Decision Tree Homework

Decision Tree Analysis

Given Data:

Instance	Meat	Crust	Veg	Quality
1	Y	Thin	N	Great
2	N	Deep	N	Bad
3	N	Stuffed	Y	Good
4	Y	Stuffed	Y	Great
5	Y	Deep	N	Good
6	Y	Deep	Y	Great
7	N	Thin	Y	Good
8	Y	Deep	N	Good
9	N	Thin	N	Bad

Calculating Entropy at the Root Node:

Step 1: Compute $Info(S)$

We have three classes: Bad (B), Good (G), Great (Gr).

Count of each class in the dataset:

- Bad: 2
- Good: 4
- Great: 3
- Total instances: 9

Compute probabilities:

- $p_{\text{Bad}} = \frac{2}{9}$
- $p_{\text{Good}} = \frac{4}{9}$

- $p_{\text{Great}} = \frac{3}{9}$

Compute entropy:

$$\begin{aligned}
Info(S) &= -\left(p_{\text{Bad}} \log_2 p_{\text{Bad}} + p_{\text{Good}} \log_2 p_{\text{Good}} + p_{\text{Great}} \log_2 p_{\text{Great}}\right) \\
&= -\left(\frac{2}{9} \log_2 \frac{2}{9} + \frac{4}{9} \log_2 \frac{4}{9} + \frac{3}{9} \log_2 \frac{3}{9}\right) \\
&\approx -\left(0.2222 \times (-2.1699) + 0.4444 \times (-1.1699) + 0.3333 \times (-1.5849)\right) \\
&\approx 1.529
\end{aligned}$$

Calculating Information Gain for Each Attribute:

Attribute: Meat

Possible values: Y, N

Splitting Data by Meat:

- **Meat = Y** (Instances: 1, 4, 5, 6, 8)
  - Class counts: Good = 2, Great = 3
- **Meat = N** (Instances: 2, 3, 7, 9)
  - Class counts: Bad = 2, Good = 2

Entropy for Each Subset:

- $Info(S_{\text{Meat=Y}})$

$$\begin{aligned}
Info(S_{\text{Meat=Y}}) &= -\left(0 \times \log_2 0 + \frac{2}{5} \log_2 \frac{2}{5} + \frac{3}{5} \log_2 \frac{3}{5}\right) \\
&= -\left(0 + 0.4 \times (-1.3219) + 0.6 \times (-0.7369)\right) \\
&\approx 0.971
\end{aligned}$$

- $Info(S_{\text{Meat=N}})$

$$\begin{aligned}
Info(S_{\text{Meat=N}}) &= -\left(\frac{2}{4} \log_2 \frac{2}{4} + \frac{2}{4} \log_2 \frac{2}{4} + 0 \times \log_2 0\right) \\
&= -\left(0.5 \times (-1) + 0.5 \times (-1) + 0\right) \\
&= 1.0
\end{aligned}$$

Weighted Entropy After Split:

$$Info_{\text{Meat}}(S) = \frac{5}{9} \times 0.971 + \frac{4}{9} \times 1.0 \approx 0.984$$

Information Gain:

$$Gain(S, \text{Meat}) = Info(S) - Info_{\text{Meat}}(S) \approx 1.529 - 0.984 = 0.545$$

Attribute: Crust

Possible values: Thin, Deep, Stuffed

[Detailed calculations omitted for brevity]

Information Gain:

$$Gain(S, \text{Crust}) \approx 0.112$$

Attribute: Veg

Possible values: Y, N

[Detailed calculations omitted for brevity]

Information Gain:

$$Gain(S, \text{Veg}) \approx 0.239$$

Best Attribute at Root Node:

- **Meat** has the highest information gain.

Splitting on Meat:

- **Root Node**: Split on Meat (Y/N)

Second Level - Leftmost Node (Meat = Y):

Remaining Attributes: Crust, Veg

Entropy at Node $S_{\text{Meat=Y}}$:

$$Info(S_{\text{Meat=Y}}) \approx 0.971$$

Calculating Information Gain for Remaining Attributes:

Attribute: Crust

Possible values: Thin, Deep, Stuffed

Splitting Data by Crust:

- **Crust = Thin** (Instance: 1)
  - Class: Great
- **Crust = Deep** (Instances: 5, 6, 8)
  - Class counts: Good = 2, Great = 1
- **Crust = Stuffed** (Instance: 4)
  - Class: Great

Entropy for Each Subset:

- $Info(S_{\text{Crust=Thin}}) = 0$ (Pure node)
- $Info(S_{\text{Crust=Deep}})$

$$Info(S_{\text{Crust=Deep}}) = -\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) \\ \approx 0.918$$

- $Info(S_{\text{Crust=Stuffed}}) = 0$ (Pure node)

Weighted Entropy After Split:

$$Info_{\text{Crust}}(S_{\text{Meat=Y}}) = \frac{1}{5} \times 0 + \frac{3}{5} \times 0.918 + \frac{1}{5} \times 0 \approx 0.551$$

Information Gain:

$$Gain(S_{\text{Meat=Y}}, \text{Crust}) = Info(S_{\text{Meat=Y}}) - Info_{\text{Crust}}(S_{\text{Meat=Y}}) \approx 0.971 - 0.551 = 0.420$$

Attribute: Veg

[Detailed calculations omitted for brevity]

Information Gain:

$$Gain(S_{\text{Meat=Y}}, \text{Veg}) \approx 0.420$$

Best Attribute at This Node:

- **Crust** (Alphabetically first among attributes with equal gain)

Splitting on Crust at Meat = Y Node:

- **Crust = Thin**: Leaf node labeled **Great**
- **Crust = Stuffed**: Leaf node labeled **Great**
- **Crust = Deep**: Leaf node labeled **Good** (Majority class)

Final Decision Tree (Up to Second Level):

1. **Root Node**: Meat
    - **Meat = Y**:
        - Split on **Crust**
            - **Crust = Deep**: **Good**
            - **Crust = Stuffed**: **Great**
            - **Crust = Thin**: **Great**
    - **Meat = N**:
        - [Further splitting can be done for practice]

Leaf Node Labels:

- Nodes are labeled with their majority class if not pure.