity in a collection of training examples.
- Entropy is a non-negative value.

$$Entropy = H(x) = -\sum_{j} p_j \log_2(p_j)$$

- When is entropy minimum?

$$H_{min} = -\sum_{j} 1 \log_2(1) = 0$$

- When is entropy maximum?

$$H_{max} = -\sum_{j=1}^{n} \frac{1}{n} \log_2\left(\frac{1}{n}\right) = -\frac{1}{n} n \log_2\left(\frac{1}{n}\right) = -\log_2\left(\frac{1}{n}\right)$$

C4.5 ALGORITHM – ENTROPY

C1	**0**
C2	**6**

$P(C1) = 0/6 = 0 \quad P(C2) = 6/6 = 1$

$\text{Entropy} = -0 \log 0 - 1 \log 1 = -0 - 0 = 0$

C1	**1**
C2	**5**

$P(C1) = 1/6 \quad P(C2) = 5/6$

$\text{Entropy} = -(1/6) \log_2 (1/6) - (5/6) \log_2 (1/6) = 0.65$

C1	**2**
C2	**4**

$P(C1) = 2/6 \quad P(C2) = 4/6$

$\text{Entropy} = -(2/6) \log_2 (2/6) - (4/6) \log_2 (4/6) = 0.92$

C4.5 ALGORITHM – INFORMATION GAIN

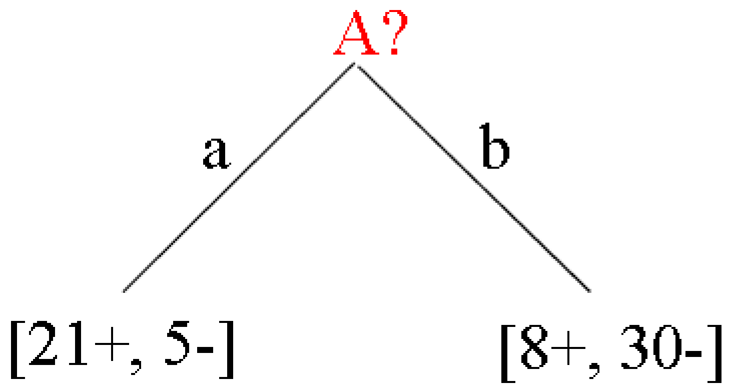
- ***Information gain*** is a measure of the effectiveness of an attribute in classifying the training data and measures the expected reduction in entropy by partitioning the examples according to an attribute.

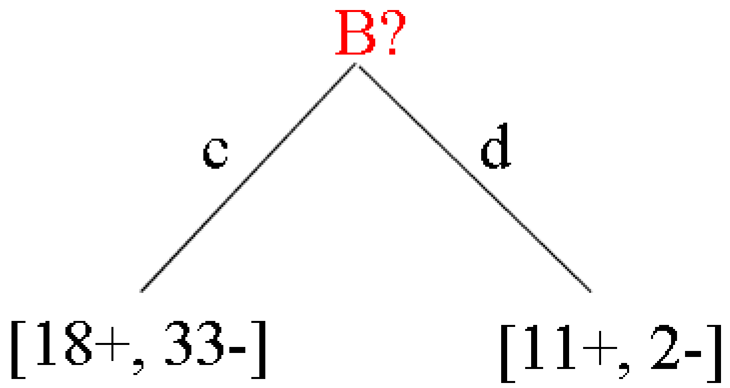
$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} (|S_v| / |S|) \, Entropy(S_v)$$

- S – a collection of examples
- A – an attribute
- Values(A) – possible values of attribute A
- Sv – the subset of S for which attribute A has value v

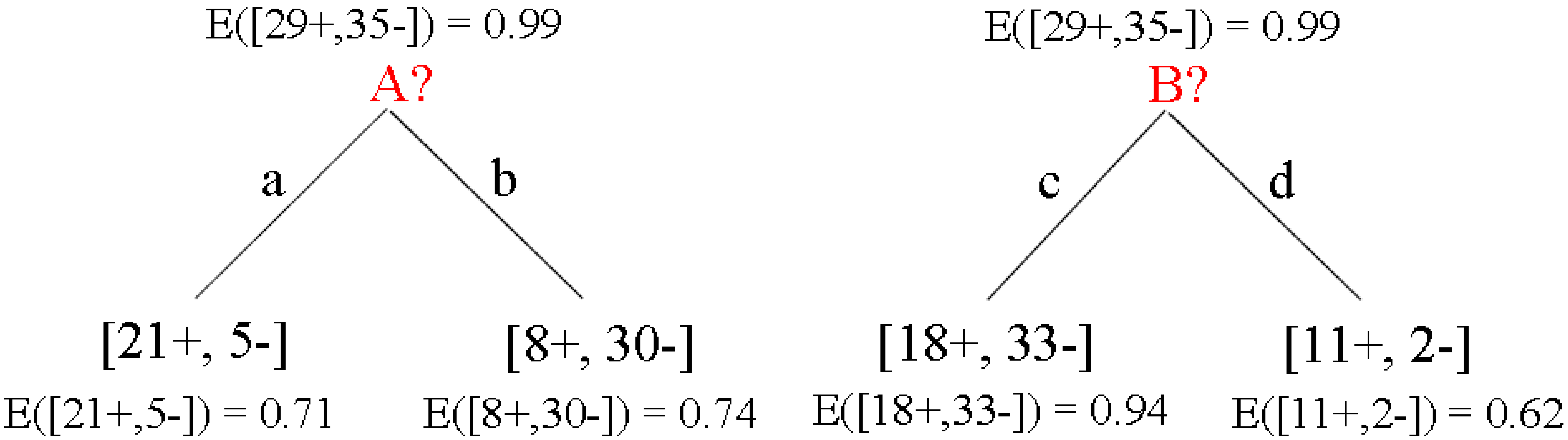
C4.5 ALGORITHM – INFORMATION GAIN

- Which attribute is the best classifier?
  - S: [29+,35-] Attributes: A and B
  - possible values for A: a,b possible values for B: c,d





C4.5 ALGORITHM – INFORMATION GAIN



$$Gain(S,A) = Ent(S) - \frac{26}{64}Ent([21+,5-]) - \frac{38}{64}Ent([8+,30-]) = 0.99 - \frac{26}{64}0.71 - \frac{38}{64}0.74 = 0.27$$

$$Gain(S,B) = Ent(S) - \frac{51}{64}Ent([18+,33-]) - \frac{13}{64}Ent([11+,2-]) = 0.99 - \frac{51}{64}0.94 - \frac{13}{64}0.62 = 0.12$$

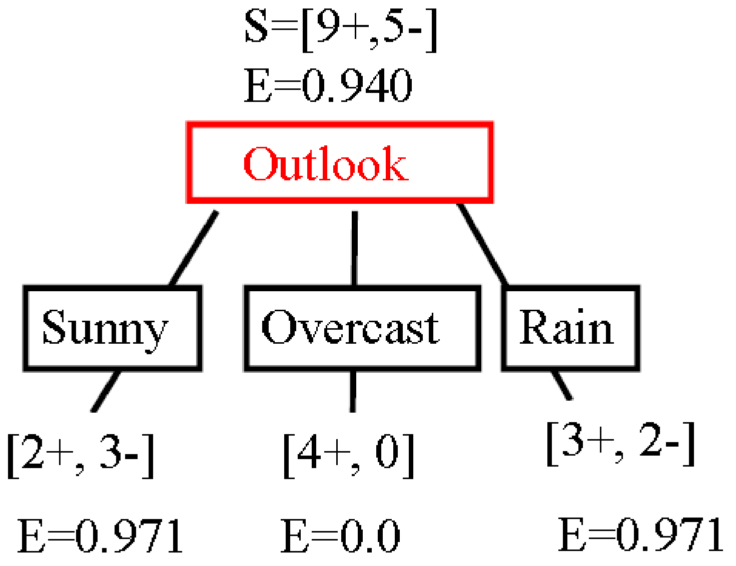
C4.5 ALGORITHM – INFORMATION GAIN

Day	Outlook	Temp.	Humidity	Wind	Play Tennis
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	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	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

C4.5 ALGORITHM – INFORMATION GAIN

- $E(S) = -(9/14)\log_2(9/14) - (5/14)\log_2(5/14)$

  $= 0.940$



Day	Outlook	Temp.	Humidity	Wind	Play Tennis
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	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	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

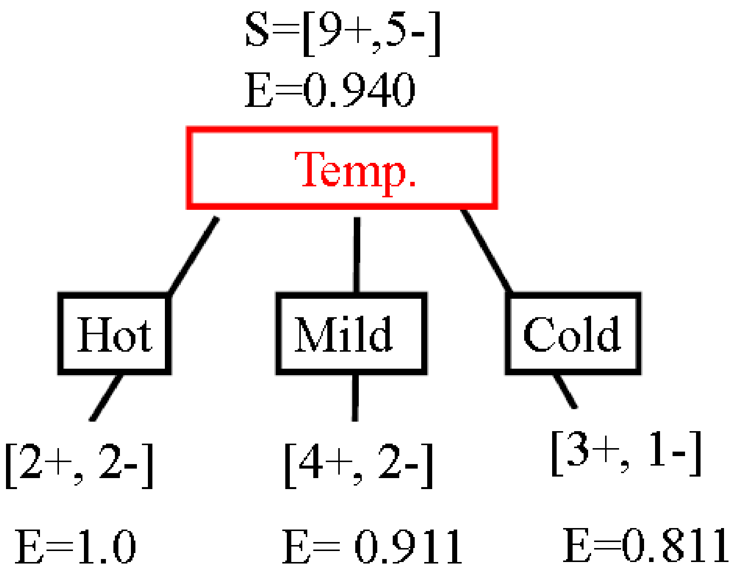
$$Gain(S, Outlook) = 0.940 - \frac{5}{14}0.971 - \frac{4}{14}0 - \frac{5}{14}0.971$$

$$Gain(S, Outlook) = 0.247$$

C4.5 ALGORITHM – INFORMATION GAIN

- $E(S) = -(9/14)\log_2(9/14) - (5/14)\log_2(5/14)$

  $= 0.940$



Day	Outlook	Temp.	Humidity	Wind	Play Tennis
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	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	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

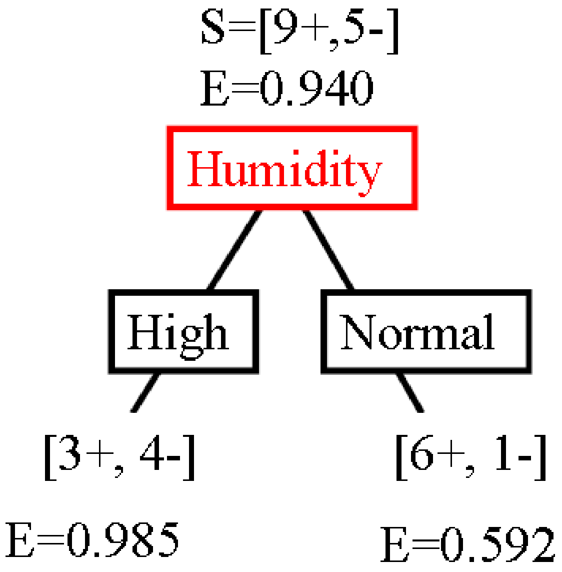
$$Gain(S,Temp) = 0.940 - \frac{4}{14}1 - \frac{6}{14}0.911 - \frac{4}{14}0.811$$

$$Gain(S,Temp) = 0.029$$

C4.5 ALGORITHM – INFORMATION GAIN

- $E(S) = -(9/14)\log_2(9/14) - (5/14)\log_2(5/14)$

  $= 0.940$



Day	Outlook	Temp.	Humidity	Wind	Play Tennis
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	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	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

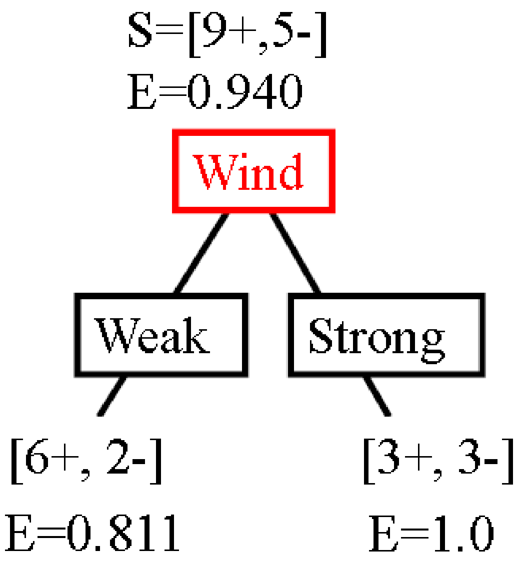
$$Gain(S, Humidity) = 0.940 - \frac{7}{14}0.985 - \frac{7}{14}0.592$$

$$Gain(S, Humidity) = 0.151$$

C4.5 ALGORITHM – INFORMATION GAIN

- $E(S) = -(9/14)\log_2(9/14) - (5/14)\log_2(5/14)$

  $= 0.940$



Day	Outlook	Temp.	Humidity	Wind	Play Tennis
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	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	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

$$Gain(S, Wind) = 0.940 - \frac{8}{14}0.811 - \frac{6}{14}1$$

$$Gain(S, Wind) = 0.048$$

C4.5 ALGORITHM – INFORMATION GAIN

Day	Outlook	Temp.	Humidity	Wind	Play Tennis
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	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	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

- $E(S) = -(9/14)\log_2(9/14) - (5/14)\log_2(5/14)$

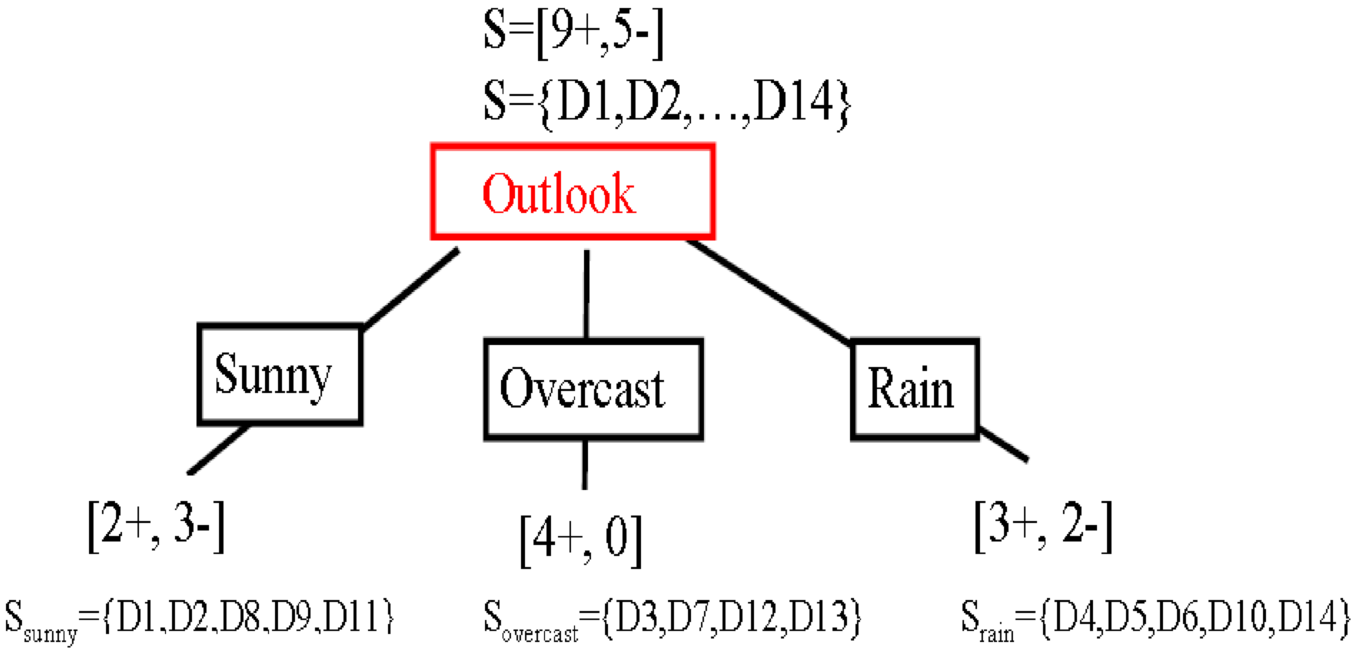
  $= 0.940$

$$Gain(S, Outlook) = 0.247$$

$$Gain(S, Temp) = 0.029$$

$$Gain(S, Humidity) = 0.151$$

$$Gain(S, Wind) = 0.048$$



C4.5 – EXAMPLE

Cus	Savings	Assets	Income ($1000s)	Credit Ri
1	Medium	High	75	Good
2	Low	Low	50	Bad
3	High	Medium	25	Bad
4	Medium	Medium	50	Good
5	Low	Medium	100	Good
6	High	High	25	Good
7	Low	Low	25	Bad
8	Medium	Medium	75	Good

C4.5 – EXAMPLE

Cust	Savings	Assets	Income ($1000s)	Credit Ri
1	Medium	High	75	Good
2	Low	Low	50	Bad
3	High	Medium	25	Bad
4	Medium	Medium	50	Good
5	Low	Medium	100	Good
6	High	High	25	Good
7	Low	Low	25	Bad
8	Medium	Medium	75	Good

Candidate Splits at Root Node for C4.5 Algorithm

Cand	Child Nodes		
1	$Savings = low$	$Savings = medium$	$Savings = high$
2	$Assets = low$	$Assets = medium$	$Assets = high$
3	$Income \leq \$25,000$	$Income > \$25,000$	
4	$Income \leq \$50,000$	$Income > \$50,000$	
5	$Income \leq \$75,000$	$Income > \$75,000$	

C4.5 – EXAMPLE

Cust	Savings	Assets	Income ($1000s)	Credit Ri
1	Medium	High	75	Good
2	Low	Low	50	Bad
3	High	Medium	25	Bad
4	Medium	Medium	50	Good
5	Low	Medium	100	Good
6	High	High	25	Good
7	Low	Low	25	Bad
8	Medium	Medium	75	Good

- 5 Good – 3 Bad Credit Risk

$$E(S) = -\sum_j p_j \log_2(p_j) = -\frac{5}{8}\log_2\left(\frac{5}{8}\right) - \frac{3}{8}\log_2\left(\frac{3}{8}\right) = 0.9544$$

$$P_{Savings} \Rightarrow P_{high} = \frac{2}{8} \qquad P_{medium} = \frac{3}{8} \qquad P_{low} = \frac{3}{8}$$

$$H_{savings}(high) = -\sum_j p_j \log_2(p_j) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$$

$$H_{savings}(medium) = -\sum_j p_j \log_2(p_j) = -\frac{3}{3}\log_2\left(\frac{3}{3}\right) - \frac{0}{3}\log_2\left(\frac{0}{3}\right) = 0$$

$$H_{savings}(low) = -\sum_j p_j \log_2(p_j) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) = 0.9183$$

C4.5 – EXAMPLE

Cust	Savings	Assets	Income ($1000s)	Credit Ri
1	Medium	High	75	Good
2	Low	Low	50	Bad
3	High	Medium	25	Bad
4	Medium	Medium	50	Good
5	Low	Medium	100	Good
6	High	High	25	Good
7	Low	Low	25	Bad
8	Medium	Medium	75	Good

- 5 Good – 3 Bad Credit Risk

$$E(S) = -\sum_j p_j \log_2(p_j) = -\frac{5}{8}\log_2\left(\frac{5}{8}\right) - \frac{3}{8}\log_2\left(\frac{3}{8}\right) = 0.9544$$

$$P_{Savings} \Rightarrow P_{high} = \frac{2}{8} \qquad P_{medium} = \frac{3}{8} \qquad P_{low} = \frac{3}{8}$$

$$H_{savings}(high) = 1$$

$$H_{Savings}(S) = \sum_{i=1}^{k} P_i H_{Savings}(S_i)$$

$$H_{savings}(medium) = 0$$

$$H_{savings}(low) = 0.9183$$

C4.5 – EXAMPLE

Cust	Savings	Assets	Income ($1000s)	Credit Ri
1	Medium	High	75	Good
2	Low	Low	50	Bad
3	High	Medium	25	Bad
4	Medium	Medium	50	Good
5	Low	Medium	100	Good
6	High	High	25	Good
7	Low	Low	25	Bad
8	Medium	Medium	75	Good

$$Gain(S, Savings) = E(S) - H_{Savings}(S)$$

$$= 0.9544 - 0.5944 = 0.36$$

C4.5 – EXAMPLE

Cust	Savings	Assets	Income ($1000s)	Credit Ri
1	Medium	High	75	Good
2	Low	Low	50	Bad
3	High	Medium	25	Bad
4	Medium	Medium	50	Good
5	Low	Medium	100	Good
6	High	High	25	Good
7	Low	Low	25	Bad
8	Medium	Medium	75	Good

- For Assets

$$P_{high} = \frac{2}{8} \qquad P_{medium} = \frac{4}{8} \qquad P_{low} = \frac{2}{8}$$

$$H_{assets}(high) = -\frac{2}{2}\log_2\left(\frac{2}{2}\right) - \frac{0}{2}\log_2\left(\frac{0}{2}\right) = 0$$

$$H_{assets}(medium) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = 0.8113$$

$$H_{assets}(low) = -\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right) = 0$$

$$H_{Assets}(S) = \left(\frac{2}{8}\times 0\right) + \left(\frac{4}{8}\times 0.8113\right) + \left(\frac{2}{8}\times 0\right) = 0.4057$$

$$Gain(S, Assets) = H(S) - H_{Assets}(S) = 0.9544 - 0.4057 = 0.5487\, bits$$

C4.5 – EXAMPLE

Cust	Savings	Assets	Income ($1000s)	Credit Ri
1	Medium	High	75	Good
2	Low	Low	50	Bad
3	High	Medium	25	Bad
4	Medium	Medium	50	Good
5	Low	Medium	100	Good
6	High	High	25	Good
7	Low	Low	25	Bad
8	Medium	Medium	75	Good

- For Income<=25000

$$P_{income \le 25K} = \frac{3}{8} \quad P_{income > 25K} = \frac{5}{8}$$

$$H_{income \le 25K}(income \le 25K) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) = 0.9183$$

$$H_{income \le 25K}(income > 25K) = -\frac{4}{5}\log_2\left(\frac{4}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right) = 0.7219$$

$$H_{income \le 25K}(S) = \left(\frac{3}{8} \times 0.9183\right) + \left(\frac{5}{8} \times 0.7219\right) = 0.7956$$

$$Gain(income \le 25K) = H(S) - H_{income \le 25K}(S) = 0.9544 - 0.7956 = 0.1588\,bits$$

C4.5 – EXAMPLE

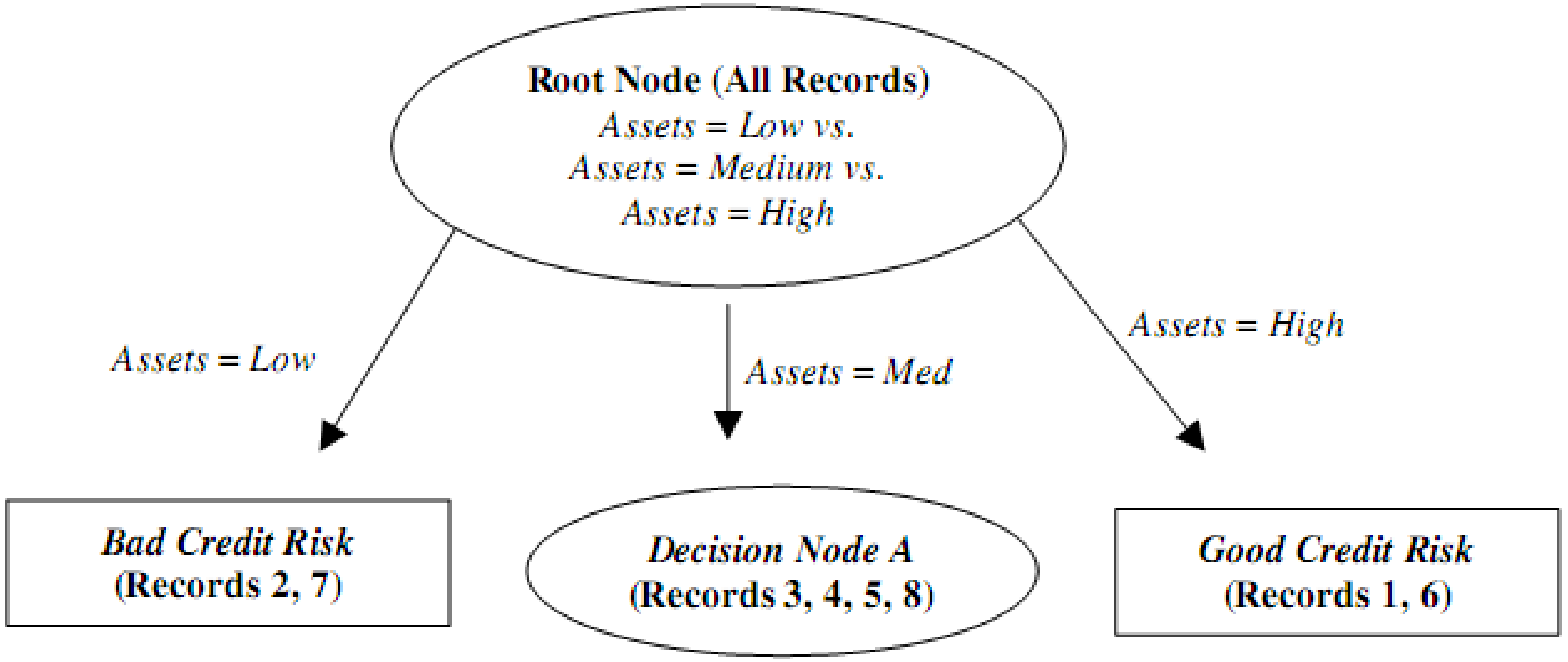
Cust	Savings	Assets	Income ($1000s)	Credit Ri
1	Medium	High	75	Good
2	Low	Low	50	Bad
3	High	Medium	25	Bad
4	Medium	Medium	50	Good
5	Low	Medium	100	Good
6	High	High	25	Good
7	Low	Low	25	Bad
8	Medium	Medium	75	Good

Information Gain for Each Candidate Split at the Root Node

Candidate Split	Child Nodes	Information Gain (Entropy Reduction)
1	$Savings = low$ $Savings = medium$ $Savings = high$	0.36 bits
2	$Assets = low$ $Assets = medium$ $Assets = high$	0.5487 bits
3	$Income \leq \$25{,}000$ $Income > \$25{,}000$	0.1588 bits
4	$Income \leq \$50{,}000$ $Income > \$50{,}000$	0.3475 bits
5	$Income \leq \$75{,}000$ $Income > \$75{,}000$	0.0923 bits

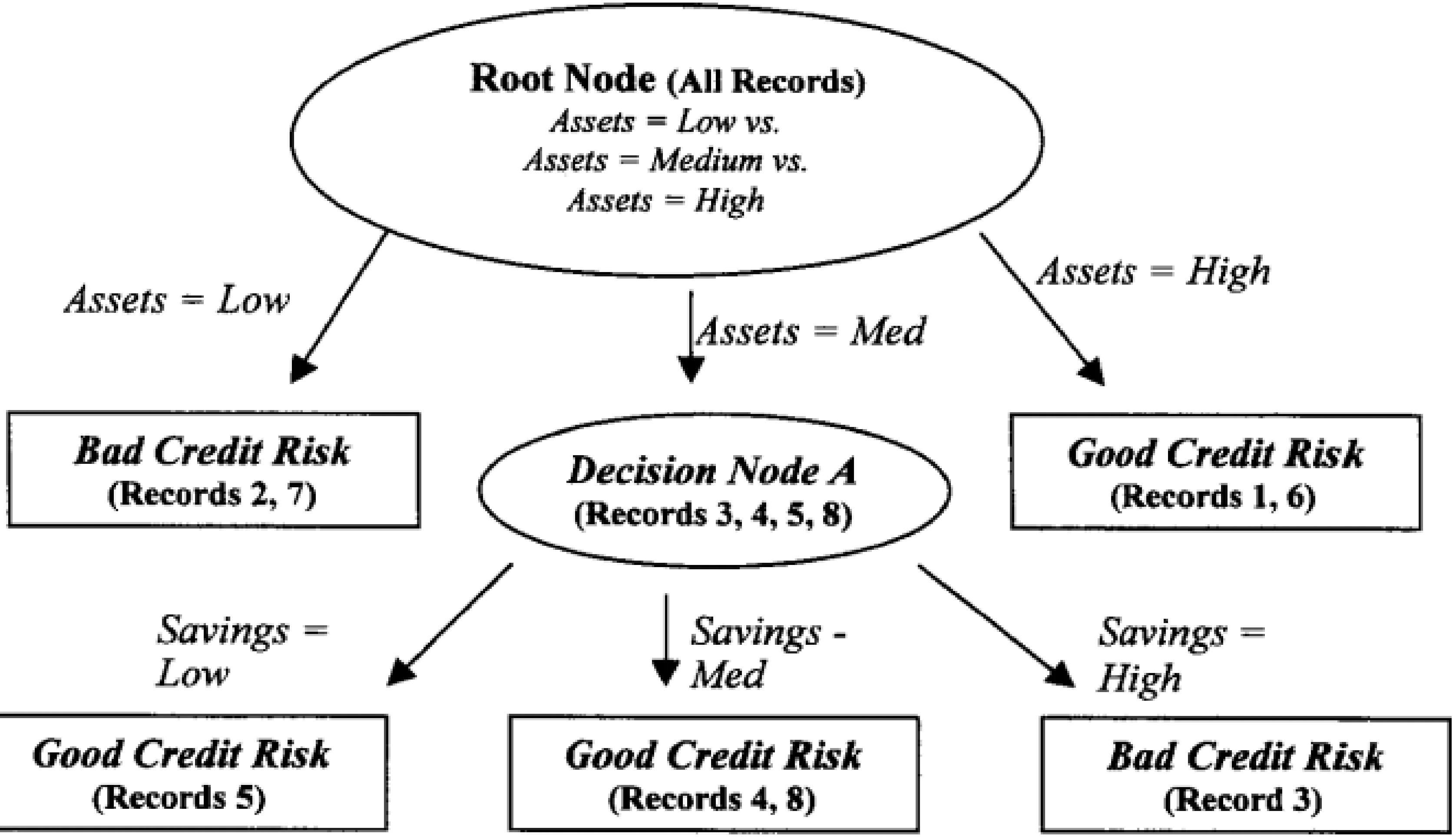
C4.5 – EXAMPLE

Cust	Savings	Assets	Income ($1000s)	Credit Ri
1	Medium	High	75	Good
2	Low	Low	50	Bad
3	High	Medium	25	Bad
4	Medium	Medium	50	Good
5	Low	Medium	100	Good
6	High	High	25	Good
7	Low	Low	25	Bad
8	Medium	Medium	75	Good



C4.5 – EXAMPLE

Cust	Savings	Assets	Income ($1000s)	Credit Ri
1	Medium	High	75	Good
2	Low	Low	50	Bad
3	High	Medium	25	Bad
4	Medium	Medium	50	Good
5	Low	Medium	100	Good
6	High	High	25	Good
7	Low	Low	25	Bad
8	Medium	Medium	75	Good



C4.5 – EXAMPLE

- Records 2 and 7 only have a "Bad" result, and records 1 and 6 have only a "Good" result. Therefore, there is no need to create a new branch from here.
- A new branch should be created only through Assets=Medium.
- To decide which attribute to use in the new decision node, the Information Gain calculation must be made on the table containing only Records 3,4,5 and 8.

THANKS
FOR
LISTENING