Decision Tree

Course: Data Mining

Professor: Dr. Tahaei

Author: Mohammad Saeed Arvenaghi

Subject: ID3 Algorithm

December 2024

Question

We have a training set of 14 examples, each with four features: Outlook, Temp, Humidity, Wind, and a binary decision Tennis?. The dataset is given below:

Day	Outlook	Temp	Humidity	Wind	Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

We want to build a decision tree using the ID3 algorithm.

Solution: Applying the ID3 Algorithm

Initial Setup

We begin with a dataset of 14 examples: 9 positives (Yes) and 5 negatives (No). Label the positives as p and the negatives as n:

$$p = 9, \quad n = 5.$$

We compute the **H** (for *entropy*) of the entire dataset as follows:

$$H\left(\tfrac{p}{p+n}, \tfrac{n}{p+n}\right) = -\tfrac{9}{14}\log_2\left(\tfrac{9}{14}\right) - \tfrac{5}{14}\log_2\left(\tfrac{5}{14}\right) \approx 0.940.$$

The features we can split on are: Outlook, Temp, Humidity, Wind. We measure the information gain by splitting on each feature, and choose the feature with the largest information gain.

Feature Splits and Information Gain

Recall, the information gain is given by:

$$\text{Gain}(A) = H\left(\tfrac{p}{p+n}, \tfrac{n}{p+n}\right) - \sum_{i=1}^{k} \tfrac{p_i+n_i}{p+n} H\left(\tfrac{p_i}{p_i+n_i}, \tfrac{n_i}{p_i+n_i}\right),$$

where A is a feature that partitions our dataset into k subsets, and each subset i has p_i positive and n_i negative examples.

Split on Outlook

$$\begin{aligned}\text{Outlook} = \text{Sunny:} &\quad + : 9, 11 \quad - : 1, 2, 8 \quad (2 \text{ positives}, 3 \text{ negatives})\\ \text{Outlook} = \text{Overcast:} &\quad + : 3, 7, 12, 13 \quad - : \text{none} \quad (4 \text{ positives}, 0 \text{ negatives})\\ \text{Outlook} = \text{Rain:} &\quad + : 4, 5, 10 \quad - : 6, 14 \quad (3 \text{ positives}, 2 \text{ negatives})\end{aligned}$$

Hence:

$$\text{Gain(Outlook)} = 0.940 - \left[\frac{5}{14} H\left(\tfrac{2}{5}, \tfrac{3}{5}\right) + \frac{4}{14} H\left(\tfrac{4}{4}, \tfrac{0}{4}\right) + \frac{5}{14} H\left(\tfrac{3}{5}, \tfrac{2}{5}\right)\right] = 0.247.$$

Split on Temp

$$\begin{aligned}\text{Temp} = \text{Cool:} &\quad + : 5, 7, 9 \quad - : 6 \quad (3, 1)\\ \text{Temp} = \text{Mild:} &\quad + : 4, 10, 11, 12 \quad - : 8, 14 \quad (4, 2)\\ \text{Temp} = \text{Hot:} &\quad + : 3, 13 \quad - : 1, 2 \quad (2, 2)\end{aligned}$$

$$\text{Gain(Temp)} = 0.940 - \left[\frac{4}{14} H\left(\tfrac{3}{4}, \tfrac{1}{4}\right) + \frac{6}{14} H\left(\tfrac{4}{6}, \tfrac{2}{6}\right) + \frac{4}{14} H\left(\tfrac{2}{4}, \tfrac{2}{4}\right)\right] = 0.029.$$

Split on Wind

$$\begin{aligned}\text{Wind} = \text{Weak:} &\quad + : 3, 4, 5, 9, 10, 13 \quad - : 1, 8 \quad (6, 2)\\ \text{Wind} = \text{Strong:} &\quad + : 7, 11, 12 \quad - : 2, 6, 14 \quad (3, 3)\end{aligned}$$

$$\text{Gain(Wind)} = 0.940 - \left[\frac{8}{14} H\left(\tfrac{6}{8}, \tfrac{2}{8}\right) + \frac{6}{14} H\left(\tfrac{3}{6}, \tfrac{3}{6}\right)\right] = 0.048.$$

Split on Humidity

$$\begin{aligned}\text{Humidity} = \text{Normal:} &\quad + : 5, 7, 9, 10, 11, 13 \quad - : 6 \quad (6, 1)\\ \text{Humidity} = \text{High:} &\quad + : 3, 4, 12 \quad - : 1, 2, 8, 14 \quad (3, 4)\end{aligned}$$

$$\text{Gain(Humidity)} = 0.940 - \left[\frac{7}{14} H\left(\tfrac{6}{7}, \tfrac{1}{7}\right) + \frac{7}{14} H\left(\tfrac{3}{7}, \tfrac{4}{7}\right)\right] = 0.151.$$

Among these, Outlook yields the highest information gain (0.247), so we choose Outlook as our root feature.

Building the Subtrees

1. Subtree: Outlook = Sunny

For Outlook = Sunny, we have 2 positive and 3 negative examples.

$$H\left(\frac{2}{5}, \frac{3}{5}\right) \approx 0.971.$$

Possible splits: Temp, Humidity, Wind.

$$\text{Gain(Temp)} = 0.971 - \left[\frac{2}{5} H\left(\frac{0}{2}, \frac{2}{2}\right) + \frac{2}{5} H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{5} H\left(\frac{1}{1}, \frac{0}{1}\right)\right] = 0.571.$$

$$\text{Gain(Humidity)} = 0.971 - \left[\frac{3}{5}\,H\left(\frac{0}{3},\frac{3}{3}\right) + \frac{2}{5}\,H\left(\frac{2}{2},\frac{0}{2}\right)\right] = 0.971.$$

$$\text{Gain(Wind)} = 0.971 - \left[\frac{2}{5}\,H\left(\frac{1}{2},\frac{1}{2}\right) + \frac{3}{5}\,H\left(\frac{1}{3},\frac{2}{3}\right)\right] = 0.020.$$

Choosing the Best Feature: Humidity yields the highest information gain (0.971), so we split on Humidity.

Splitting on Humidity for Outlook = Sunny

Humidity = Normal:

$$+ : 9, 11 \quad - : \text{none} \quad \Rightarrow \text{Yes}$$

Humidity = High:

$$+ : \text{none} \quad - : 1, 2, 8 \quad \Rightarrow \text{No}$$

Thus, the subtree for Outlook = Sunny is:

$$\text{Outlook} = \text{Sunny} \quad \longrightarrow \quad \begin{cases} \text{Humidity} = \text{Normal} \quad \rightarrow \quad \text{Yes} \\ \text{Humidity} = \text{High} \quad \rightarrow \quad \text{No} \end{cases}$$

2. Subtree: Outlook = Overcast

All 4 examples are positive:

$$\text{Outlook} = \text{Overcast} \quad \rightarrow \quad \text{Yes}$$

3. Subtree: Outlook = Rain

For Outlook = Rain, we have 3 positives and 2 negatives:

$$H\left(\frac{3}{5},\frac{2}{5}\right) \approx 0.971.$$

Possible splits: Temp, Humidity, Wind.

$$\text{Gain(Temp)} = 0.971 - \left[\frac{3}{5}\,H\left(\frac{2}{3},\frac{1}{3}\right) + \frac{2}{5}\,H\left(\frac{1}{2},\frac{1}{2}\right)\right] = 0.020.$$

$$\text{Gain(Humidity)} = 0.971 - \left[\frac{3}{5}\,H\left(\frac{2}{3},\frac{1}{3}\right) + \frac{2}{5}\,H\left(\frac{1}{2},\frac{1}{2}\right)\right] = 0.020.$$

$$\text{Gain(Wind)} = 0.971 - \left[\frac{3}{5}\,H\left(\frac{3}{3},\frac{0}{3}\right) + \frac{2}{5}\,H\left(\frac{0}{2},\frac{2}{2}\right)\right] = 0.971.$$

Choosing the Best Feature: Wind yields the highest information gain (0.971), so we split on Wind.

Splitting on Wind for Outlook = Rain

Wind = Weak:

$$+ : 4, 5, 10 \quad - : \text{none} \quad \Rightarrow \text{Yes}$$

Wind = Strong:

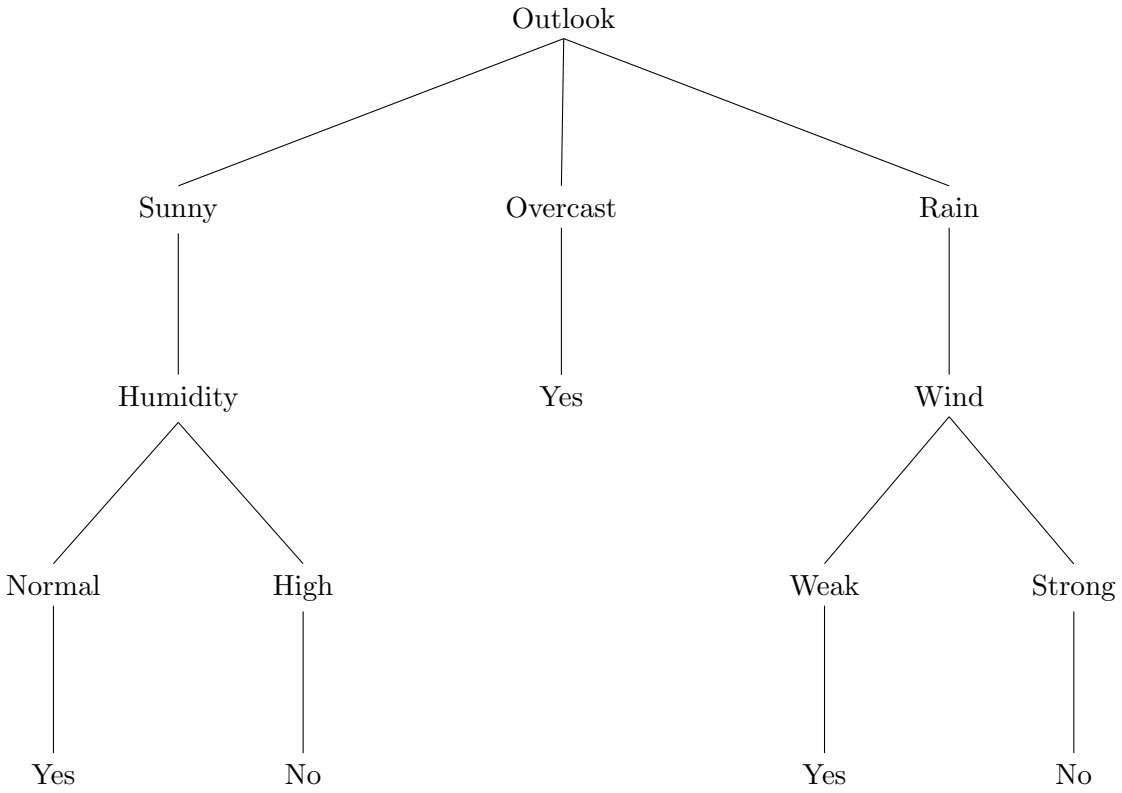
$$+ : \text{none} \quad - : 6, 14 \quad \Rightarrow \text{No}$$

Thus, the subtree for Outlook = Rain is:

$$\text{Outlook} = \text{Rain} \longrightarrow \begin{cases} \text{Wind} = \text{Weak} & \rightarrow \quad \text{Yes} \\ \text{Wind} = \text{Strong} & \rightarrow \quad \text{No} \end{cases}$$

Final Decision Tree

Summarizing all the subtrees, the final decision tree is depicted below:



Explanation:

- **Root Node:** Outlook is chosen as the root node since it has the highest information gain (0.247).
- **Sunny Branch:**
  - Humidity is chosen as the next split with an information gain of 0.971.
  - Humidity = Normal leads to all positives (Yes).
  - Humidity = High leads to all negatives (No).
- **Overcast Branch:**

  - All examples are positive (Yes), so it's a leaf node.

- **Rain Branch:**
  - Wind is chosen as the next split with an information gain of 0.971.
  - Wind = Weak leads to all positives (Yes).
  - Wind = Strong leads to all negatives (No).

This simple decision tree perfectly classifies the training data and is expected to perform well on the test set, thus serving as a good hypothesis for predicting whether Bertie will play tennis based on the given weather attributes.