Data Mining

Decision Tree

Outline

A. Introduction to Decision Tree ?

B. Measurement of impurity

C. Decision Tree Models and their variations

(Tan, Steinbach and Kumar, Chapter 4.3)

Data Mining

A. Decision Tree

- We are given a set of records. Each record has the same structure, consisting of a number of attribute/value pairs.
- One of these attributes represents the goal of the record.
- Usually the goal attribute takes only the binary values true, false, or success, failure, or something equivalent.

Data Mining

Example : Golf Playing

- We are dealing with records reporting on weather conditions for playing golf.
- The goal attribute specifies whether or not to Play.
- The non-goal attributes are:

ATTRIBUTE	POSSIBLE VALUES
Outlook	sunny, overcast, rain
temperature	continuous
humidity	continuous
windy	true, false

Data Mining

Outlook	Temperature	Humidity	Windy	Play
sunny	85	85	false	Don't Play
sunny	80	90	true	Don't Play
overcast	83	78	false	Play
rain	70	96	false	Play
rain	68	80	false	Play
rain	65	70	true	Don't Play
overcast	64	65	true	Play
sunny	72	95	false	Don't Play
sunny	69	70	false	Play
rain	75	80	false	Play
sunny	75	70	true	Play
overcast	72	90	true	Play
overcast	81	75	false	Play
rain	71	80	true	Don't Play

Data Mining

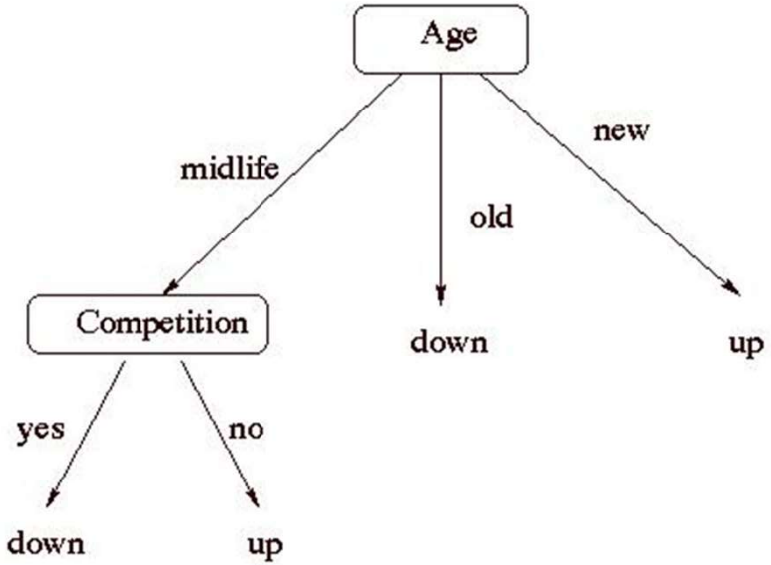
Example : Stock Market

- A simpler example from the stock market involving only discrete ranges has Profit as goal attribute, with values {up, down}. Its non-goal attributes are:

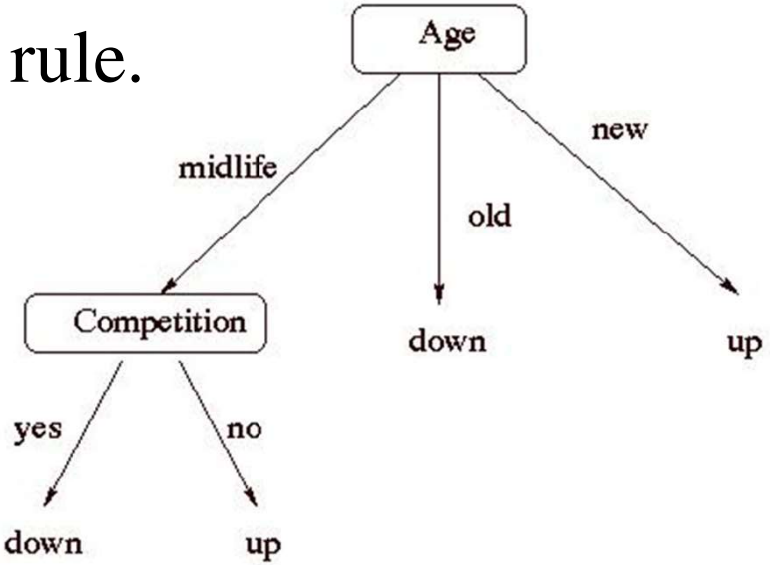
Attribute	Possible Values
age	old, midlife, new
competitive	no, yes
type	software, hardware

Data Mining

	age	competition	type	profit
1	old	yes	software	down
2	old	no	software	down
3	old	no	hardware	down
4	midlife	yes	software	down
5	midlife	yes	hardware	down
6	midlife	no	hardware	up
7	midlife	no	software	up
8	new	yes	software	up
9	new	no	hardware	up
10	new	no	software	up



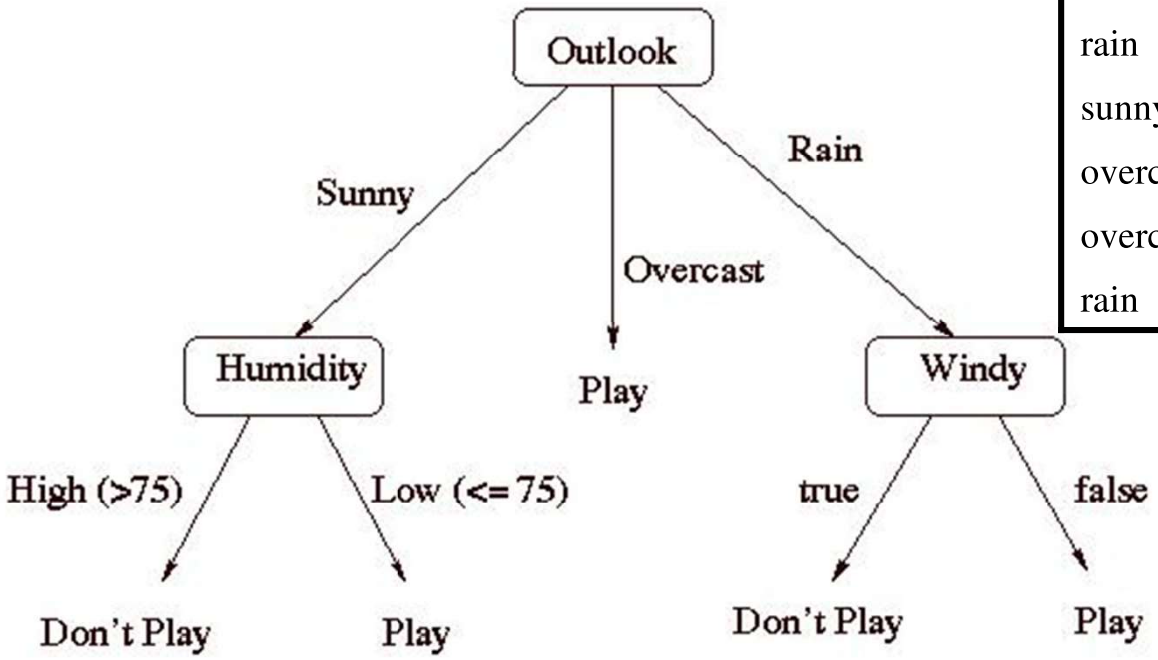
- Each internal node of the tree tests an attribute.
- Each branch corresponds to a possible value for that attribute.
- Each leaf node provides a classification.
- Each tree path corresponds to a rule.



Data Mining

Golf Playing Example

Outlook	Temp.	Humi	Windy	Play
sunny	85	85	false	Don't Play
sunny	80	90	true	Don't Play
overcast	83	78	false	Play
rain	70	96	false	Play
rain	68	80	false	Play
rain	65	70	true	Don't Play
overcast	64	65	true	Play
sunny	72	95	false	Don't Play
sunny	69	70	false	Play
rain	75	80	false	Play
sunny	75	70	true	Play
overcast	72	90	true	Play
overcast	81	75	false	Play
rain	71	80	true	Don't Play

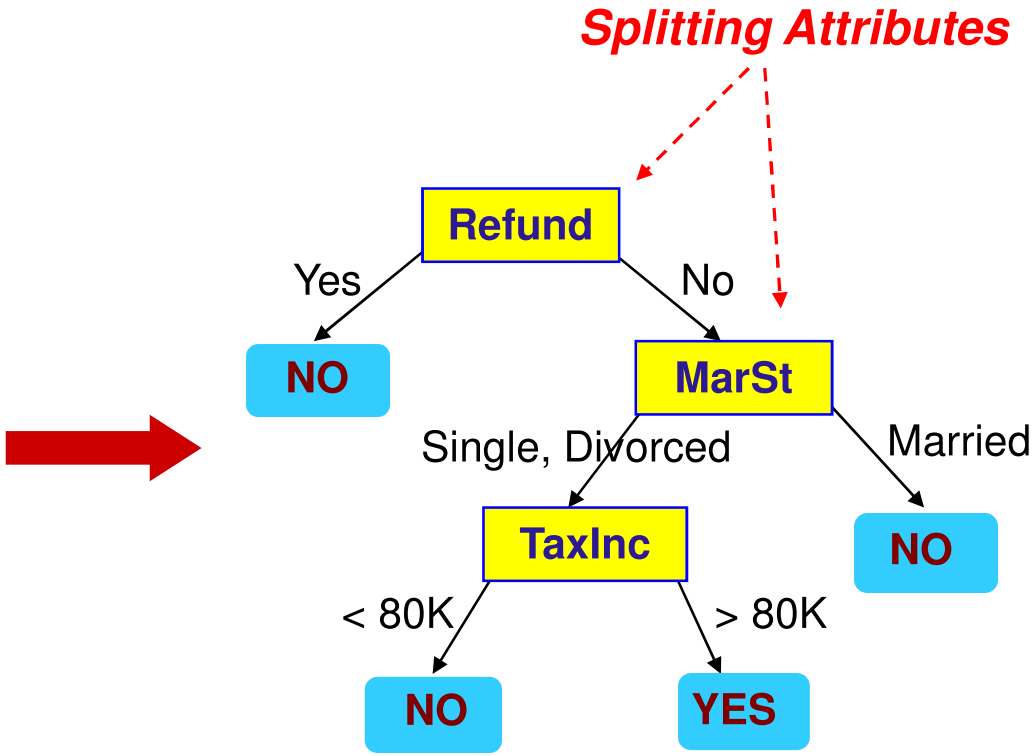


Example of a Decision Tree

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

categorical categorical continuous class

Training Data

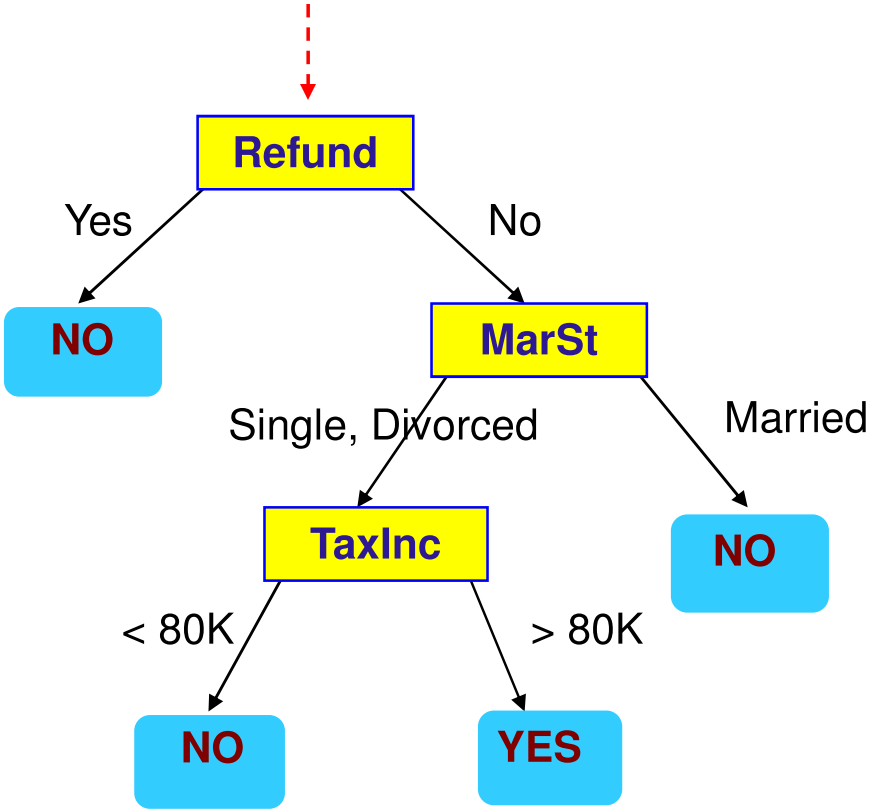


Model: Decision Tree

Data Mining

Apply Model to Test Data

Start from the root of tree.



Test Data

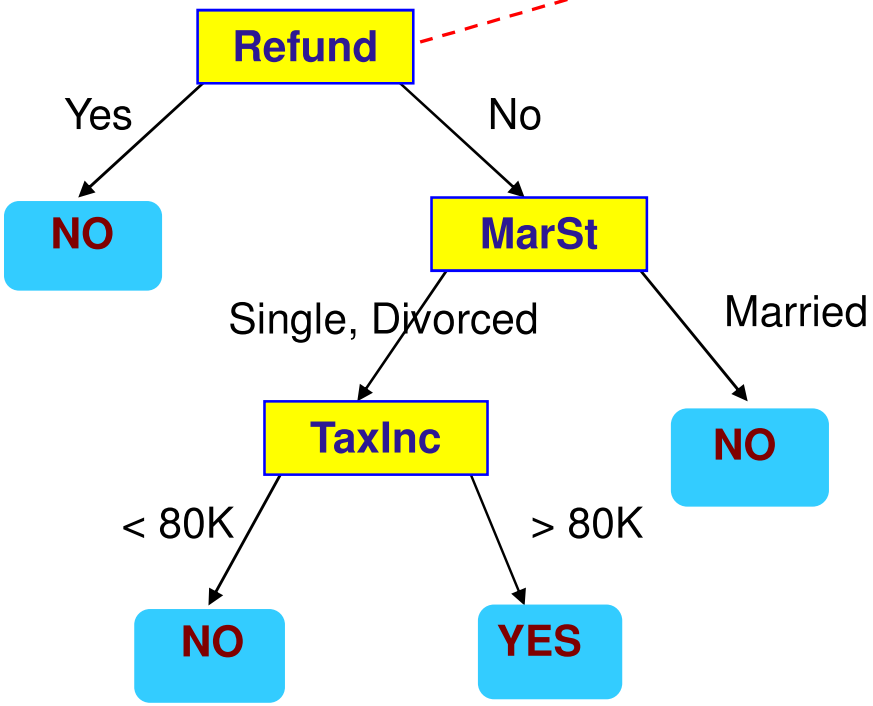
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Data Mining

Apply Model to Test Data

Test Data

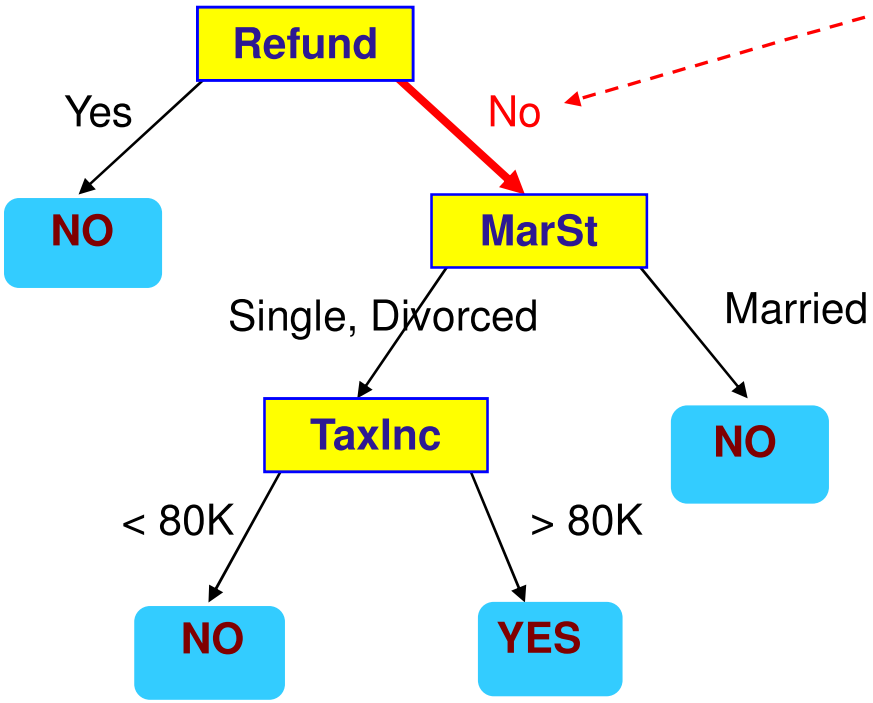
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

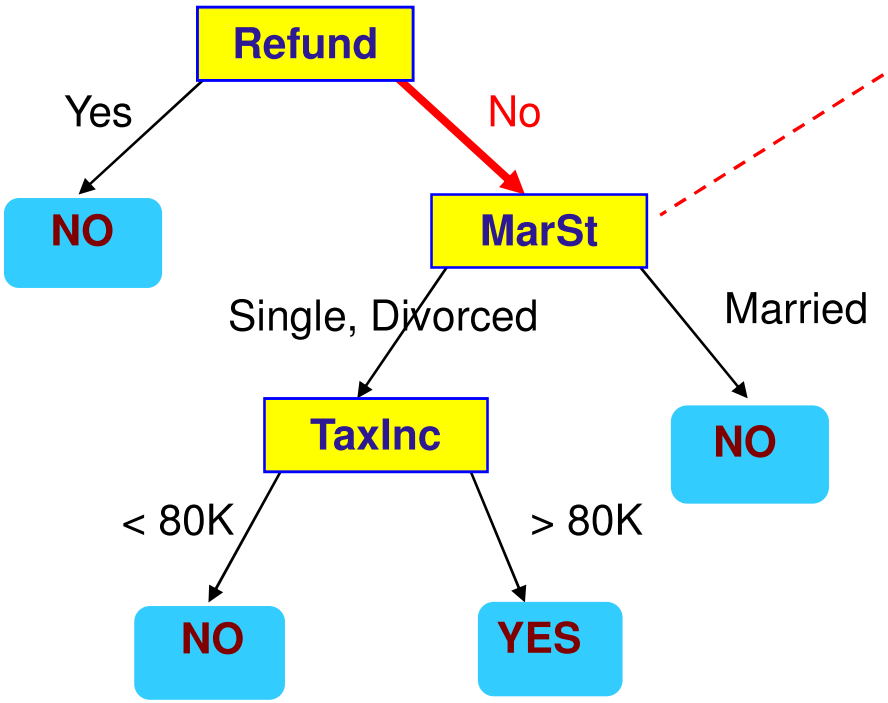


Data Mining

Apply Model to Test Data

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

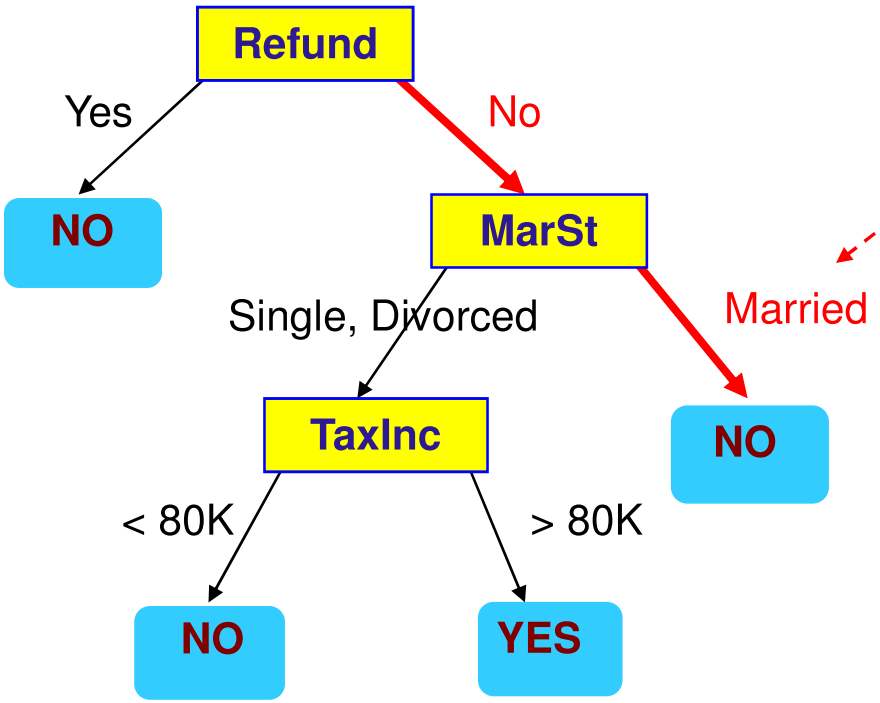


Data Mining

Apply Model to Test Data

Test Data

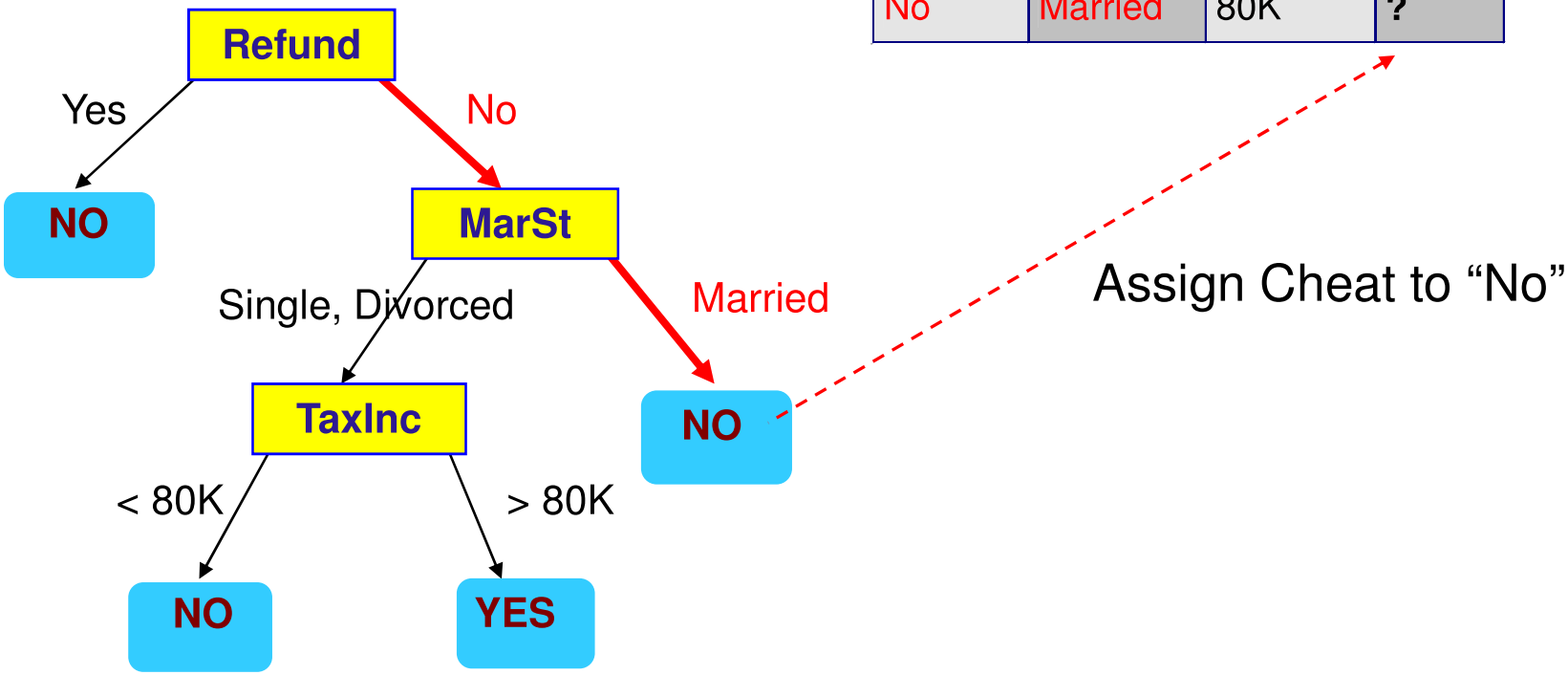
Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Apply Model to Test Data

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Data Mining

How to Specify Test Condition?

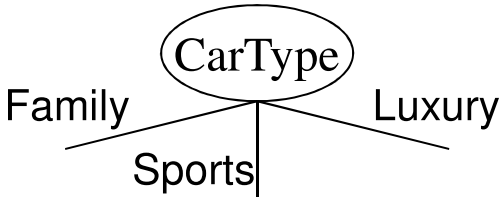
- Depends on attribute types
  - Nominal
  - Ordinal
  - Continuous

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

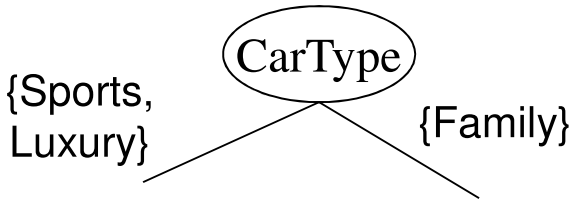
Data Mining

Splitting Based on Nominal Attributes

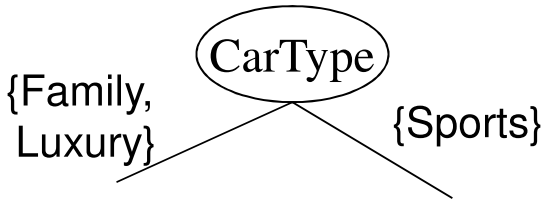
- Multi-way split: Use as many partitions as distinct values.



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

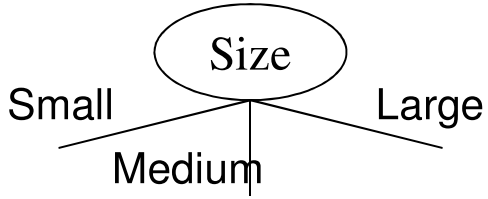


OR

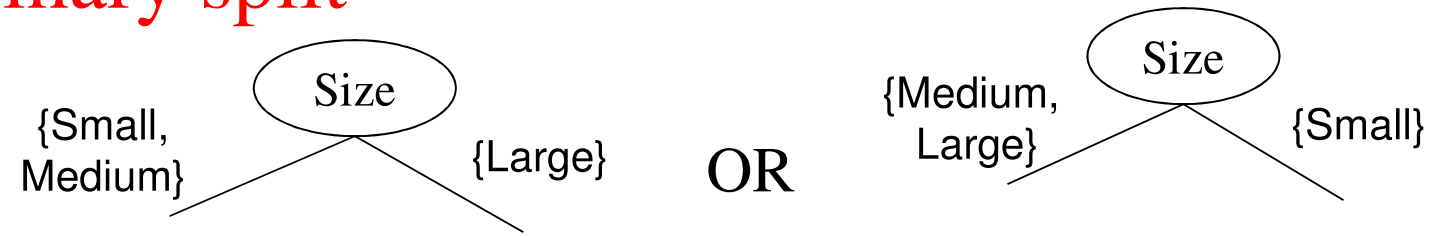


Splitting Based on Ordinal Attributes

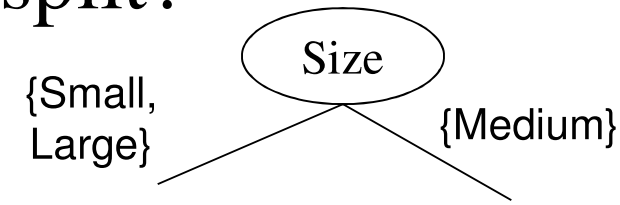
- Multi-way split



- Binary split



- What about this split?

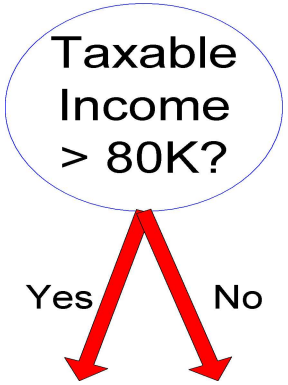


Data Mining

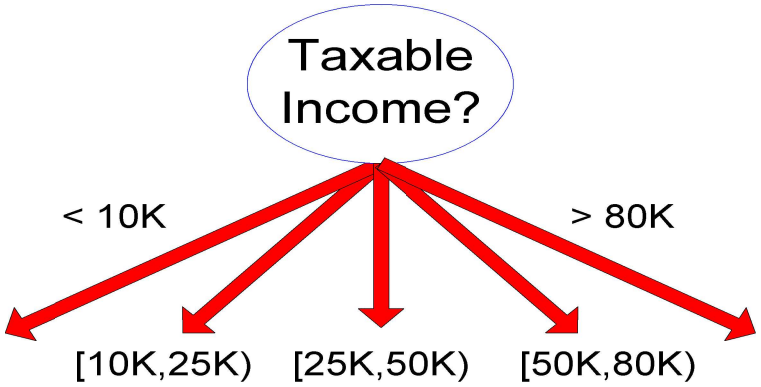
Splitting Based on Continuous Attributes

- Different ways of handling
  - Discretization to form an ordinal categorical attribute
    - Static – discretize once at the beginning
    - Dynamic – ranges can be found by equal interval bucketing, equal frequency bucketing (percentiles), or clustering.
  - Binary Decision: $(A < v)$ or $(A \geq v)$
    - consider all possible splits and finds the best cut
    - can be more compute intensive

Splitting Based on Continuous Attributes



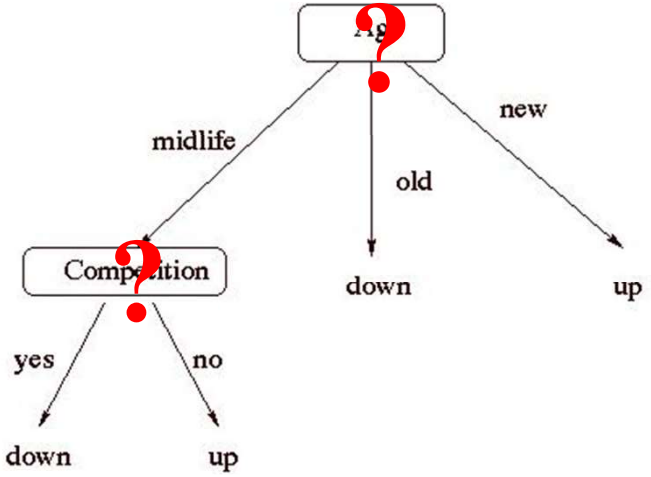
(i) Binary split



(ii) Multi-way split

Decision Tree

- The problem is to determine a decision tree that on the basis of answers to questions about the non-goal attributes predicts correctly the value of the goal attribute



Basic Decision Tree Learning Method

1. Choose best attribute
2. Extend tree by adding branch for each attribute value
3. Sort training examples to leaf nodes
4. If all/most training examples are being classified, then stop, else repeat steps 1-4 for leaf nodes.

Data Mining

Variations of decision trees

- Which attribute is the best ?
- When to stop the learning ?

Decision Tree Models

- ID3 (by Quinlan in 1979)
- C4.5 (by Quinlan in 1993)
- CART (by Breiman 1984)

Data Mining

B. How to determine the Best Split

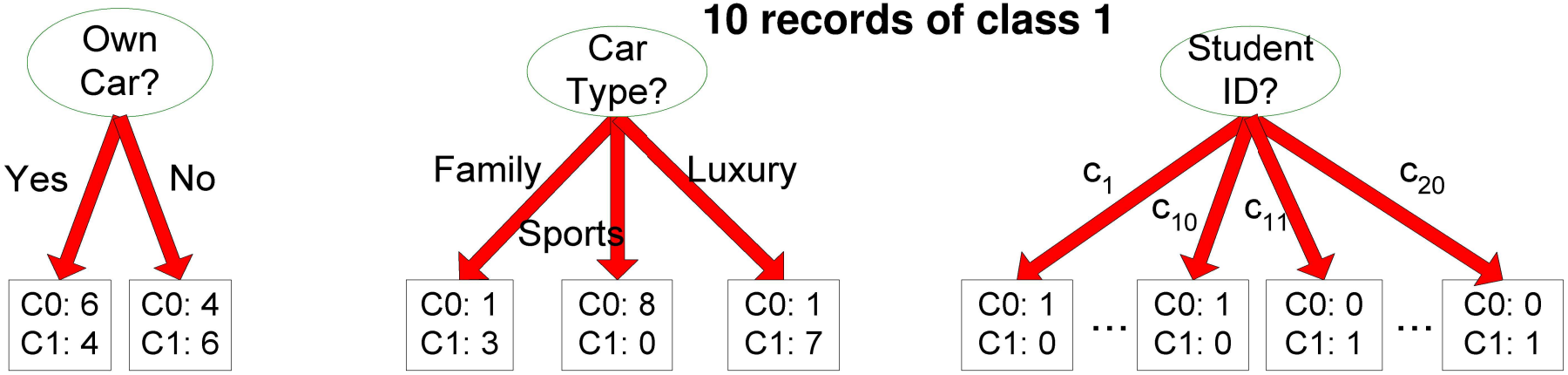
- Greedy approach:
  - Nodes with homogeneous class distribution are preferred
- Need a measure of node impurity:

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

Data Mining

How to determine the Best Split

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



Which test condition is the best?

Data Mining

How to Find the Best Split

Before Splitting:

C0	**N00**
C1	**N01**

→ **M0**

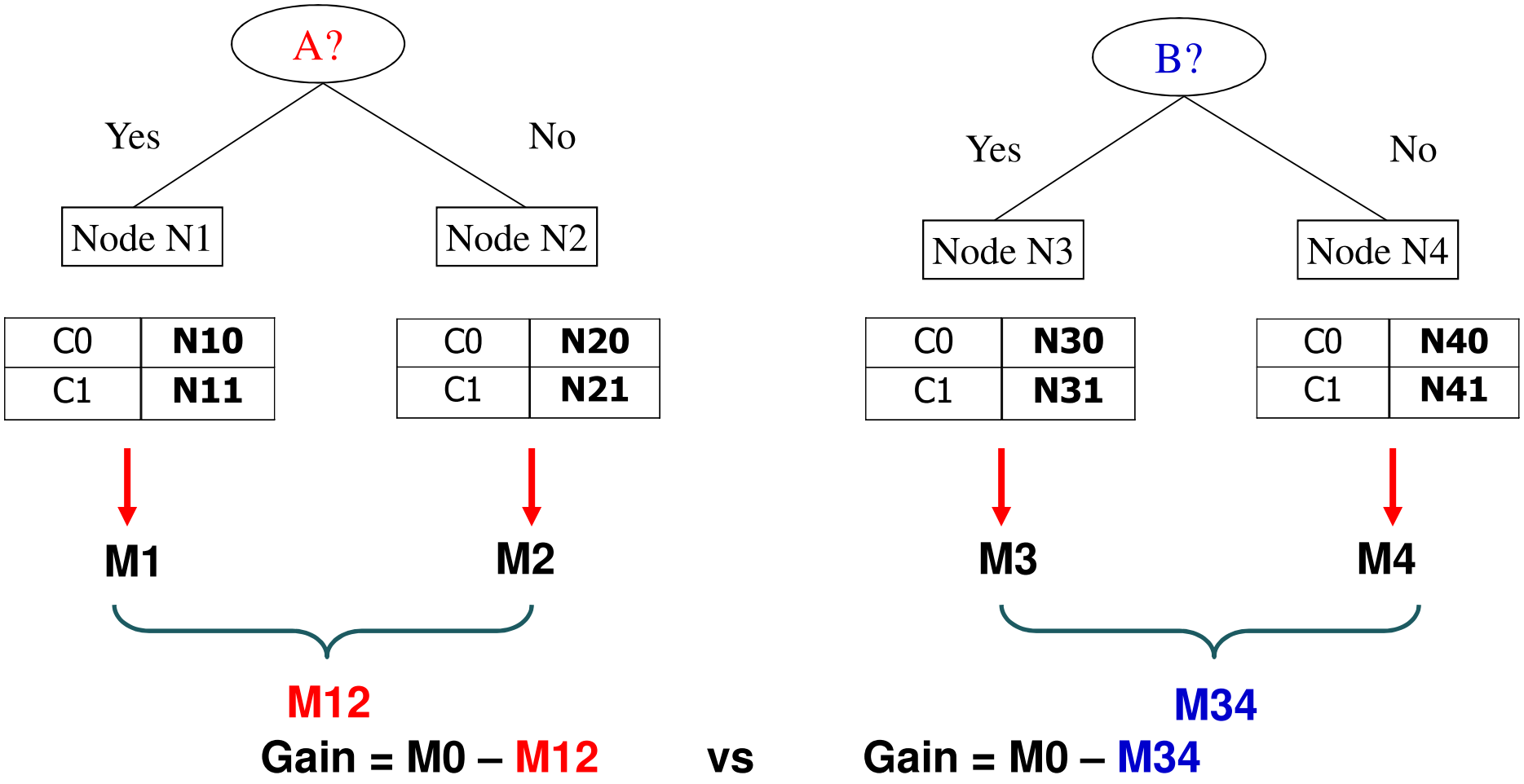
Data Mining

How to Find the Best Split

Before Splitting:

C0	**N00**
C1	**N01**

→ **M0**



Gain = M0 – M12 **vs** **Gain = M0 – M34**

Measure of the "Impurity" in general

- The measurement should be
  - zero if p_j is concentrated on one class.
  - maximal if p_j is uniform.
  - a symmetric function of its arguments.

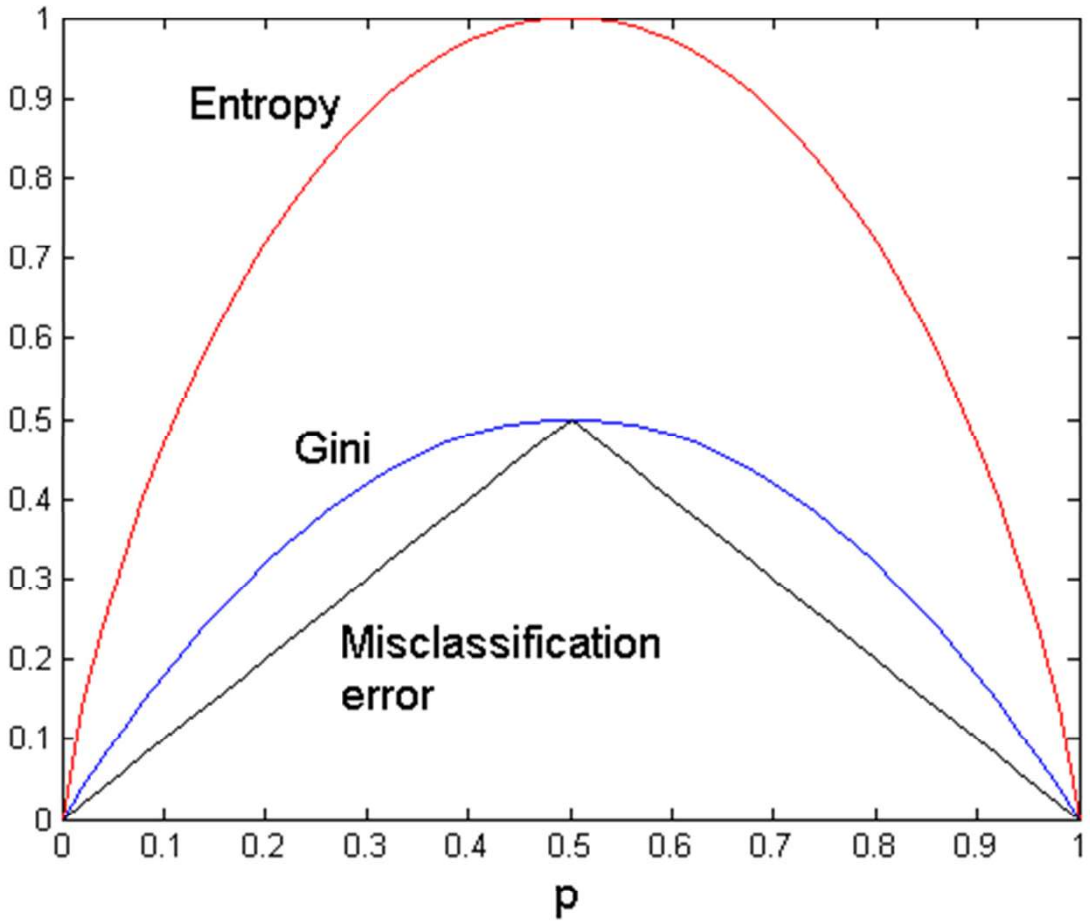
Data Mining

Measures of Node Impurity

- Gini Index
- Entropy
- Misclassification error

Plot of various measurement

For a 2-class problem:



Examples of the measurement

- Entropy

$$I(p) = -(p_1 \times \log(p_1) + p_2 \times \log(p_2) + .. + p_n \times \log(p_n))$$

- Misclassification cost

$$I(p) = 1 - \max_j p_j$$

- Gini

$$I(p) = \Sigma_{i \neq j} p_i p_j = 1 - \Sigma_j p_j^2$$

Entropy

Example 1 :

- Consider a random variable which has a uniform distribution over 32 outcomes.
- To identify an outcome, we need a label that takes on 32 different values. Thus 5 bit strings suffice as labels.

Entropy (cont')

Example 2 :

- Suppose we have a horse race with eight horses taking part.
- Assume that the probabilities of winning for the eight horses are
- ½ , ¼ , ⅛, 1/16, 1/64 , 1/64, 1/64, 1/64

- Suppose we wish to send a message to another person indicating which horse won the race. One method is to send a 3 bit string to denote the index of the winning horse.

- Another method is to use a variable length coding set (i.e. 0, 10, 110, 1110, 111100, 111101, 111110, 111111) to represent the eight horses.
- The average description length is

$$\tfrac{1}{2} \times 1 + \tfrac{1}{4} \times 2 + \tfrac{1}{8} \times 3 + \{1/16\} \times 4 + 4 \times \{1/64\} \times 5$$

$$= \tfrac{1}{2} \log 2 + \tfrac{1}{4} \log 4 + \tfrac{1}{8} \log 8 + \{1/16\} \log\{16\} + 4 \times \{1/64\} \log\{64\}$$

$$= -\tfrac{1}{2} \log \tfrac{1}{2} - \tfrac{1}{4} \log \tfrac{1}{4} - \tfrac{1}{8} \log \tfrac{1}{8} - \{1/16\} \log\{1/16\}$$

$$- 4 \times \{1/64\} \log\{1/64\}$$

$$= 2 \text{ bits}$$

- All logarithms here are in base 2.

- Entropy is a way to measure the amount of information.
- If we are given a probability distribution

  $P = (p_1, p_2, .., p_n)$ then the Information conveyed by this distribution, also called the Entropy of *P*, is:

  $$I(P) = -(p_1 \times \log(p_1) + p_2 \times \log(p_2) + .. + p_n \times \log(p_n))$$
- All logarithms here are in base 2.

Data Mining

Examples for computing Entropy

$$Entropy = -\sum_j p_j \log_2 p_j$$

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

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

Entropy = – 0 log 0 – 1 log 1 = – 0 – 0 = 0

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

P(C1) = 1/6 P(C2) = 5/6

Entropy = – (1/6) $\log_2$ (1/6) – (5/6) $\log_2$ (1/6) = 0.65

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

P(C1) = 2/6 P(C2) = 4/6

Entropy = – (2/6) $\log_2$ (2/6) – (4/6) $\log_2$ (4/6) = 0.92

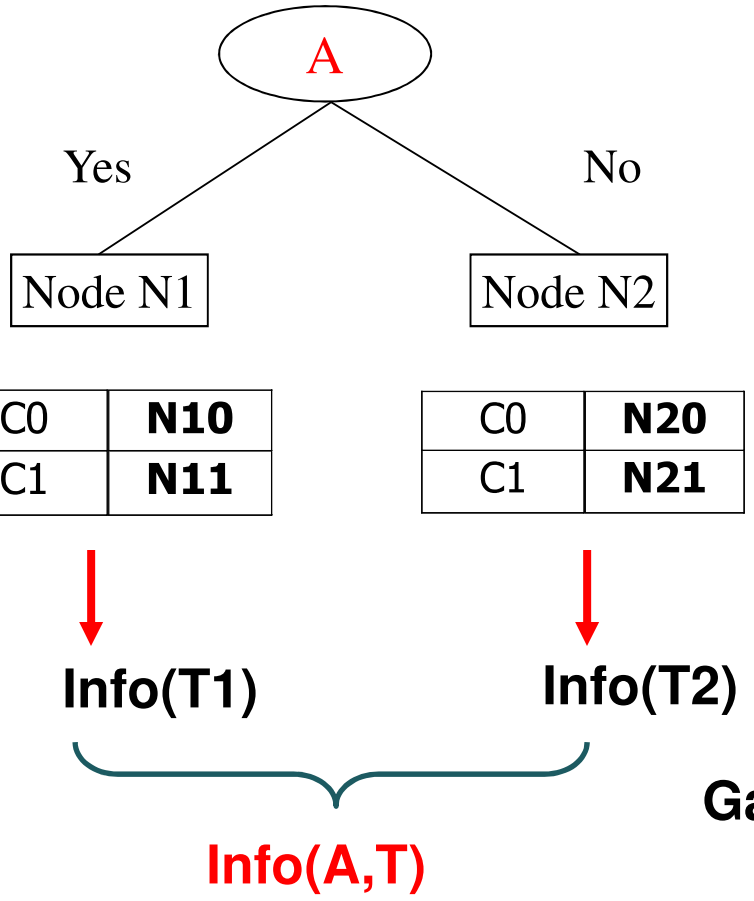
Data Mining

How to Find the Best Split

Before Splitting:

C0	**N00**
C1	**N01**

→ **Info(T)=I(P_T)**



I(P_T)= Entropy, Gini, or misclassication cost

Gain(A,T) = Info(T) – Info(A,T)

Data Mining

Info(T)

- If a set *T* of records is partitioned into disjoint exhaustive classes $C_1, C_2, .., C_k$ on the basis of the value of the goal attribute,
- then the information needed to identify the class of an element of *T* is Info(T) = I(P), where *P* is the probability distribution of the partition ($C_1, C_2, .., C_k$):

$$P = \left(\frac{|C_1|}{|T|}, \frac{|C_2|}{|T|}, \cdots, \frac{|C_k|}{|T|} \right)$$

$$\text{Info}(T) = -\left(\frac{|C_1|}{|T|} \log \frac{|C_1|}{|T|} + \frac{|C_2|}{|T|} \log \frac{|C_2|}{|T|} + \cdots \frac{|C_k|}{|T|} \log \frac{|C_k|}{|T|} \right)$$

- In the golfing example, 9 out 14 examples are in the class “Play” and the other 5 are in the class “Don't Play”. We have

  Info(T) = I(9/14, 5/14) = 0.94
- In the stock market example, 5 out of 10 examples are in the class “down” and the other 5 are in the class “up”. We have

  Info(T) = I(5/10,5/10) = 1.0

Info(X,T)

- If we first partition T on the basis of the value of a non-goal attribute X into sets $T_1, T_2, .., T_n$ then the information needed to identify the class of an element of T becomes the weighted average of the information needed to identify the class of an element of T_i, i.e. the weighted average of Info(T_i):

$$\text{Info}(X,T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \text{Info}(T_i)$$

In the golfing example

- for the attribute Outlook, we have

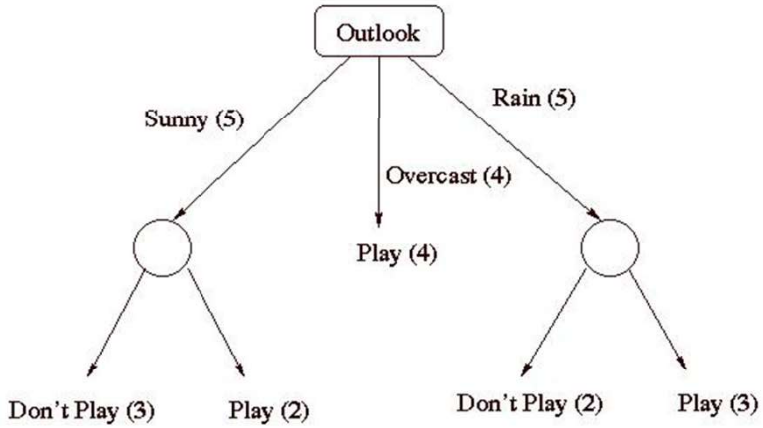
  Info(Outlook,T)

  $= 5/14 \times I(2/5, 3/5) + 4/14 \times I(4/4, 0)$

  $+ 5/14 \times I(3/5, 2/5) = 0.694$

- for the attribute Windy, we have

  Info(Windy,T) = 0.892



Gain(X,T)

- Consider the quantity Gain(X,T) defined as

  Gain(X,T) = Info(T) - Info(X,T)

- This represents the difference between the information needed to identify an element of T and the information needed to identify an element of T after the value of attribute X has been obtained,
- i.e., this is the gain in information due to attribute *X*.

In the golfing example

- For the Outlook attribute the gain is:

  Gain(Outlook,T) = Info(T) - Info(Outlook,T)

  = 0.94 - 0.694 = 0.246

- For the attribute Windy, the gain is :

  Gain(Windy,T) = Info(T) - Info(Windy,T)

  = 0.94 - 0.892 = 0.048

- Thus, Outlook offers a greater informational gain than Windy.

- We can use this notion of gain to rank attributes and to build decision trees where at each node is located the attribute with greatest gain among the attributes not yet considered in the path from the root.

Other measurements

- Entropy

$$I(p) = -(p_1 \times \log(p_1) + p_2 \times \log(p_2) + .. + p_n \times \log(p_n))$$

- Misclassification cost

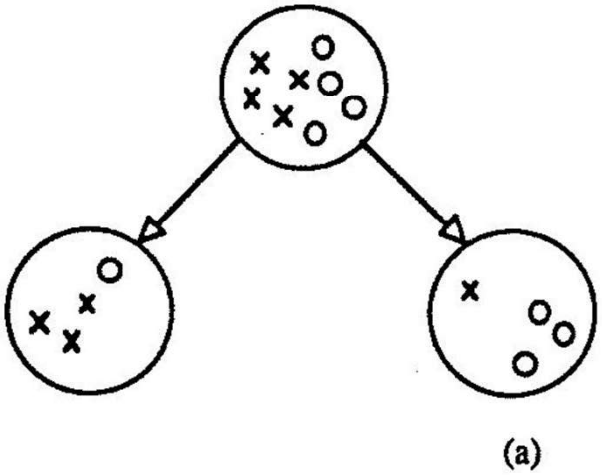
$$I(p) = 1 - \max_j p_j$$

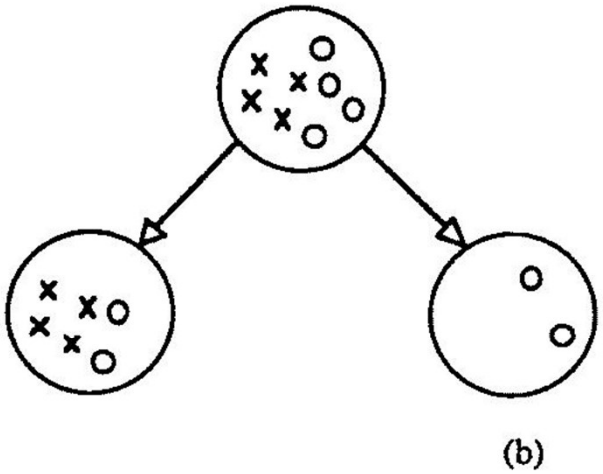
- Gini

$$I(p) = \Sigma_{i \neq j} p_i p_j = 1 - \Sigma_j p_j^2$$

Dat

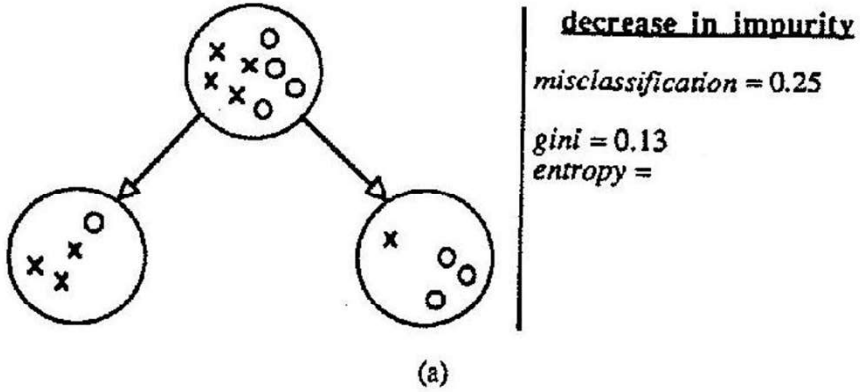
Example





- 8 examples from 2 classes (denoted by o and x in the diagrams).
- 2 attributes to consider (diagrams (a) and (b) respectively)

Misclassification in (a)



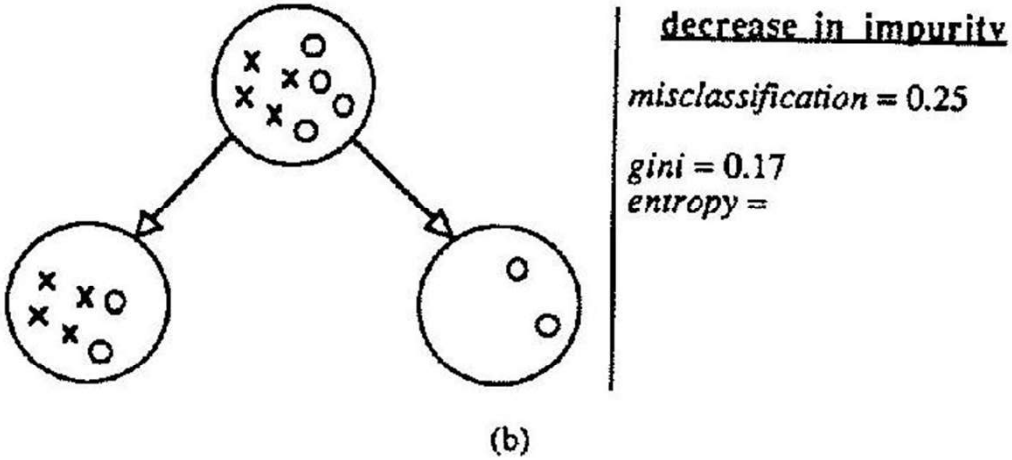
- Info(T) = 1 - 4/8 = ½

In case (a)

- $\{|T_L| / |T|\}$ Info(T_L) = (1 - ¾) × ½ = ⅛
- $\{|T_R| / |T|\}$ Info(T_R) = (1 - ¾) × ½ = ⅛

- Decrease in Impurity
  Gain (A,T) = Info(T) - Info(A,T) = ½ - ⅛ - ⅛ = ¼

Misclassification in (b)



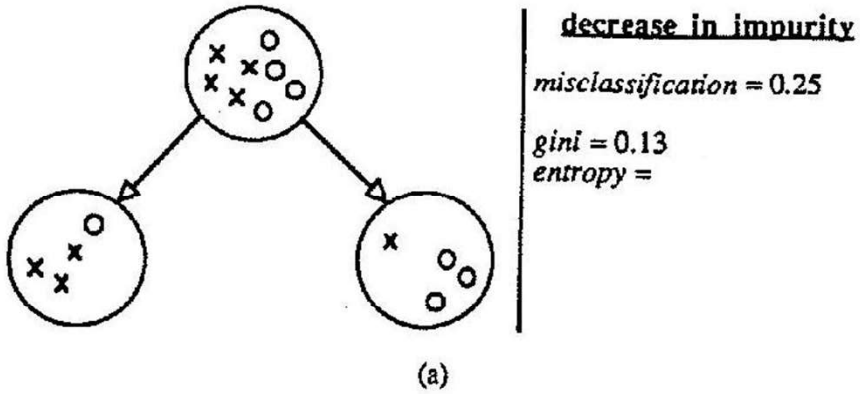
In case (b)

- $\{|T_L| / |T|\}$ Info(T_L) $= (1 - 4/6) \times 6/8 = \frac{1}{4}$
- $\{|T_R| / |T|\}$ Info(T_R) $= (1 - 2/2) \times 2/8 = 0$

- Decrease in Impurity

  Gain (B,T) = Info(T) - Info(B,T) = ½ - ¼ - 0 = ¼

Gini in (a)



- Info(T) = 1 - (½)^2 - (½)^2 = ½

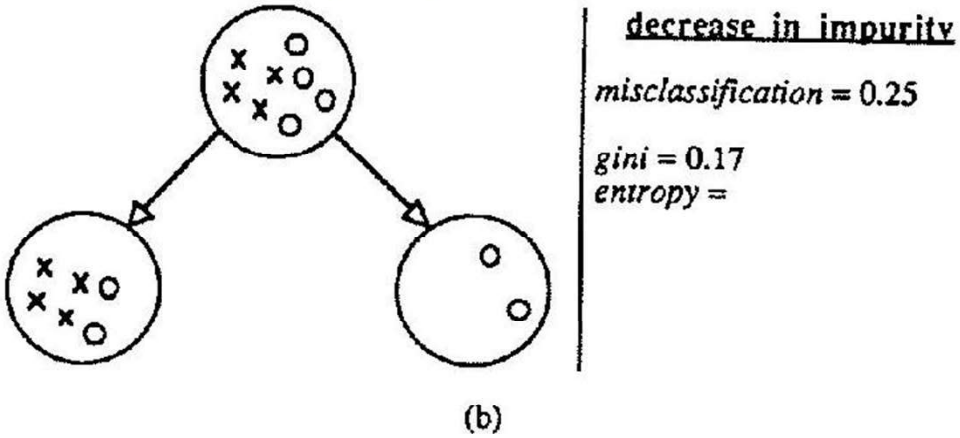
In case (a)

- {|T_L| / |T|} Info(T_L) = (1 - (¾)^2 - (¼)^2)×4/8 = 3/16
- {|T_R| / |T|} Info(T_R) = (1 - (¾)^2 - (¼)^2)×4/8 = 3/16

- Decrease in Impurity

  Gain (A, T) = Info(T) - Info(A,T) = ½ - 3/16 - 3/16 = 1/8

Gini in (b)



In case (b)

- $\{|T_L| / |T|\}\ \text{Info}(T_L) = (1 - (1/3)^2 - (2/3)^2) \times 3/4 = 1/3$
- $\{|T_R| / |T|\}\ \text{Info}(T_R) = (1 - 1^2 - 0^2) \times 1/4 = 0$

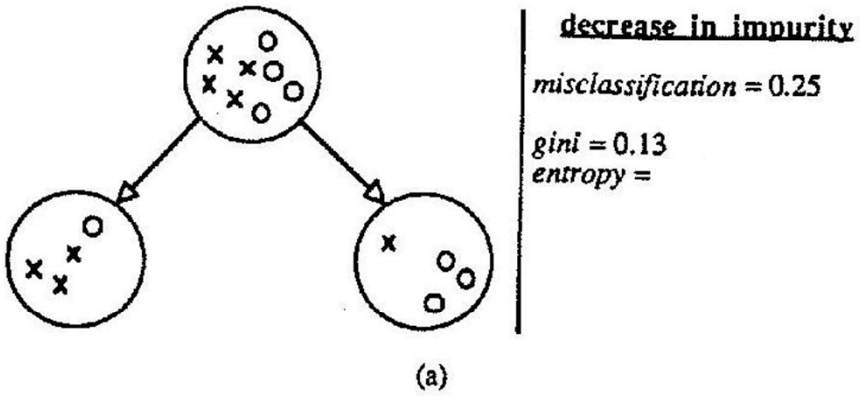
- Decrease in Impurity

  $\text{Gain}(B,T) = \text{Info}(T) - \text{Info}(B,T) = 1/2 - 1/3 - 0 = 1/6$

Entropy in (a)

- Info(T) = - ½ log ½ – ½ log ½
  = 1

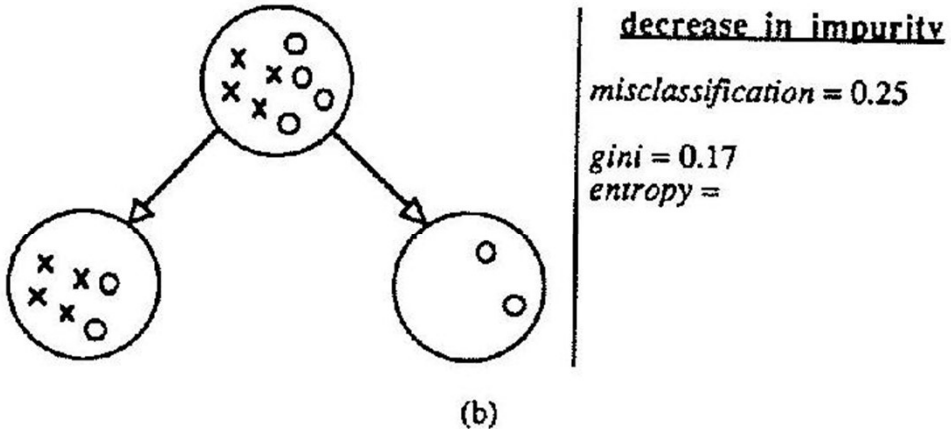
In case (a)



- $\{|T_L| / |T|\}$ Info(T_L) = (- ¾ log ¾ – ¼ log ¼)× ½ = 0.41
- $\{|T_R| / |T|\}$ Info(T_R) = (- ¾ log ¾ - ¼ log ¼) × ½ = 0.41

- Decrease in Impurity
  Gain (A,T) = Info(T) - Info(A,T) = 1 - 0.82 = 0.18

Entropy in (b)



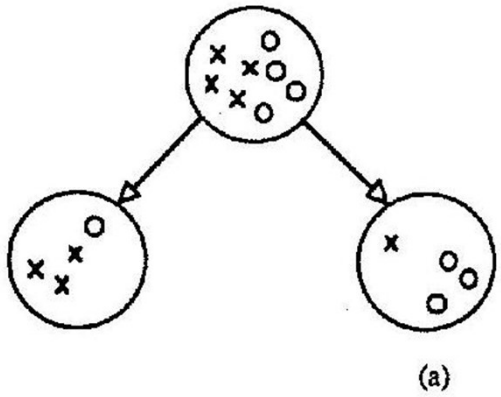
In case (b)

- $\{|T_L|/|T|\}$ Info$(T_L) = (- \frac{1}{3} \log \frac{1}{3} - \frac{2}{3} \log \frac{2}{3}) \times \frac{3}{4} = 0.73$
- $\{|T_R|/|T|\}$ Info$(T_R) = (- 1 \log 1 - 0 \log 0) \times \frac{1}{4} = 0$

- Decrease in Impurity

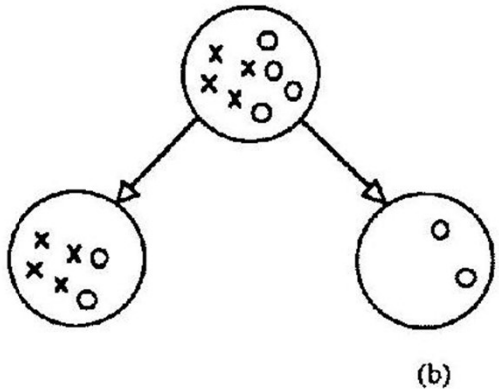
Gain (B,T) = Info(T) - Info(B,T) = 1 - 0.73 - 0 = 0.27

Decrease in impurity



misclassification
gini
entropy

(a) 0.25 ✓
(a) 0.13
(a) 0.18



(b) 0.25 ✓
(b) 0.17 ✓
(b) 0.27 ✓