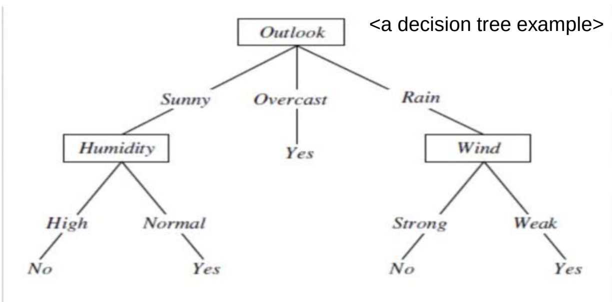
Lecture 6: Decision Tree

Decision Tree

■ A **decision tree** is a decision support tool that uses a tree of decision

■ **Representation**
- ○ Each internal node tests an attribute variable
- ○ Each branch corresponds to an attribute value
- ○ Each leaf node assigns a classification value



<a decision tree example>

An Example

- Do you want to play tennis ?

- Suppose that you have the following rule in your brain to answer to this question?

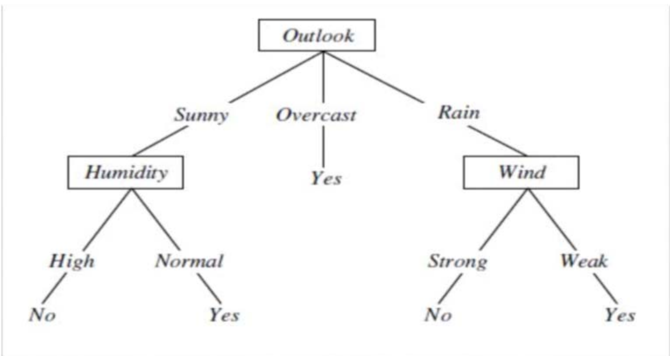
- PlayTennis Rule in your brain :

if (*Outlook =Sunny* ∧ *Humidity = Normal*)
∨ (*Outlook = Overcast*)
∨ (*Outlook =Rain* ∧ *Wind = Weak*) **then** Yes
else No

3

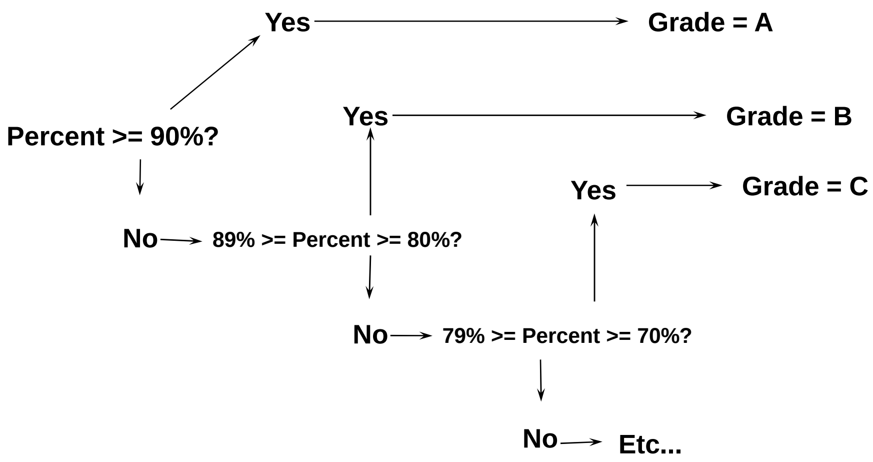
An Example of Decision Tree

if (*Outlook =Sunny* ∧ *Humidity = Normal*)
∨ (*Outlook = Overcast*)
∨ (*Outlook =Rain* ∧ *Wind = Weak*) **then** Yes
else No



4

Another Example - Grading



5

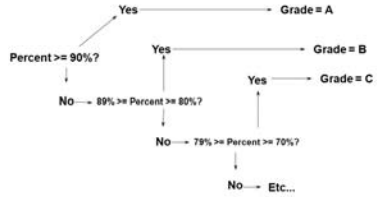
Introduction

- Decision Trees
  - Powerful/popular for classification & prediction
  - Useful to explore data to gain insight into relationships of a large number of attribute variables to a target(classification) variable
- You may often use mental decision trees in practice
  - Remember PlayTennis example!

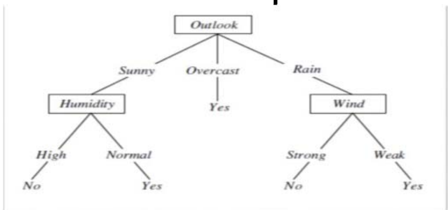
6

Decision Tree Types

- Binary decision trees – only two choices in each split.



- N-way or ternary decision trees – three or more choices in at least one of its splits (3-way, 4-way, etc.)



7

Decision Tree – More details

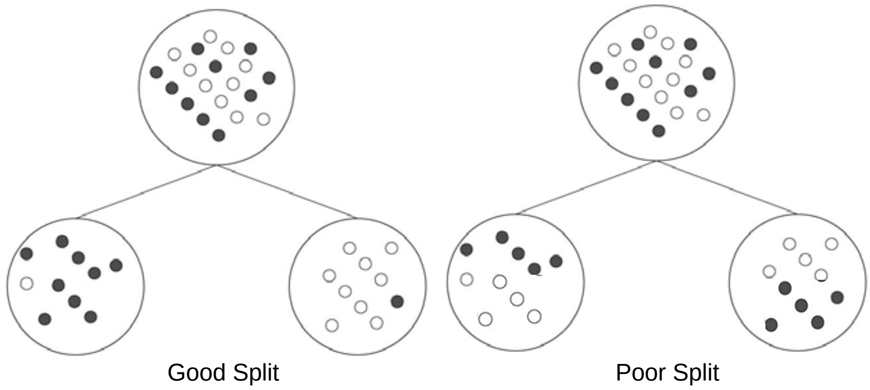
- A tree structure that can be used to split a large set of records into successively smaller sets of records by applying a sequence of simple decision rules
- A decision tree model consists of a set of split rules for dividing a large heterogeneous population into smaller, more homogeneous groups with respect to a particular target variable

8

Decision Tree Splits

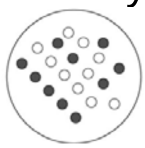
- The best split at root or child nodes is defined as one that does the best job of separating the data into groups each of which is homogeneous
  - Homogeneous means that each data has the same target value as the other
- Just split data according to the above “best split” rule !

Example: Good & Poor Splits

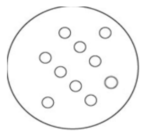


Split Criteria

- The measure used to evaluate how good a potential split is **purity!**
- If a data group contains several classes(several target values), then we say it is impure



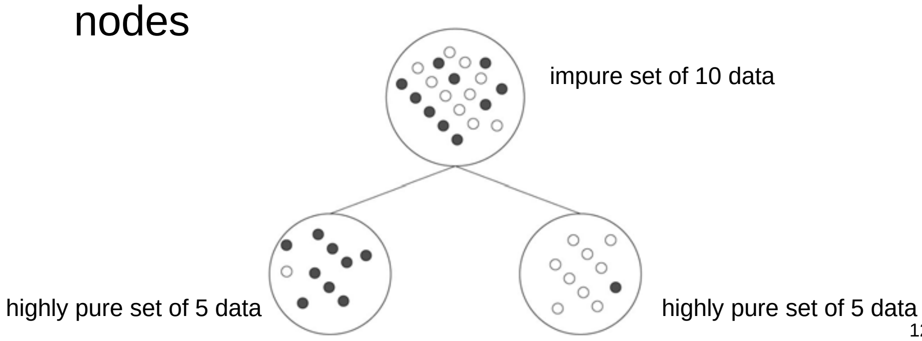
- If a data group contains one class(one target value), then we say it is pure



11

Split Criteria

- The best split is one that increases purity of the sub-sets by the greatest amount
- A good split also creates nodes of similar size or at least does not create very small nodes



12

Impurity (or Diversity) Measures

- Impurity Measures for Choosing Best Split :
  - **Information Gain** based on Entropy
  - **Gini** (population diversity)
  - Information Gain Ratio (as a simple variation of Information Gain)
  - Chi-square Test based (on chi-square distribution in Statistics)

We will only explore Information Gain and Gini in this class !

13

Information Gain

■ Entropy

To measure degree of impurity

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

where p_j values of probability of class j

■ Information gain

○ To compare the difference of impurity degrees between an original data set S and its split subsets S_v

$$Gain(S,\ A) = Entropy\ (S) - \sum_{\upsilon \in Values\ (A)} \frac{|S_\upsilon|}{|S|} Entropy\ (S_\upsilon)$$

14

Information Gain: Example

■ Data Set

Target Variable

Gender	Car Owner Ship	Travel Cost($)/km	Cost Income Level	Transportation
Male	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	1	Cheap	Medium	Train
Female	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Male	0	Standard	Medium	Train
Female	1	Standard	Medium	Train
Female	1	Expensive	High	Car
Male	2	Expensive	Medium	Car
Female	2	Expensive	High	Car

4B,3C,3T

■ Data Set impurity

$$Entropy = -0.4\log(0.4) - 0.3\log(0.3) - 0.3\log(0.3) = 1.571$$

Gender	Class
Male	Bus
Male	Bus
Male	Bus
Male	Car
Male	Train

$$Entropy = -\frac{3}{5}\log(\frac{3}{5}) - \frac{1}{5}\log(\frac{1}{5}) - \frac{1}{5}\log(\frac{1}{5}) = 1.371$$

3B,1C,1T

Gender	Class
Female	Bus
Female	Car
Female	Car
Female	Train
Female	Train

$$Entropy = -\frac{1}{5}\log(\frac{1}{5}) - \frac{2}{5}\log(\frac{2}{5}) - \frac{2}{5}\log(\frac{2}{5}) = 1.522$$

1B,2C,2T

■ **Information Gain of attribute *Gender***

1.574 - (5/10 * 1.371 + 5/10 * 1.522)

= 0.125

15

15

Maximum Info gain is to be root node

Information Gain: Example

■ ***Gender***

Gender	Class
Male	Bus
Male	Bus
Male	Bus
Male	Car
Male	Train

- 3B,1C,1T
- Entropy = 1.522

Gender	Class
Female	Bus
Female	Car
Female	Car
Female	Train
Female	Train

- 1B,2C,2T
- Entropy = 1.371

Gain = 0.125

■ ***Car Ownership***

Car Owner	Class
0	Bus
0	Bus
0	Train

- 2B, 1T
- Entropy = 0.918

Car Owner	Class
1	Bus
1	Bus
1	Car
1	Train
1	Train

- 1B,1C,2T
- Entropy = 1.522

Car Owner	Class
2	Car
2	Car

- 1C
- Entropy = 0.000

Gain = 0.534

■ ***Travel Cost***

Travel Cost	Class
Cheap	Bus
Cheap	Bus
Cheap	Bus
Cheap	Bus
Cheap	Train

- 4B, 1T
- Entropy = 0.722

Travel Cost	Class
Expensive	Car
Expensive	Car
Expensive	Car

- 3C
- Entropy = 0.000

Travel Cost	Class
Standard	Train
Standard	Train

- 2T
- Entropy = 0.00

Gain = 1.210

■ ***Income Level***

Income Level	Class
High	Bus
High	Bus

- 2B
- Entropy = 0.000

Income Level	Class
Low	Bus
Low	Bus

- 2B
- Entropy = 0.000

Income Level	Class
Medium	Bus
Medium	Bus
Medium	Car
Medium	Train
Medium	Train
Medium	Train

- 2B,1C,3T
- Entropy = 1.459

Gain = 0.695

16

16

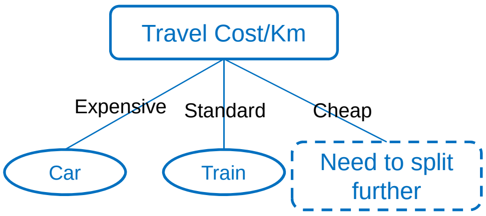
Information Gain: Example

■ Split Data Set based on Optimum attribute(Travel Cost)

Gender	Car Owner Ship	Travel Cost($)/km	Cost Income Level	Transportation
Male	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Female	1	Cheap	Medium	Train
Female	0	Cheap	Low	Bus
Male	1	Cheap	Medium	Bus
Male	0	Standard	Medium	Train
Female	1	Standard	Medium	Train
Female	1	Expensive	High	Car
Male	2	Expensive	Medium	Car
Female	2	Expensive	High	Car

Pure Classes !

■ Pure class is assigned into leaf node



17

Information Gain: Example

■ Second iteration

- Attribute *Travel cost* is not needed any more so it is removed
- In the same way as the previous, we repeat the computations of Impurity and Information Gain for each of the three attributes

Target Variable

Gender	Car Owner Ship	Cost Income Level	Transportation
Female	0	Low	Bus
Male	0	Low	Bus
Male	1	Medium	Bus
Male	1	Medium	Bus
Female	1	Medium	Train

4B,1T

$$Entropy = -\frac{1}{5}\log(\frac{1}{5}) - \frac{4}{5}\log(\frac{4}{5}) = 0.722$$

18

Information Gain: Example

- *Gender*

Gender	Class
Male	Bus
Male	Bus
Male	Bus

- 3B
- Entropy = 0.000

Gender	Class
Female	Bus
Female	Train

- 1B,1T
- Entropy = 1.000

Gain = 0.322

Maximum Info gain is to be next node

- *Car Ownership*

Car Owner	Class
0	Bus
0	Bus

- 2B
- Entropy = 0.000

Car Owner	Class
1	Bus
1	Bus
1	Train

- 2B, 1T
- Entropy = 0.918

Gain = 0.171

- *Income Level*

Income Level	Class
Low	Bus
Low	Bus

- 2B
- Entropy = 0.000

Income Level	Class
Medium	Bus
Medium	Bus
Medium	Train

- 2B, 3T
- Entropy = 1.459

Gain = 0.171

19

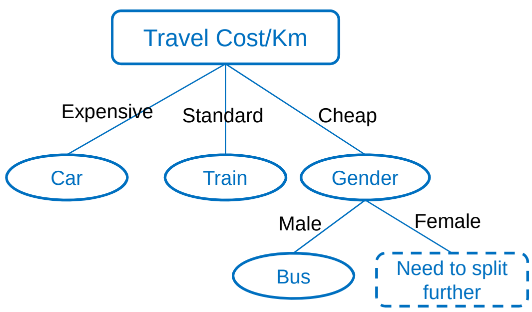
Information Gain: Example

- **Split Data Set based on Optimum attribute(Gender)**

Gender	Car Owner Ship	Cost Income Level	Transportation
Female	0	Low	Bus
Male	0	Low	Bus
Male	1	Medium	Bus
Male	1	Medium	Bus
Female	1	Medium	Train

Pure Class !

- **Pure class is assigned into leaf node**



20

Information Gain: Example

■ Third iteration

• Attribute *Gender* cost is not needed any more so it is removed

• In the same way as the previous, we repeat the computations of Impurity and Information Gain for each of the two attributes

Car Owner Ship	Cost Income Level	Transportation
0	Low	Bus
1	Medium	Train

21

Information Gain: Example

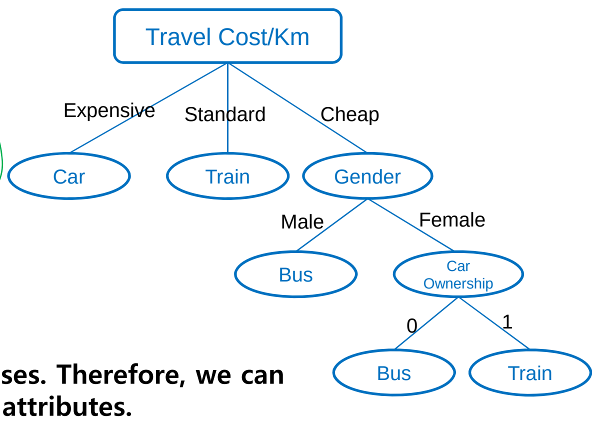
■ Split Data Set based on each attribute

Car Owner Ship	Cost Income Level	Transportation
0	Low	Bus
1	Medium	Train

Pure Classes !

• Both tuples are pure classes. Therefore, we can use either one of the two attributes.

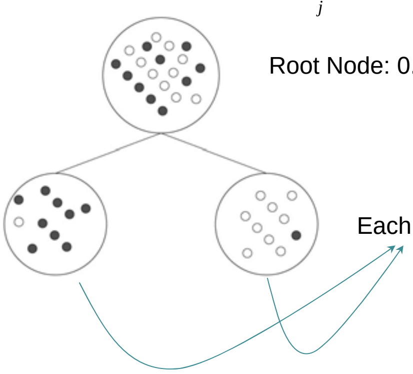
■ Final version decision tree



22

Gini Purity (Population Diversity)

- The Gini measure of a node is the sum of the squares of the proportions of the classes: $Gini = \sum_j p_j^{\,2}$



Root Node: $0.5^2 + 0.5^2 = 0.5$ (evenly balanced: impure)

Each Leaf Node: $0.1^2 + 0.9^2 = 0.82$ (close to pure)

23

Gini Purity (Population Diversity)

- **Just replacing Entropy() with Gini() is enough to determine the best splitting attribute for a decision tree node**

$$Gain(S,A) = \sum_{\upsilon \in Values(A)} \frac{|S_\upsilon|}{|S|} Gini(S_\upsilon) - Gini(S)$$

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

24

Decision Tree Advantages

1. Easy to understand
2. Mapped nicely to a set of business rules
3. Applied to a variety of real classification and prediction problems
4. Make no prior assumptions about the data
5. Able to process both numerical and categorial attributes in data

25

Decision Tree Disadvantages

1. Target(Classification) attribute must be categorial
2. Limited to one target attribute (one class)
3. Decision tree algorithms are sometimes unstable (similar to local search !)
4. Trees created from numerical-attribute datasets can be very complex

26

End of Decision Tree

27