Classification: Decision Trees

Dr. Faisal Kamiran

Tree Induction

- Greedy strategy.
    - Split the records based on an attribute test that optimizes certain criterion.

- 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

Tree Induction

- Greedy strategy.
  - Split the records based on an attribute test that optimizes certain criterion.

- 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
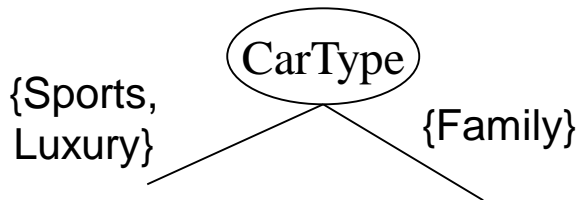
How to Specify Test Condition?

- Depends on attribute types
  - Nominal
  - Ordinal
  - Continuous

- Depends on number of ways to split
  - 2-way split
  - Multi-way split
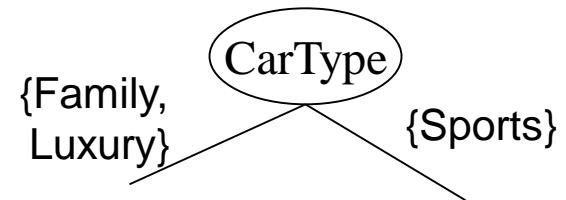
Splitting Based on Nominal Attributes

- Multi-way split: Use as many partitions as distinct values.
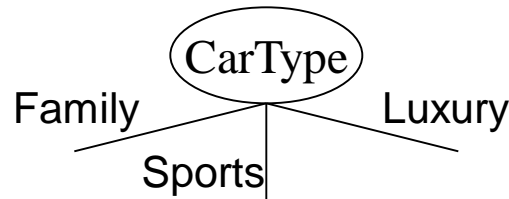
```
        CarType
Family  /  |  \  Luxury
          Sports
```

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

```
     CarType                              CarType
{Sports,  /  \  {Family}      OR     {Family,  /  \  {Sports}
 Luxury}                              Luxury}
```
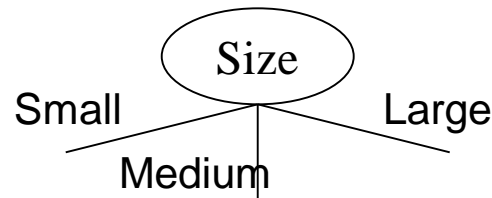
Splitting Based on Ordinal Attributes

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

```
        ( Size )
Small  /   |   \  Large
          Medium
```

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

```
{Small,   ( Size )              OR      {Medium,  ( Size )
Medium}  /       \  {Large}             Large}  /       \  {Small}
```
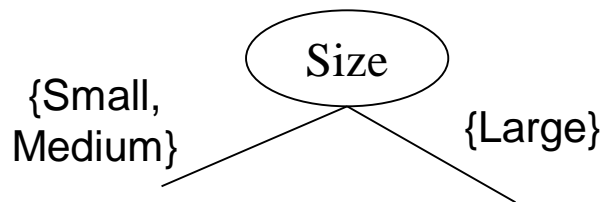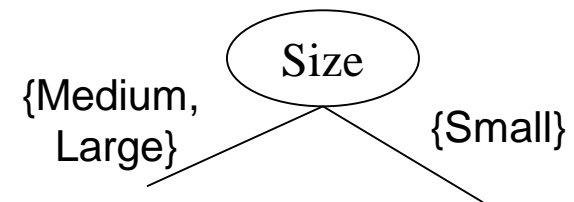
● What about this split?
```
{Small,   ( Size )
Large}   /       \  {Medium}
```
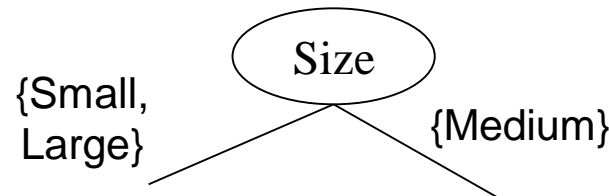
Splitting Based on Continuous Attributes

- 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

Taxable Income > 80K?

Yes   No

(i) Binary split

Taxable Income?

< 10K   [10K,25K)   [25K,50K)   [50K,80K)   > 80K
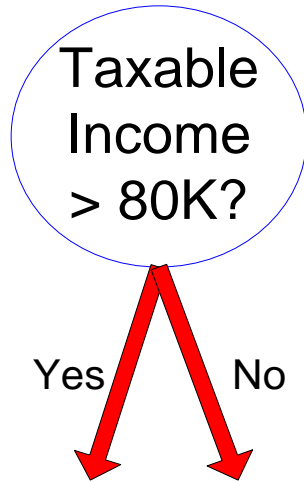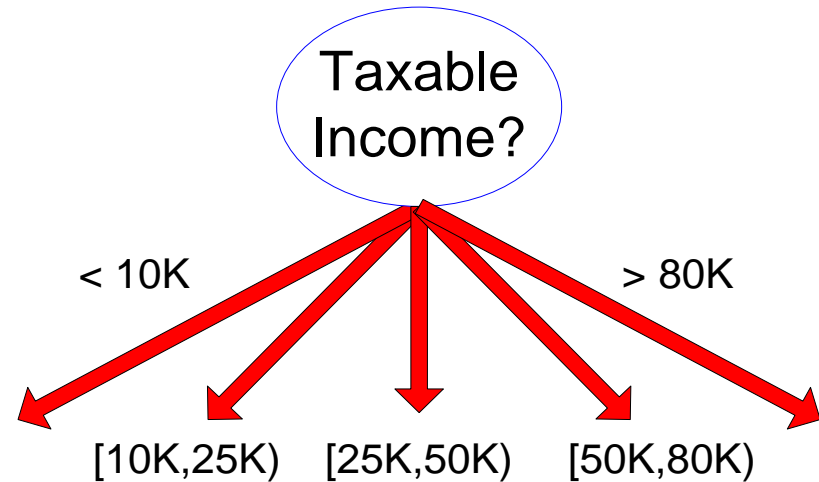
(ii) Multi-way split

Tree Induction

- Greedy strategy.
  - Split the records based on an attribute test that optimizes certain criterion.

- 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 determine the Best Split

Before Splitting: 10 records of class 0,
10 records of class 1
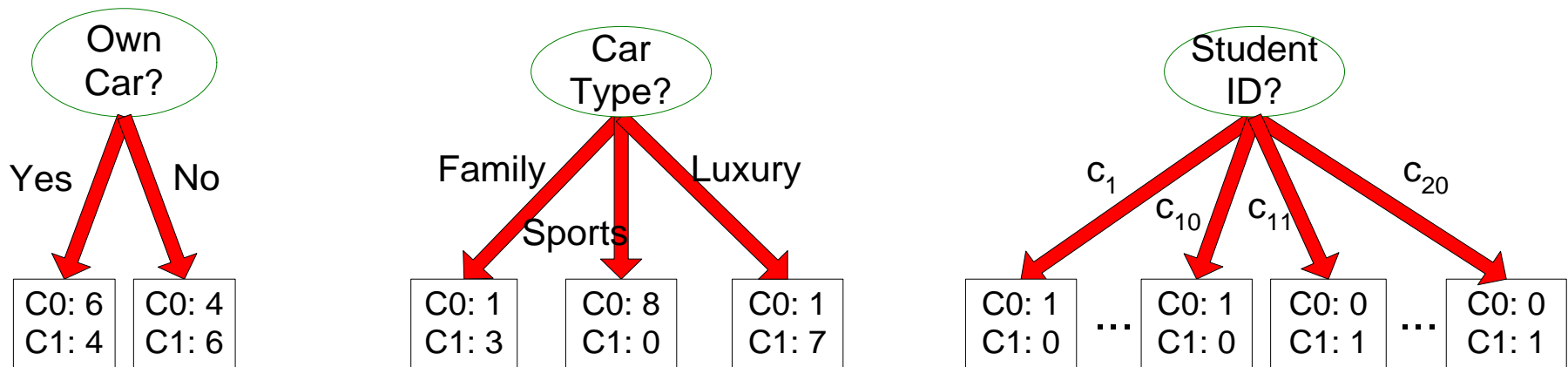
Which test condition is the best?

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

Measures of Node Impurity

- Gini Index

- Entropy

- Misclassification error

How to Find the Best Split

Before Splitting:

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

→ **M0**

A?

Yes — **Node N1**   No — **Node N2**

C0	**N10**
C1	**N11**

C0	**N20**
C1	**N21**

M1   **M2**

M12

B?

Yes — **Node N3**   No — **Node N4**

C0	**N30**
C1	**N31**

C0	**N40**
C1	**N41**

M3   **M4**

M34

Gain = M0 – M12 vs M0 – M34

Measure of Impurity: GINI

- Gini Index for a given node t :

$$GINI(t) = 1 - \sum_{j} [p(j \mid t)]^2$$

(NOTE: $p(j \mid t)$ is the relative frequency of class j at node t).

- – Maximum ($1 - 1/n_c$) when records are equally distributed among all classes, implying least interesting information
- – Minimum (0.0) when all records belong to one class, implying most interesting information

C1	0
C2	6
Gini=0.000	

C1	1
C2	5
Gini=0.278	

C1	2
C2	4
Gini=0.444	

C1	3
C2	3
Gini=0.500	

Examples for computing GINI

$$GINI(t) = 1 - \sum_{j} [p(j \mid t)]^2$$

C1	0
C2	6

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

Gini = 1 – P(C1)^2 – P(C2)^2 = 1 – 0 – 1 = 0

C1	1
C2	5

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

Gini = 1 – (1/6)^2 – (5/6)^2 = 0.278

C1	2
C2	4

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

Gini = 1 – (2/6)^2 – (4/6)^2 = 0.444

Splitting Based on GINI

- Used in CART, SLIQ, SPRINT.
- When a node p is split into k partitions (children), the quality of split is computed as,
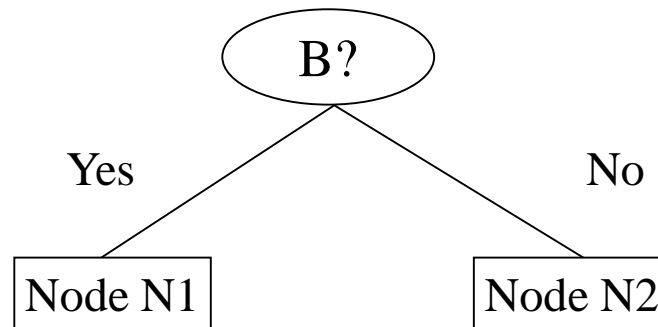
$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

where, n_i = number of records at child i,

n = number of records at node p.

$$GINI(i) = 1 - \sum_{j} [p(j \mid i)]^2$$

Binary Attributes: Computing GINI Index

- Splits into two partitions
- Effect of Weighing partitions:
  – Larger and Purer Partitions are sought for.

	Parent
C1	6
C2	6

Gain (B) = 0.5

B?

Yes   No

Node N1   Node N2

Gini(N1)
= 1 – (5/7)² – (2/7)²
= 0.408

	N1	N2
C1	5	1
C2	2	4

Gini (Split) = 0.371

Gini(Children)
= 7/12 * 0.408 +
**   5/12 * 0.32**
= 0.371

Gini(N2)
= 1 – (1/5)² – (4/5)²
= 0.32

Gain (B)=0.5 – 0.371 = 0.129

Categorical Attributes: Computing Gini Index

- For each distinct value, gather counts for each class in the dataset
- Use the count matrix to make decisions

Multi-way split

CarType		
Family	**Sports**	**Luxury**
1	2	1
4	1	1

	CarType		
C1	1	2	1
C2	4	1	1
Gini	0.393		

Two-way split
(find best partition of values)

	CarType	
	{Sports, Luxury}	**{Family}**
C1	3	1
C2	2	4
Gini	0.400	

	CarType	
	{Sports}	**{Family, Luxury}**
C1	2	2
C2	1	5
Gini	0.419	