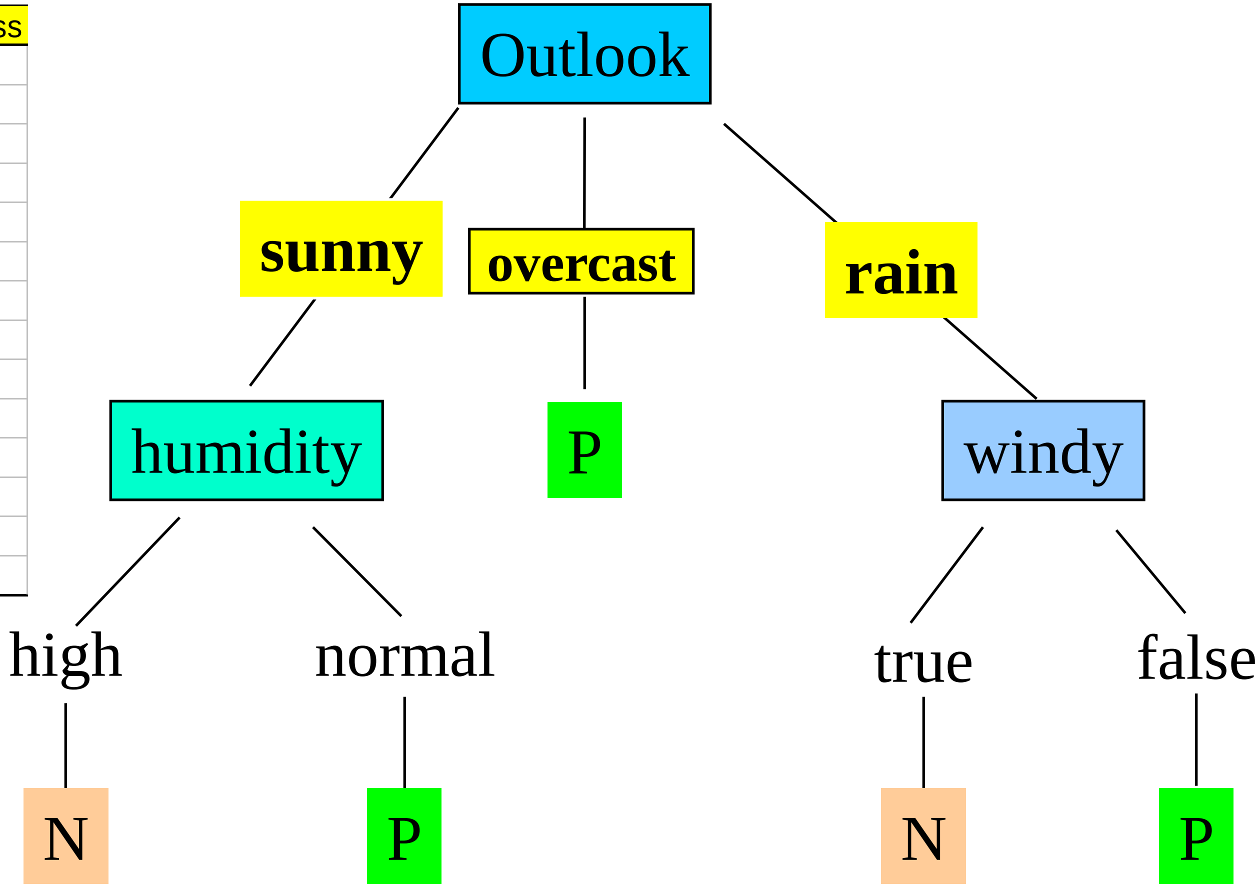
Chapter 6
Decision Trees

An Example

Outlook	Tempreature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N



Another Example - Grades

Percent >= 90%?

Yes → Grade = A

No → 89% >= Percent >= 80%?

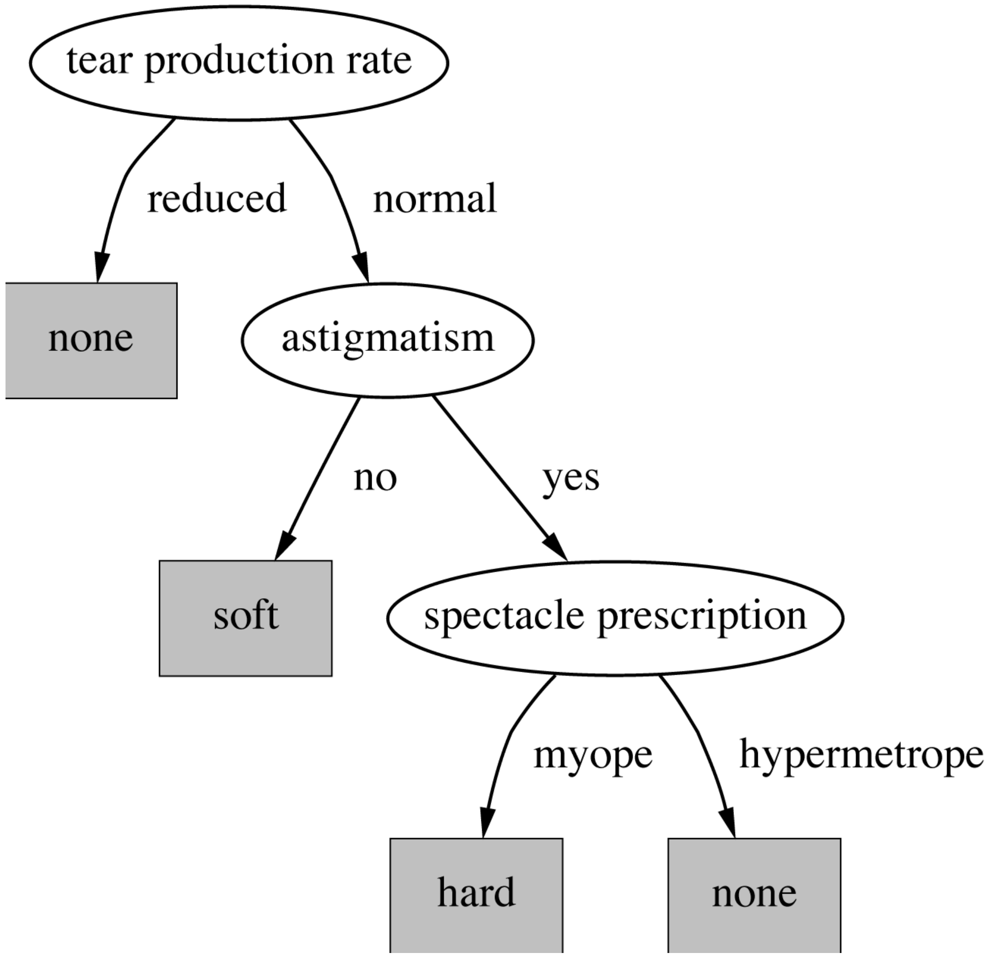
Yes → Grade = B

No → 79% >= Percent >= 70%?

Yes → Grade = C

No → Etc...

Yet Another Example



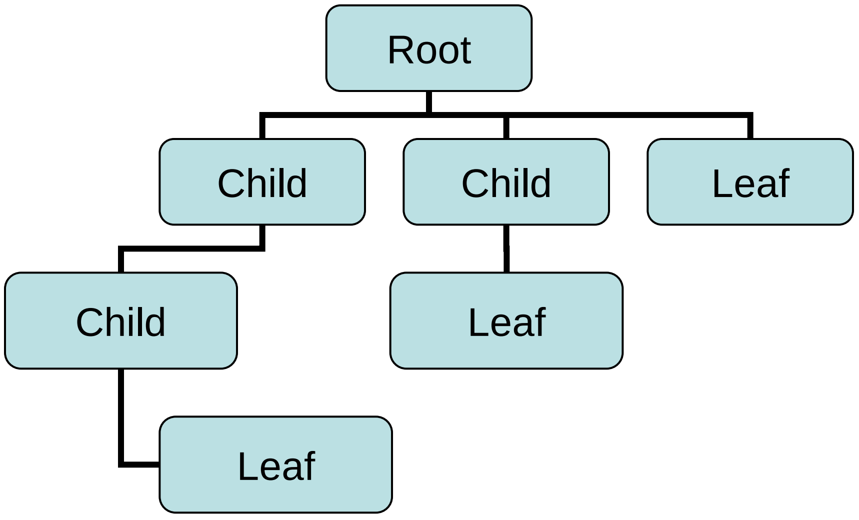
Yet Another Example

- English Rules (for example):

```
If tear production rate = reduced then recommendation = none.
If age = young and astigmatic = no and tear production rate = normal
   then recommendation = soft
If age = pre-presbyopic and astigmatic = no and tear production
   rate = normal then recommendation = soft
If age = presbyopic and spectacle prescription = myope and
   astigmatic = no then recommendation = none
If spectacle prescription = hypermetrope and astigmatic = no and
   tear production rate = normal then recommendation = soft
If spectacle prescription = myope and astigmatic = yes and
   tear production rate = normal then recommendation = hard
If age = young and astigmatic = yes and tear production rate =
    normal
   then recommendation = hard
If age = pre-presbyopic and spectacle prescription = hypermetrope
   and astigmatic = yes then recommendation = none
If age = presbyopic and spectacle prescription = hypermetrope
   and astigmatic = yes then recommendation = none
```

Decision Tree Template

- Drawn top-to-bottom or left-to-right
- Top (or left-most) node = **Root Node**
- Descendent node(s) = **Child Node(s)**
- Bottom (or right-most) node(s) = **Leaf Node(s)**
- Unique path from root to each leaf = **Rule**



Introduction

- Decision Trees
  - Powerful/popular for classification & prediction
  - Represent rules
    - Rules can be expressed in English
      - IF Age <=43 & Sex = Male
        & Credit Card Insurance = No
        THEN Life Insurance Promotion = No
    - Rules can be expressed using SQL for query
  - Useful to explore data to gain insight into relationships of a large number of candidate input variables to a target (output) variable
- You use mental decision trees often!
- Game: “I’m thinking of…” “Is it …?”

Decision Tree – What is it?

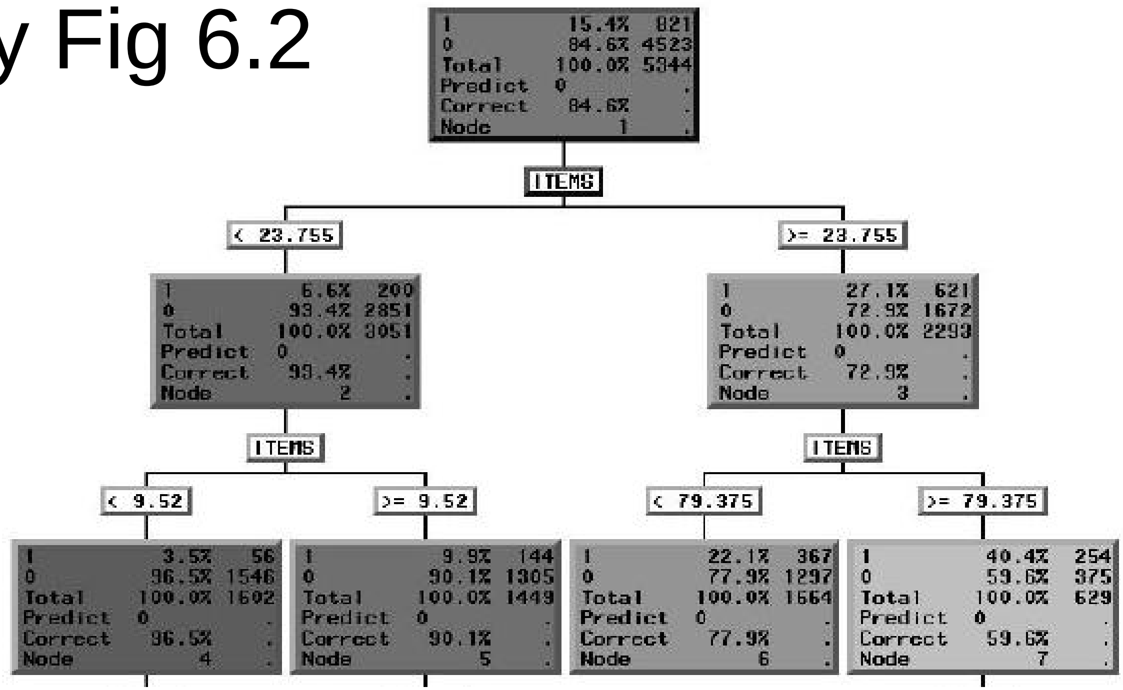
- A structure that can be used to divide up a large collection of records into successively smaller sets of records by applying a sequence of simple decision rules
- A decision tree model consists of a set of rules for dividing a large heterogeneous population into smaller, more homogeneous groups with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

Decision Tree Types

- Binary trees – only two choices in each split. Can be non-uniform (uneven) in depth
- N-way trees or ternary trees – three or more choices in at least one of its splits (3-way, 4-way, etc.)

Scoring

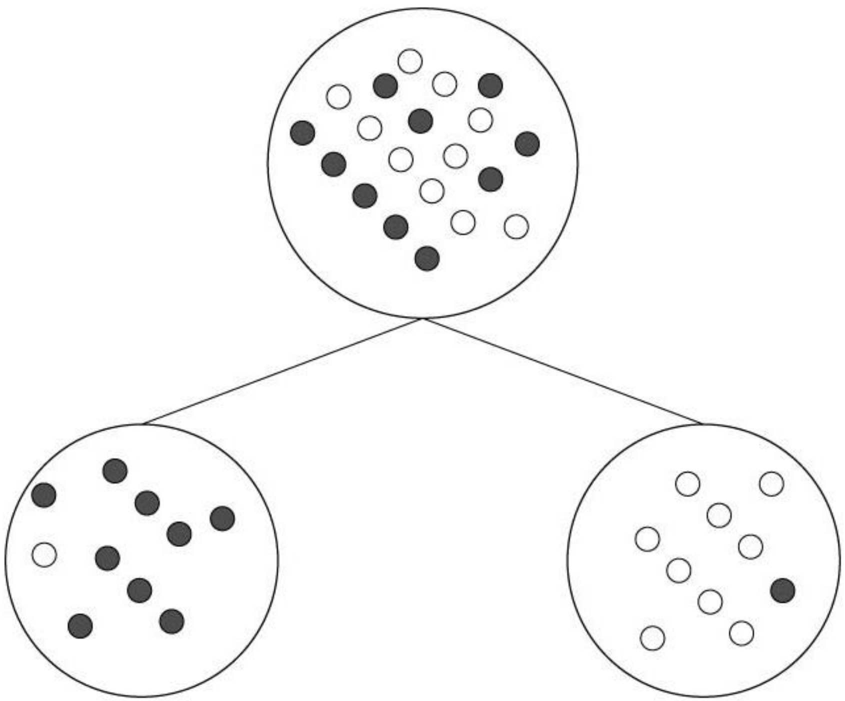
- Often it is useful to show the proportion of the data in each of the desired classes
- Clarify Fig 6.2



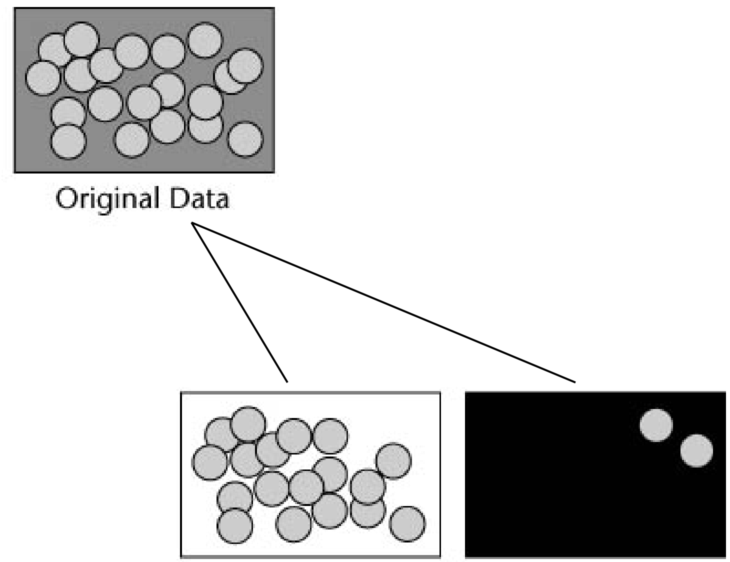
Decision Tree Splits (Growth)

- The best split at root or child nodes is defined as one that does the best job of separating the data into groups where a single class predominates in each group
  - Example: US Population data input categorical variables/attributes include:
    - Zip code
    - Gender
    - Age
  - Split the above according to the above “best split” rule

Example: Good & Poor Splits



Good Split



Poor Split

Split Criteria

- The best split is defined as one that does the best job of separating the data into groups where a single class predominates in each group
- Measure used to evaluate a potential split is **purity**
  - 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

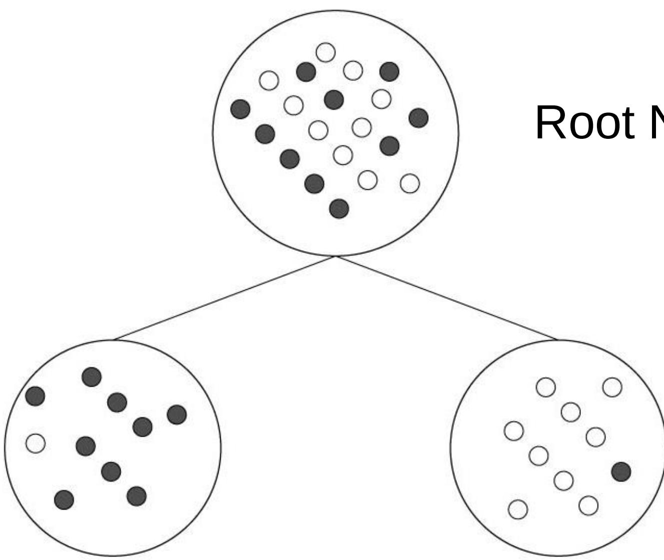
Tests for Choosing Best Split

- Purity (Diversity) Measures:
  - Gini (population diversity)
  - Entropy (information gain)
  - Information Gain Ratio
  - Chi-square Test

We will only explore Gini in class

Gini (Population Diversity)

- The Gini measure of a node is the sum of the squares of the proportions of the classes.

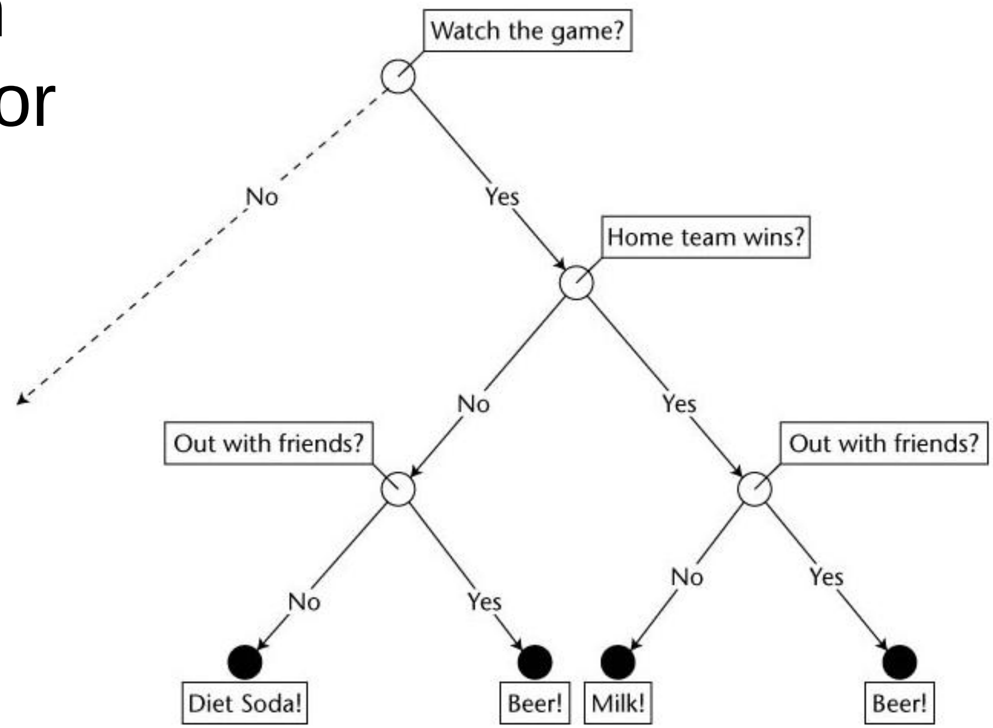


Root Node: 0.5^2 + 0.5^2 = 0.5 (even balance)

Leaf Nodes: 0.1^2 + 0.9^2 = 0.82 (close to pure)

Pruning

- Decision Trees can often be simplified or pruned:
  - CART
  - C5
  - Stability-based



We will not cover these in detail

Decision Tree Advantages

1. Easy to understand
2. Map nicely to a set of business rules
3. Applied to real problems
4. Make no prior assumptions about the data
5. Able to process both numerical and categorical data

Decision Tree Disadvantages

1. Output attribute must be categorical
2. Limited to one output attribute
3. Decision tree algorithms are unstable
4. Trees created from numeric datasets can be complex

Alternative Representations

- Box Diagram
- Tree Ring Diagram
- Decision Table
- Supplementary Material

End of Chapter 6