ECE/CS/ME 539 – Fall 2024 — Activity 10

1.

Using the table provided, answer the following questions to construct the decision tree.

Outlook	**Temperature**	**Humidity**	**Wind**	**Played football(yes/no)**
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Table 1: Weather conditions and football play decisions

(a) Compute the entropy of the entire dataset for the target variable "Played football (yes/no)". Show your work.

The dataset has 14 instances. The target variable "Played football (yes/no)" has 9 "Yes" and 5 "No".

Entropy (S) for the entire dataset is given by:

$$\text{Entropy}(S) = -\sum_{i=1}^{c} p_i \log_2(p_i)$$

where p_i is the proportion of instances in class i.

$$\text{Entropy}(S) = -(p_{\text{Yes}} \log_2(p_{\text{Yes}}) + p_{\text{No}} \log_2(p_{\text{No}}))$$

$$p_{\text{Yes}} = \frac{9}{14}, \quad p_{\text{No}} = \frac{5}{14}$$

$$\begin{aligned}\text{Entropy}(S) &= -\left(\frac{9}{14}\log_2\frac{9}{14} + \frac{5}{14}\log_2\frac{5}{14}\right)\\ &= -\left(\frac{9}{14}\cdot(-0.530) + \frac{5}{14}\cdot(-0.737)\right)\\ &= 0.940\end{aligned}$$

(b) Calculate the entropy for the subsets of the "Outlook" attribute. Use these subsets to determine the information gain for each possible split: "Sunny", "Overcast", and "Rain". Which attribute value of "Outlook" gives the maximum information gain?

To determine the best split based on the "Outlook" attribute, we first calculate the entropy for each subset and then compute the information gain.

Entropy for each subset of "Outlook":

- Sunny: 5 instances (2 Yes, 3 No)

$$H(\text{Sunny}) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) \approx 0.971$$

- Overcast: 4 instances (4 Yes, 0 No)

$$H(\text{Overcast}) = -\left(\frac{4}{4}\log_2\frac{4}{4} + \frac{0}{4}\log_2\frac{0}{4}\right) = 0$$

- Rain: 5 instances (3 Yes, 2 No)

$$H(\text{Rain}) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) \approx 0.971$$

Information Gain for each split:

$$IG(\text{Outlook}) = H(S) - \sum_i \frac{|S_i|}{|S|}H(S_i)$$

- Sunny:

$$IG(\text{Sunny}) = 0.94 - \left(\frac{5}{14}\times 0.971\right) \approx 0.246$$

- Overcast:

$$IG(\text{Overcast}) = 0.94 - \left(\frac{4}{14}\times 0\right) = 0.94$$

- Rain:

$$IG(\text{Rain}) = 0.94 - \left(\frac{5}{14}\times 0.971\right) \approx 0.246$$

The "Overcast" value of the "Outlook" attribute provides the maximum information gain of **0.94**.

(c) After selecting the "Outlook" attribute and performing the first split, repeat the calculation of entropy, information gain, and resulting splits for the next attribute, focusing on the "Sunny" subset. Should the next split be on "Humidity" or "Wind"? Provide the calculations and rationale behind your choice.

After selecting "Outlook" as the first split, we focus on the "Sunny" subset to determine the next best attribute to split on. The "Sunny" subset contains 5 instances:

- 2 instances where "Played football" is "Yes"
- 3 instances where "Played football" is "No"

The entropy for the "Sunny" subset is:

$$H(\text{Sunny}) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) \approx 0.971$$

We now calculate the entropy and information gain for the possible splits on "Humidity" and "Wind".

Split on "Humidity":

- High: 3 instances (0 Yes, 3 No)

$$H(\text{High}) = -\left(\frac{0}{3}\log_2\frac{0}{3} + \frac{3}{3}\log_2\frac{3}{3}\right) = 0$$

- Normal: 2 instances (2 Yes, 0 No)

$$H(\text{Normal}) = -\left(\frac{2}{2}\log_2\frac{2}{2} + \frac{0}{2}\log_2\frac{0}{2}\right) = 0$$

The information gain for splitting on "Humidity" is:

$$IG(\text{Humidity}) = H(\text{Sunny}) - \left(\frac{3}{5}\times 0 + \frac{2}{5}\times 0\right) = 0.971$$

Split on "Wind":

- Weak: 3 instances (2 Yes, 1 No)

$$H(\text{Weak}) = -\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) \approx 0.918$$

- Strong: 2 instances (0 Yes, 2 No)

$$H(\text{Strong}) = -\left(\frac{0}{2}\log_2\frac{0}{2} + \frac{2}{2}\log_2\frac{2}{2}\right) = 0$$

The information gain for splitting on "Wind" is:

$$IG(\text{Wind}) = H(\text{Sunny}) - \left(\frac{3}{5}\times 0.918 + \frac{2}{5}\times 0\right) \approx 0.550$$

The next split should be on the "Humidity" attribute, as it provides a higher information gain (**0.971**) compared to "Wind" (**0.550**).

(d) Perform the necessary calculations to determine the splits for the remaining subsets of "Outlook" (i.e., "Overcast" and "Rain"), considering "Wind" and "Humidity" as potential attributes for splitting. Include all your entropy and information gain calculations.

After splitting on "Outlook", we now focus on the remaining subsets: "Overcast" and "Rain". We will calculate the entropy and information gain for potential splits on "Wind" and "Humidity" for each subset.

1. Overcast Subset:

The "Overcast" subset contains 4 instances, all of which are "Yes" for "Played football". Therefore, the entropy for this subset is:

$$H(\text{Overcast}) = -\left(\frac{4}{4}\log_2\frac{4}{4} + \frac{0}{4}\log_2\frac{0}{4}\right) = 0$$

Since the entropy is 0, no further splitting is necessary for the "Overcast" subset.

2. Rain Subset:

The "Rain" subset contains 5 instances:

- 3 instances where "Played football" is "Yes"
- 2 instances where "Played football" is "No"

The entropy for the "Rain" subset is:

$$H(\text{Rain}) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) \approx 0.971$$

We now calculate the entropy and information gain for potential splits on "Humidity" and "Wind".

Split on "Humidity":

- High: 2 instances (0 Yes, 2 No)

$$H(\text{High}) = -\left(\frac{0}{2}\log_2\frac{0}{2} + \frac{2}{2}\log_2\frac{2}{2}\right) = 0$$

- Normal: 3 instances (3 Yes, 0 No)

$$H(\text{Normal}) = -\left(\frac{3}{3}\log_2\frac{3}{3} + \frac{0}{3}\log_2\frac{0}{3}\right) = 0$$

The information gain for splitting on "Humidity" is:

$$IG(\text{Humidity}) = H(\text{Rain}) - \left(\frac{2}{5}\times 0 + \frac{3}{5}\times 0\right) = 0.971$$

Split on "Wind":

- Weak: 3 instances (3 Yes, 0 No)

$$H(\text{Weak}) = -\left(\frac{3}{3}\log_2\frac{3}{3} + \frac{0}{3}\log_2\frac{0}{3}\right) = 0$$

- Strong: 2 instances (0 Yes, 2 No)

$$H(\text{Strong}) = -\left(\frac{0}{2}\log_2\frac{0}{2} + \frac{2}{2}\log_2\frac{2}{2}\right) = 0$$

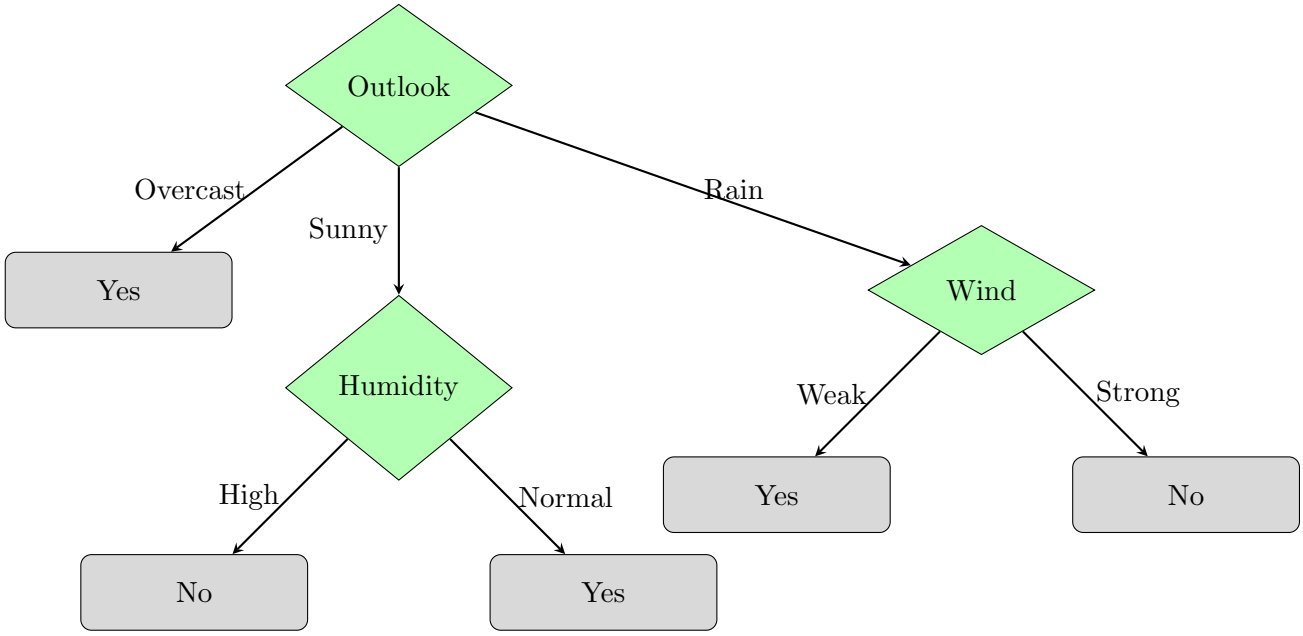
The information gain for splitting on "Wind" is:

$$IG(\text{Wind}) = H(\text{Rain}) - \left(\frac{3}{5}\times 0 + \frac{2}{5}\times 0\right) = 0.971$$

For the "Rain" subset, both "Humidity" and "Wind" provide the same information gain of **0.971**. Therefore, either attribute can be chosen for the split, as they both perfectly separate the data.

(e) Based on your calculations, construct the decision tree. Verify if your final decision tree matches the tree shown in the figure provided in the problem statement. If there are any discrepancies, explain them.

Based on the calculations from the previous parts, the decision tree is constructed as follows:



The decision tree is constructed as follows:

- The root node is the attribute **Outlook**.
- If **Outlook** is **Overcast**, the decision is **Yes**.
- If **Outlook** is **Sunny**, the next split is on the attribute **Humidity**:
  - If **Humidity** is **High**, the decision is **No**.
  - If **Humidity** is **Normal**, the decision is **Yes**.
- If **Outlook** is **Rain**, the next split is on the attribute **Wind**:
  - If **Wind** is **Weak**, the decision is **Yes**.
  - If **Wind** is **Strong**, the decision is **No**.