Review

FEATURE SELECTION METHODS

Feature Selection Methods

❑Filter Methods

- Measure feature importance via univariate statistical methods
  - Information Gain (or mutual information)
  - Chi-squared Test (For categorical features)
  - Fisher Score
  - Correlation Coefficient (Features should have high correlation with label but should be uncorrelated among themselves)
  - Variance Threshold (There should be some variance for a feature i.e., feature with same value for all records has zero-variance thus, should be removed).

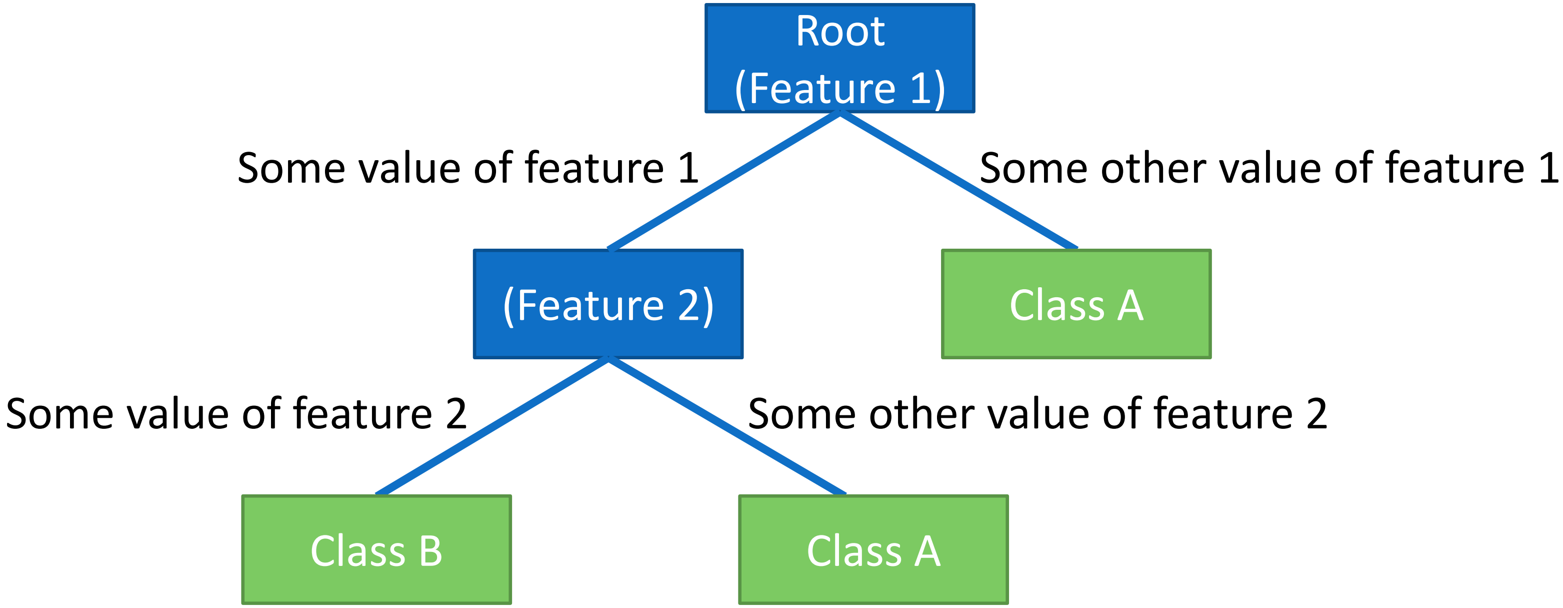
❑Wrapper Methods

- Search all possible subsets of features and assess the quality by training and evaluating a classifier
  - Forward Feature Selection
  - Backward Feature Elimination
  - Recursive Feature Elimination (Recursively eliminate least important feature until desired number of features are selected)

Decision Tree

Decision Tree

- A decision tree is **a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks**.
- It has a hierarchical, tree structure, which consists of a **root node**, **branches**, **internal nodes** and **leaf nodes**.
- Root node and internal node has attributes.
- Leaf nodes have class labels.



Which attribute should be the root node?

The attribute that has the maximum predictability power with respect to the class label!

Let’s use Information Gain

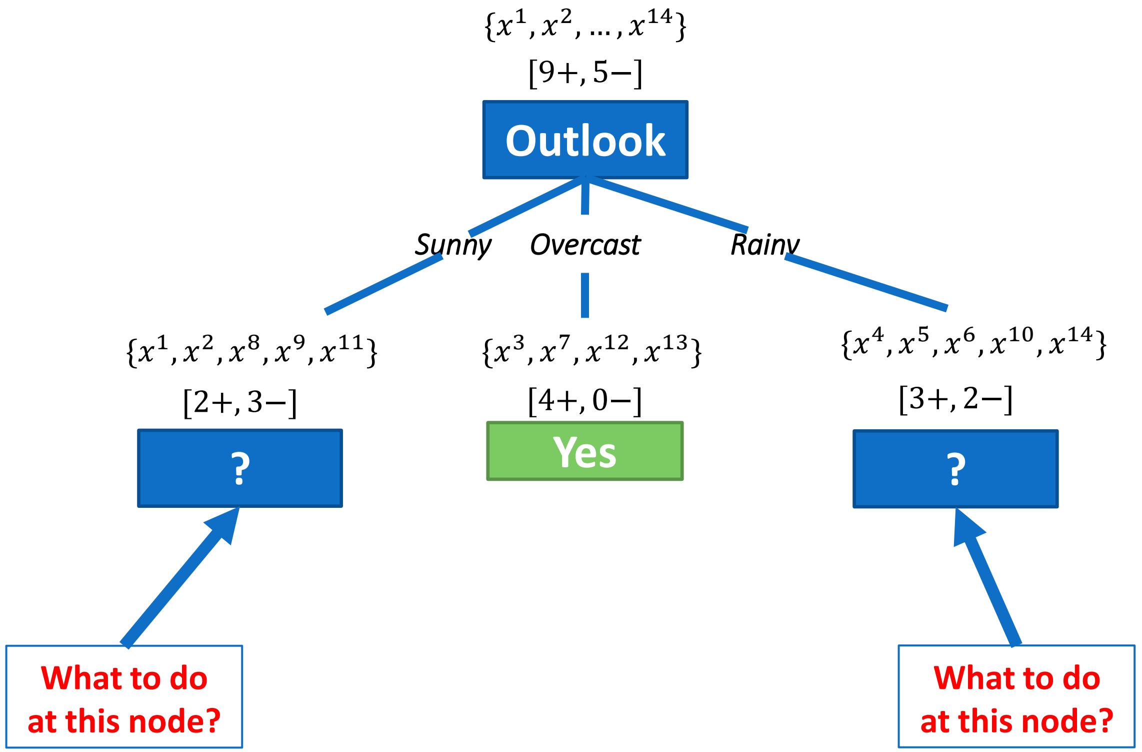
Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rainy	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

$\boldsymbol{Gain\ (S, Outlook) = 0.2464}$

$\boldsymbol{Gain\ (S, Temp) = 0.0289}$

$\boldsymbol{Gain\ (S, Wind) = 0.048}$

$\boldsymbol{Gain\ (S, Humidity) = 0.151}$



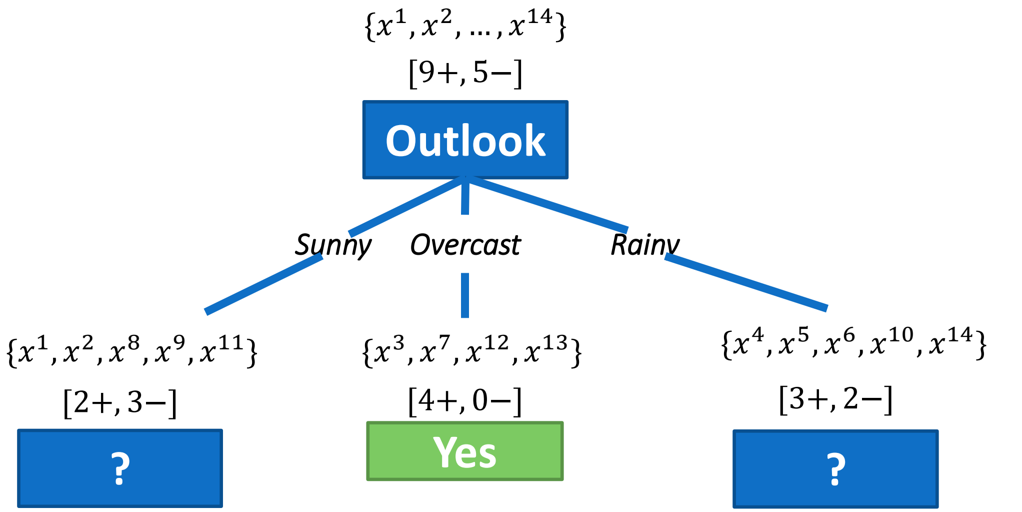
Let's use Information Gain

Make a subset of dataset where outlook=sunny and repeat the process!

Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes

$$Entropy\ (S) = 0.94$$

$$Entropy\ (S_{sunny}) = -\frac{2}{5} \times log_2 \frac{2}{5} - \frac{3}{5} \times log_2 \frac{3}{5} = 0.971$$



Let's use Information Gain

Make a subset of dataset where outlook=sunny and repeat the process!

Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes

$$Entropy\ (S) = 0.94$$

$$Entropy\ (S_{sunny}) = -\frac{2}{5} \times log_2 \frac{2}{5} - \frac{3}{5} \times log_2 \frac{3}{5} = 0.971$$

$$Gain\ (S_{sunny}, Temp) = 0.57$$

Attribute: Temp
Values (Temp) = Hot, Mild, Cool

$$S_{Hot} = [0+, 2-] \qquad Entropy\ (S_{Hot}) = 0$$

$$S_{Mild} = [1+, 1-] \qquad Entropy\ (S_{Mild}) = 1$$

$$S_{Cool} = [1+, 0-] \qquad Entropy\ (S_{Cool}) = 0$$

$$Gain\ (S_{sunny}, Temp) = Entropy\ (S_{sunny}) - \sum_{v \in \{Hot, Mild, Cool\}} \frac{|S_v|}{|S|} \times Entropy\ (S_v)$$

$$Gain\ (S_{sunny}, Temp) = 0.97 - \frac{2}{5} \times Entropy\ (S_{Hot}) - \frac{2}{5} \times Entropy (S_{Mild}) - \frac{1}{5} \times Entropy\ (S_{Cool})$$

$$Gain\ (S_{sunny}, Temp) = 0.97 - \frac{2}{5} \times 0 - \frac{2}{5} \times 1 - \frac{1}{5} \times 0 = 0.57$$

Let’s use Information Gain

Make a subset of dataset where outlook=sunny and repeat the process!

Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes

$$Entropy\ (S) = 0.94$$

$$Entropy\ (S_{sunny}) = -\frac{2}{5} \times log_2 \frac{2}{5} - \frac{3}{5} \times log_2 \frac{3}{5} = 0.971$$

$$Gain\ (S_{sunny}, Temp) = 0.57$$

$$Gain\ (S_{sunny}, Humidity) = 0.97$$

Attribute: Humidity
Values (Humidity) = High, Normal

$$S_{High} = [0+, 3-]$$

$$Entropy\ (S_{Hot}) = 0$$

$$S_{Normal} = [2+, 0-]$$

$$Entropy\ (S_{Mild}) = 0$$

$$Gain\ (S_{sunny}, Humidity) = Entropy\ (S_{sunny}) - \sum_{v \in \{High, Normal\}} \frac{|S_v|}{|S|} \times Entropy\ (S_v)$$

$$Gain\ (S_{sunny}, Humidity) = 0.97\ - \frac{3}{5} \times Entropy\ (S_{High}) - \frac{2}{5} \times Entropy(S_{Normal})$$

$$Gain\ (S_{sunny}, Humidity) = 0.97\ - \frac{3}{5} \times 0\ - \frac{2}{5} \times 0\ = 0.97$$

Let’s use Information Gain

Make a subset of dataset where outlook=sunny and repeat the process!

Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes

$$Entropy\ (S) = 0.94$$

$$Entropy\ (S_{sunny}) = -\frac{2}{5} \times log_2 \frac{2}{5} - \frac{3}{5} \times log_2 \frac{3}{5} = 0.971$$

$$Gain\ (S_{sunny}, Temp) = 0.57$$

$$Gain\ (S_{sunny}, Humidity) = 0.97$$

Attribute: Wind
Values (Wind) = Strong, Weak

$$S_{Strong} = [1+, 1-]$$

$$Entropy\ (S_{Strong}) = 1$$

$$S_{Weak} = [1+, 2-]$$

$$Entropy\ (S_{Weak}) = -\frac{1}{3} log\left(\frac{1}{3}\right) - \frac{2}{3} log\left(\frac{2}{3}\right) = 0.9183$$

$$Gain\ (S_{sunny}, Wind) = Entropy\ (S_{sunny}) - \sum_{v \in \{Strong, Weak\}} \frac{|S_v|}{|S|} \times Entropy\ (S_v)$$

$$Gain\ (S_{sunny}, Wind) = 0.97 - \frac{2}{5} \times Entropy\ (S_{Strong}) - \frac{3}{5} \times Entropy(S_{Weak})$$

$$Gain\ (S_{sunny}, Wind) = 0.97 - \frac{2}{5} \times 1 - \frac{3}{5} \times 0.918 = 0.0192$$

Let’s use Information Gain

$$\boldsymbol{Entropy\,(S_{sunny}) = 0.971}$$

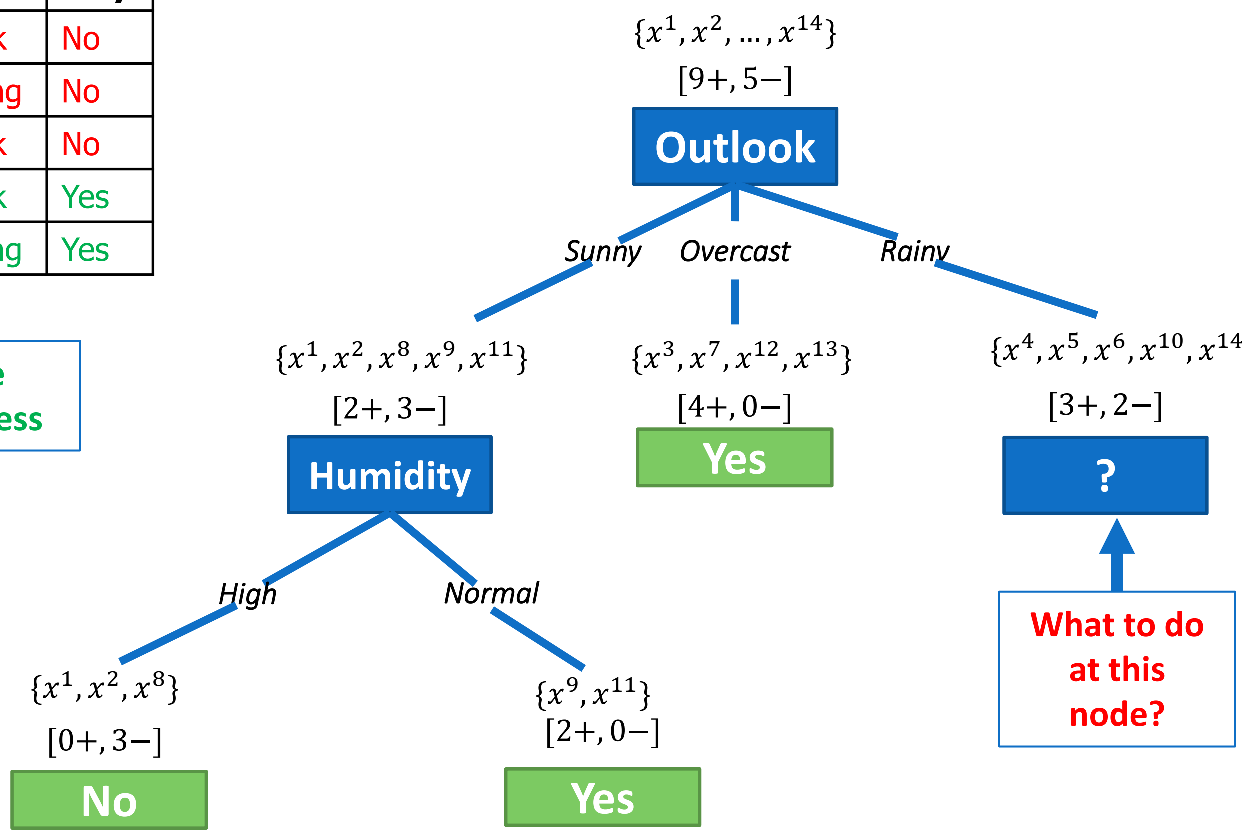
$$\boldsymbol{Gain\,(S_{sunny}, Temp) = 0.57}$$

$$\boldsymbol{Gain\,(S_{sunny}, Humidity) = 0.97}$$

$$\boldsymbol{Gain\,(S_{sunny}, Wind) = 0.192}$$

Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes

Make a subset of dataset where outlook=Rain and repeat the process



Let's use Information Gain

Make a subset of dataset where outlook=Rain and repeat the process

$$Entropy\ (S_{Rain}) = 0.97$$

$$Gain\ (S_{Rainy}, Temp) = 0.0192$$

Outlook	Temp	Humidity	Wind	Play
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Rainy	Mild	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

Attribute: Temp
Values (Temp) = Hot, Mild, Cool

$$S_{Hot} = [0+, 0-]$$

$$Entropy\ (S_{Hot}) = 0$$

$$S_{Cool} = [1+, 1-]$$

$$Entropy\ (S_{Cool}) = 1$$

$$S_{Mild} = [2+, 1-]$$

$$Entropy\ (S_{Mild}) = -\frac{2}{3} log\left(\frac{2}{3}\right) - \frac{1}{3} log\left(\frac{1}{3}\right) = 0.9183$$

$$Gain\ (S_{Rain}, Temp) = Entropy\ (S_{Rain}) - \sum_{v \in \{Hot, Mild, Cool\}} \frac{|S_v|}{|S|} \times Entropy\ (S_v)$$

$$Gain\ (S_{Rain}, Temp) = 0.97 - \frac{0}{5} \times Entropy\ (S_{Hot}) - \frac{3}{5} \times Entropy(S_{Mild}) - \frac{2}{5} \times Entropy(S_{Mild})$$

$$Gain\ (S_{sunny}, Temp) = 0.97 - \frac{0}{5} \times 0 - \frac{3}{5} \times 0.918 - \frac{2}{5} \times 1 = 0.0192$$

Let’s use Information Gain

Make a subset of dataset where outlook=Rain and repeat the process

$$Entropy\ (S_{Rain}) = 0.97$$

$$Gain\ (S_{Rainy}, Temp) = 0.0192$$

$$Gain\ (S_{Rainy}, Humidity) = 0.0192$$

Outlook	Temp	Humidity	Wind	Play
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Rainy	Mild	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

Attribute: Humidity
Values (Temp) = High, Normal

$$S_{High} = [1+, 1-]$$

$$Entropy\ (S_{High}) = 1$$

$$S_{Normal} = [2+, 1-]$$

$$Entropy\ (S_{Normal}) = -\frac{2}{3} log\left(\frac{2}{3}\right) - \frac{1}{3} log\left(\frac{1}{3}\right) = 0.9183$$

$$Gain\ (S_{Rain}, Humidity) = Entropy\ (S_{Rain}) - \sum_{v \in \{High, Normal\}} \frac{|S_v|}{|S|} \times Entropy\ (S_v)$$

$$Gain\ (S_{Rain}, Humidity) = 0.97 - \frac{2}{5} \times Entropy\ (S_{High}) - \frac{3}{5} \times Entropy(S_{Normal})$$

$$Gain\ (S_{Rain}, Humidity) = 0.97 - \frac{2}{5} \times 1 - \frac{3}{5} \times 0.918 = 0.0192$$

Let's use Information Gain

Make a subset of dataset where outlook=Rain and repeat the process

$$Entropy\ (S_{Rain}) = 0.97$$

$$Gain\ (S_{Rainy}, Temp) = 0.0192$$

$$Gain\ (S_{Rainy}, Humidity) = 0.0192$$

Outlook	Temp	Humidity	Wind	Play
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Rainy	Mild	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

Attribute: Wind
Values (Temp) = Strong, Weak

$$S_{Strong} = [0+, 2-]$$

$$Entropy\ (S_{Strong}) = 0$$

$$S_{Weak} = [3+, 0-]$$

$$Entropy\ (S_{Weak}) = 0$$

$$Gain\ (S_{Rain}, Wind) = Entropy\ (S_{Rain}) - \sum_{v \in \{Strong, Weak\}} \frac{|S_v|}{|S|} \times Entropy\ (S_v)$$

$$Gain\ (S_{Rain}, Wind) = 0.97 - \frac{2}{5} \times Entropy\ (S_{Strong}) - \frac{3}{5} \times Entropy(S_{Weak})$$

$$Gain\ (S_{Rain}, Wind) = 0.97 - \frac{2}{5} \times 0 - \frac{3}{5} \times 0.0 = 0.97$$

Let’s use Information Gain

$Entropy\ (S_{Rain}) = 0.97$

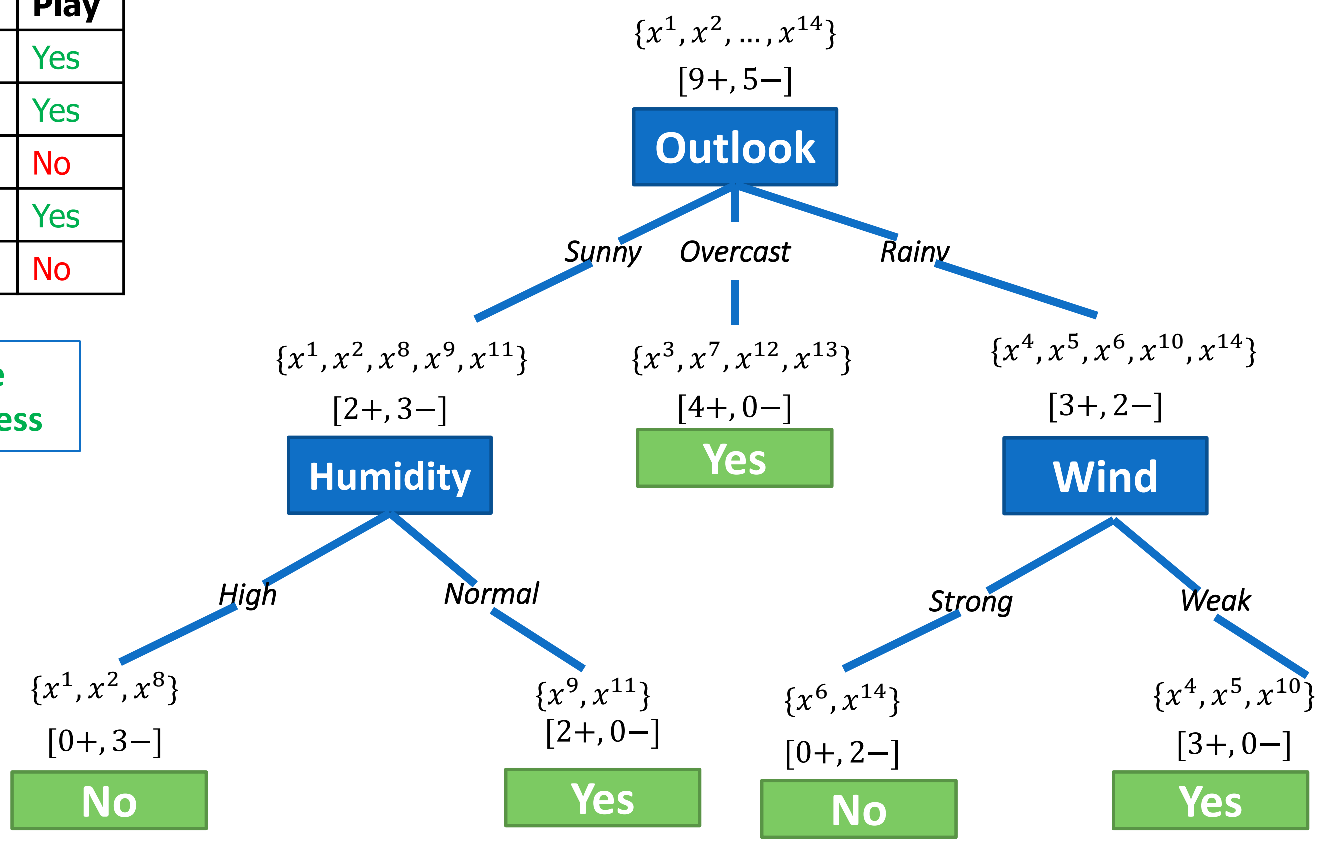
$Gain\ (S_{Rainy}, Temp) = 0.0192$

$Gain\ (S_{Rainy}, Humidity) = 0.0192$

$Gain\ (S_{Rainy}, Wind) = 0.97$

Outlook	Temp	Humidity	Wind	Play
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Rainy	Mild	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

Make a subset of dataset where outlook=Rain and repeat the process



Let’s use Information Gain

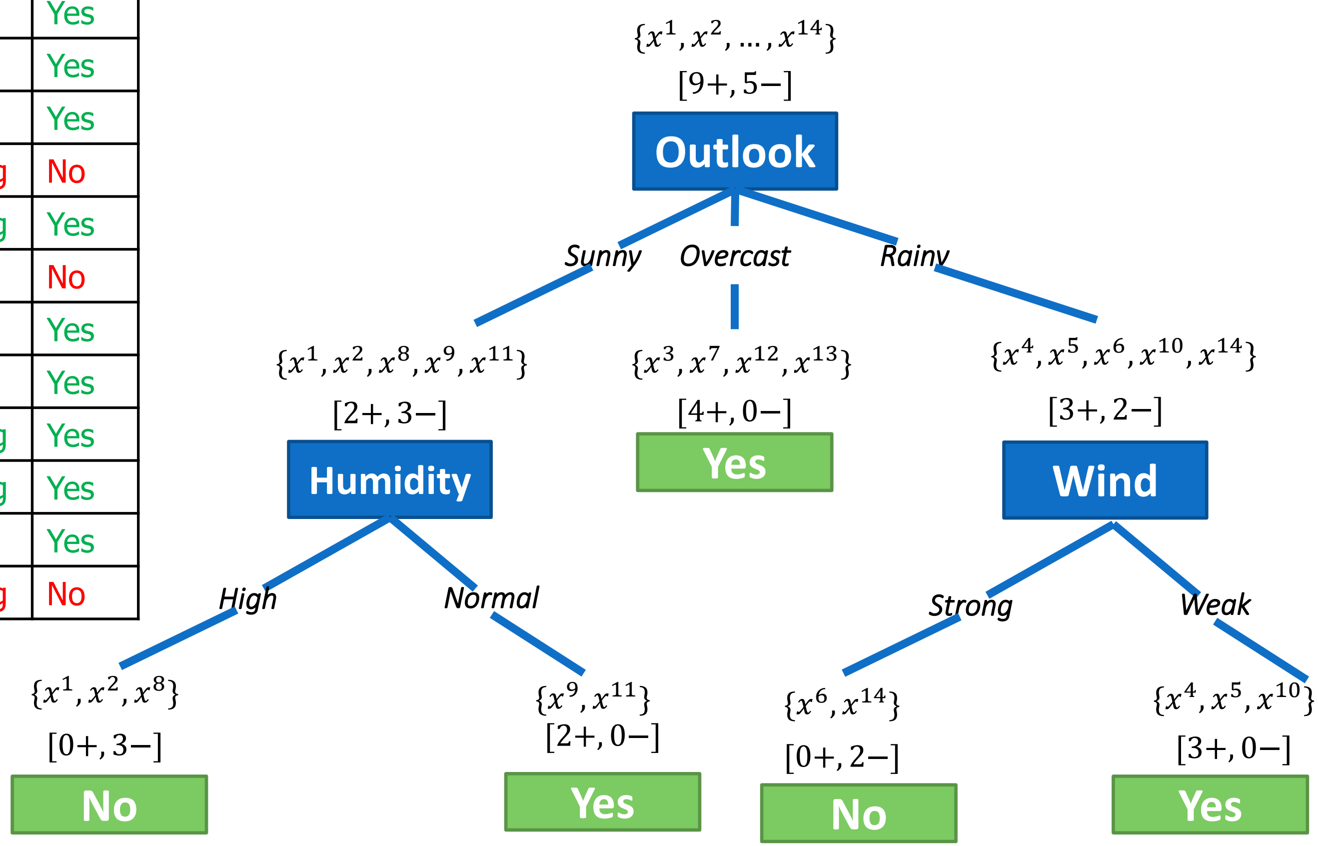
Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rainy	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

$\boldsymbol{Gain\ (S, Outlook) = 0.2464}$

$\boldsymbol{Gain\ (S, Temp) = 0.0289}$

$\boldsymbol{Gain\ (S, Wind) = 0.048}$

$\boldsymbol{Gain\ (S, Humidity) = 0.151}$



ID3 Tree Learning Algorithm

- Create a ***Root*** node for the tree
- If all ***Examples*** in training data are ***positive***, return the single-node tree ***Root***, with label $+$
- If all ***Examples*** in training data are ***negative***, return the single-node tree ***Root***, with label $-$
- If ***Attributes*** are not present, return the single-node tree ***Root***, ***with the most common value of Target Attribute in Examples***
- Otherwise Begin

ID3 Tree Learning Algorithm

- Otherwise Begin
  - $\boldsymbol{A}$ ←the attribute from ***Attributes*** that best classifies Examples
  - The decision attribute for ***Root*** $\leftarrow \boldsymbol{A}$
  - For each possible value of ***A***,
    - Add a new tree branch below Root, corresponding to the test $\boldsymbol{A = v_i}$
    - Let $\boldsymbol{Examples_{v_i}}$ be the subset of Examples that have value $\boldsymbol{v_i}$ for ***A***
    - If $\boldsymbol{Examples_{v_i}}$ is empty
      - Then below this new branch, add a leaf node with the most common value of ***Target Attribute*** in ***Examples***
    - Else
      - Below this new branch, add the subtree
      - Recursively call $\boldsymbol{ID3(Examples_{vi}, Target\ Attribute, Attributes - \{A\})}$

Inductive Bias of Decision Trees

Inductive bias

- A set of assumptions made by a learning algorithm to predict outputs of a given input that it has not seen before; that is, to generalize a finite set of observation (training data) into a general model of the domain.

Inductive Bias of Decision Trees

- Selects in favor of shorter trees over longer ones
- Selects trees that place the attributes with highest information gain closest to the root

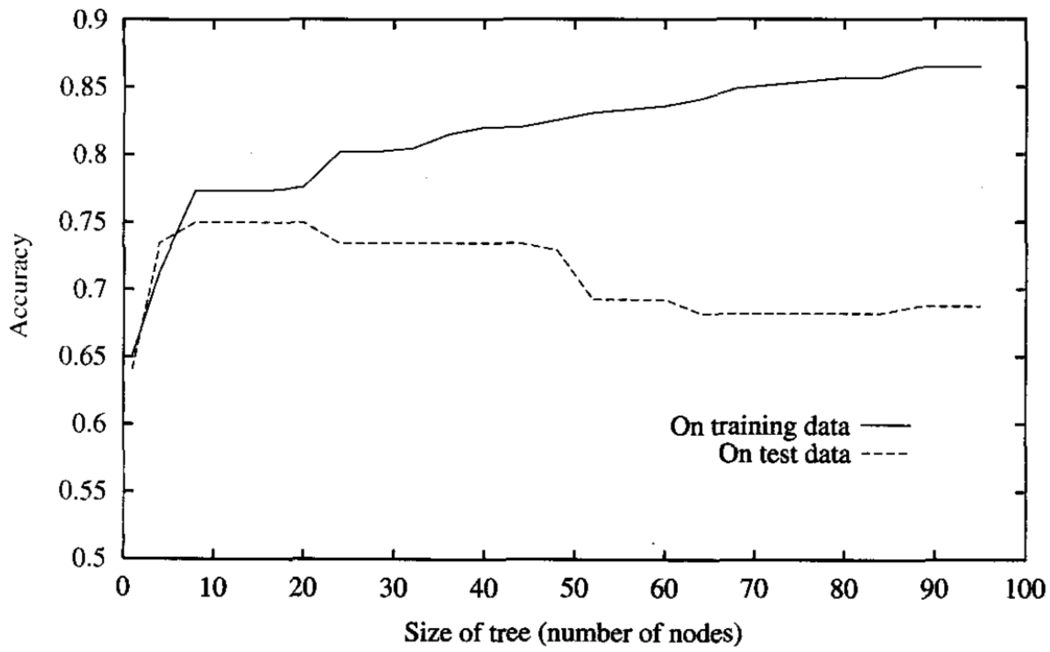
Prefer the simplest hypothesis that **fits** the data!

Problems with Decision Tree

- **Overfitting**
  - The model learns the **training data** “too well”
  - Performs poorly on testing/validation data.
- **Attributes with large number of disjoint possible values**
  - i.e., instead of $Temperature \in \{Hot, Mild, Cold\}$, $Temperature$ can take any numerical value
- **Output/Label is a continuous value**
  - How to do regression in trees?
- **Training data with missing attribute values**

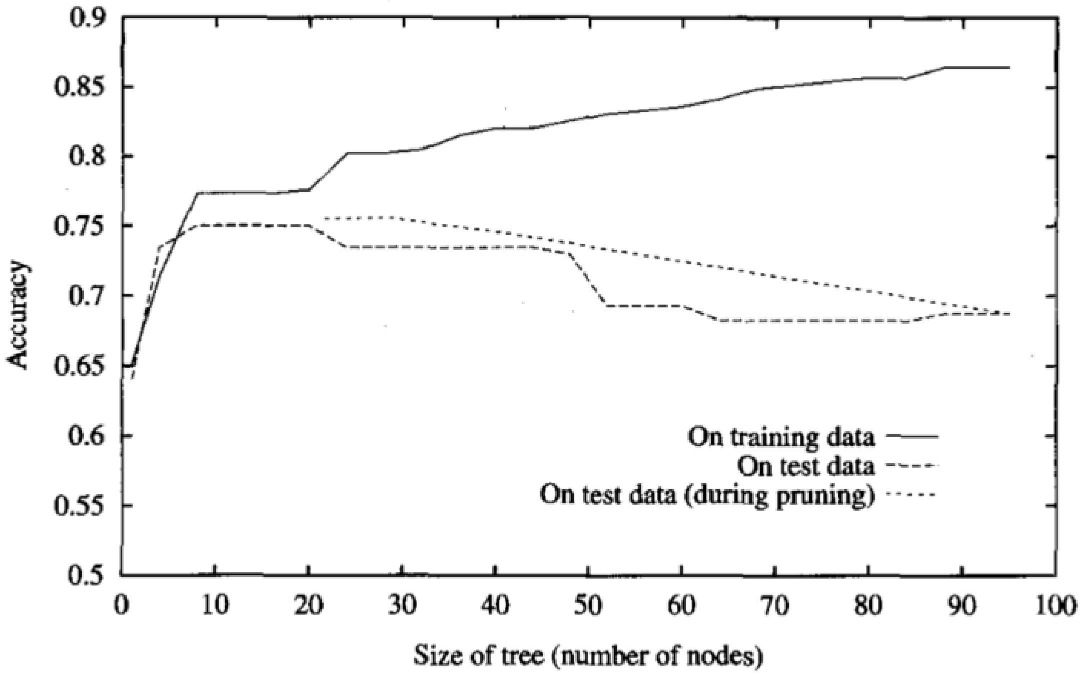
Avoiding Overfitting

- **Overfitting:** Given a hypothesis space H, a hypothesis $h \in H$ is said to **overfit** the training data, if there exists some alternative hypothesis $h' \in H$, such that h has smaller error than h' over the training examples, but h' has a smaller error than h over the entire distribution of instances (including testing/validation data).



Avoiding Overfitting

- 1- Stop growing the tree earlier, before it reaches the point where it perfectly classifies the training data
- 2- Allow the tree to overfit the data, then post-prune the tree
  - More effective than the first approach



- 3- Build Forest!

Assignment 1: Task 3

- **From Task 2**, you should have two versions of the dataset.
  - Without feature selection
  - With feature selection (Use number of features as per your liking)
- **From Task 1**, you should have your images dataset with binary labels (You vs Unknown)
- **For Task 3:**
  - Train decision tree with entropy as criterion with all three datasets (i.e., two versions of text dataset and one images dataset)
  - Output the accuracy of all three decisions trees on respective test splits.
- **Submission:**
  - Submit all three tasks together as single zip file on Moodle. All three tasks should have a separate folder.
  - Only one member of the group has to submit on Moodle.
  - Mention Roll numbers in the zip file name e.g, 23131323_2344545.zip
  - A viva will be conducted during office hours after the submission.

Book Reading

- Murphy – Chapter 1, Chapter 14
- Tom Mitchel (TM) – Chapter 3