Week 11-2: Decision Trees

Ekarat Rattagan

April 4, 2025

Decision Tree

- Dataset

X_1	X_2	Y
M	Y	1
F	Y	1
M	N	0

- Rule Representation:
  - If $X_1 = F$ then $Y = 1$
  - If $X_1 = M$ and $X_2 = N$ then $Y = 0$

Example of data set

day	outlook	temp	humidity	wind	play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Information Entropy

- A measure of impurity in a set of examples
- How much variance the data has?
- **Binary class:**

$$H(S) = -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$

- **Multi-class:**

$$H(S) = -\sum_{i=1}^{k} p_i \log_2 p_i \quad \text{where } k = \#\text{classes}$$

- **Example:**

$$H(S) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} \approx 0.940$$

- Ref: `https://victorzhou.com/blog/information-gain/`

Information Gain

- How can we quantify the quality of a split?
- Higher Gain $\Rightarrow$ better split.

Definition:

$$\begin{aligned} \text{Gain}(S, A) &= H(S) - H(S|A) \\ &= H(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} H(S_v) \end{aligned}$$

- S: set of examples
- A: attribute
- Values(A): possible values of A
- S_v: subset of S where $A = v$
- Ref: `https://victorzhou.com/blog/information-gain/`

Example: Gain(S, Humidity)

$$
\begin{aligned}
\text{Gain}(S, \text{Humidity}) &= H(S) - \sum_{v \in \{\text{high,normal}\}} \frac{|S_v|}{|S|} H(S_v) \\
&= 0.94 - (\frac{7}{14}(-\frac{3}{7} log_2 \frac{3}{7} - \frac{4}{7} log_2 \frac{4}{7}) + \\
&\qquad \frac{7}{14}(-\frac{3}{7} log_2 \frac{3}{7} - \frac{4}{7} log_2 \frac{4}{7}) \\
&= 0.151
\end{aligned}
$$

ID3 algorithm (part 1)

```
def id3(examples, classification_attribute, attributes):
    create a root node for the tree
    if all examples are positive/yes:
        return root node with positive/yes label
    else if all examples are negative/no:
        return root node with negative/no label
    else if there are no attributes left:
        return root node with most popular
classification_attribute label
    else:
```

ID3 algorithm (part 2)

```
best_attribute = attribute from attributes that best
                 classifies examples
assign best_attribute to root node
for each value in best_attribute:
    add branch below root node for the value
    branch_examples = [examples that have that value
                       for best_attribute]
    if branch_examples is empty:
        add leaf node with most popular
            classification_attribute label
    else:

        add subtree id3(branch_examples,
                        classification_attribute,
                        attributes - best_attribute)
```

Information Gain for each attribute

IG(S, outlook) = 0.248

IG(S, humidity) = 0.151

IG(S, wind) = 0.048

IG(S, temp) = 0.029

Choose attribute with **highest IG** for root.

Subset Example for Outlook (Sunny) and humidity

ID3([D1, D2, D8, D9, D11],
[N, N, N, Y, Y],
[Humidity, Wind, temp])

Compute:

$$IG(S_{\text{sunny}}, \text{Humidity}) = H(S_{\text{sunny}}) - \sum_{v \in \{\text{high,normal}\}} \frac{|S_v|}{|S|} H(S_v)$$

$$= -\frac{2}{5} log_2(\frac{2}{5}) - \frac{3}{5} log_2(\frac{3}{5})$$

$$- \left[\frac{3}{5}(0) + \frac{2}{5}(0)\right]$$

$$= 0.97$$

Subset Example for Outlook (Sunny) and temp

$$IG(S_{\text{sunny}}, \text{temp}) = H(S_{\text{sunny}}) - \sum_{v \in \{\text{hot,mild,cool}\}} \frac{|S_v|}{|S|} H(S_v)$$

$$= -\frac{2}{5} log_2(\frac{2}{5}) - \frac{3}{5} log_2(\frac{3}{5})$$

$$- \left[\frac{2}{5}(0) + \frac{2}{5}(1) + \frac{1}{5}(0)\right]$$

$$= 0.57$$

Subset Example for Outlook (Sunny) and wind

$$IG(S_{\text{sunny}}, \text{temp}) = H(S_{\text{sunny}}) - \sum_{v \in \{\text{weak,strong}\}} \frac{|S_v|}{|S|} H(S_v)$$

$$= -\frac{2}{5} log_2(\frac{2}{5}) - \frac{3}{5} log_2(\frac{3}{5})$$

$$- \left[\frac{3}{5}(0.92) + \frac{2}{5}(1)\right]$$

$$= 0.018$$

Decision Tree Overview

- ID3 $\rightarrow$ Not available in scikit-learn
- C4.5 $\rightarrow$ Handles continuous data
- CART (Classification and Regression Tree)
  - Binary tree
  - Uses Information Gain (IG), Gini Index (Impurity)
  - Used as splitting criterion

Splitting Criteria

1. **Information Gain** (based on Entropy)
2. **Gini Index** (Impurity)

Gini Index formula:

$$GI = 1 - \sum_{i=1}^{K} p_i^2 \quad \text{where } p_i = \frac{\text{value of each class}}{\text{Total}}$$

Gini Index: Outlook Example

$$\mathrm{GI}(\text{outlook} = \text{sunny}) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0.48$$

$$\mathrm{GI}(\text{outlook} = \text{overcast}) = 1 - \left(\frac{4}{4}\right)^2 = 0$$

$$\mathrm{GI}(\text{outlook} = \text{rain}) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$$

$$GI(\text{outlook}) = \frac{5}{14}(0.48) + \frac{4}{14}(0) + \frac{5}{14}(0.48) = 0.342$$

Gini Index for Other Attributes

$$GI(\text{temperature}) = \frac{4}{14}(0.5) + \frac{6}{14}(0.44) + \frac{4}{14}(0.375) = 0.439$$

$$GI(\text{humidity}) = \frac{7}{14}(0.48) + \frac{7}{14}(0.24) = 0.367$$

$$GI(\text{wind}) = \frac{8}{14}(0.375) + \frac{6}{14}(0.5) = 0.428$$

$\therefore$ Pick **outlook** as the root node.

Overfitting in Decision Trees

- High depth $\Rightarrow$ Too many nodes
- **Solution to overfit:**
  1. Limit the number of depths or iterations (e.g., ID3).
  2. Pruning (Post-pruning): adjust after training.
  3. Bagging VS Boosting.
  4. Ref1: `https://www.kaggle.com/code/prashant111/bagging-vs-boosting`
  5. Ref2: `https://www.kaggle.com/code/satishgunjal/ensemble-learning-bagging-boosting-stacking`

Pruning: Bottom-Up Walk

- Case 1: Replace internal node with a leaf.
- Case 2: Replace internal node with a subtree.