# ISOM3360 Assignment: Decision Tree

Zhang Yichen

March 24, 2020

## 1 Answer

The feature ***Snow*** should be the first split condition for root node as it will creat a highest information gain after split.

## 2 Procedures

According to ID3 algorithm, the feature to split the root should be able to split it into subsets for which the information gain is maximum. Hence, I will compare the information gain by using different feature as the root node splitting condition. I denote the root node set as $\boldsymbol{S}$ and the two subsets as $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$.

### Select *Snow* as the root node

$$\begin{aligned} IG(feature = \textbf{\textit{Snow}}, \boldsymbol{S}) &= H(\boldsymbol{S}) - p(\boldsymbol{T}_1)H(\boldsymbol{T}_1) - p(\boldsymbol{T}_2)H(\boldsymbol{T}_2) \\ &= -(\frac{6}{10}\log_2(\frac{6}{10}) + \frac{4}{10}\log_2(\frac{4}{10})) + \frac{6}{10} * (\frac{1}{6}\log_2(\frac{1}{6}) \\ &\quad + \frac{5}{6}\log_2(\frac{5}{6})) + \frac{4}{10} * (\frac{1}{4}\log_2(\frac{1}{4}) + \frac{3}{4}\log_2(\frac{3}{4})) \\ &= 0.971 - 0.390 - 0.325 \\ &= \textbf{0.256} \end{aligned}$$

### Select *Season* as the root node

$$\begin{aligned} IG(feature = \textbf{\textit{Season}}, \boldsymbol{S}) &= H(\boldsymbol{S}) - p(\boldsymbol{T}_1)H(\boldsymbol{T}_1) - p(\boldsymbol{T}_2)H(\boldsymbol{T}_2) \\ &= -(\frac{6}{10}\log_2(\frac{6}{10}) + \frac{4}{10}\log_2(\frac{4}{10})) + \frac{5}{10} * (\frac{1}{5}\log_2(\frac{1}{5}) \\ &\quad + \frac{4}{5}\log_2(\frac{4}{5})) + \frac{5}{10} * (\frac{2}{5}\log_2(\frac{2}{5}) + \frac{3}{5}\log_2(\frac{3}{5})) \\ &= 0.971 - 0.361 - 0.485 \\ &= \textbf{0.125} \end{aligned}$$

### Select *Weather* as the root node

$$
\begin{aligned}
IG(feature = \boldsymbol{Weather}, \boldsymbol{S}) &= H(\boldsymbol{S}) - p(\boldsymbol{T}_1)H(\boldsymbol{T}_1) - p(\boldsymbol{T}_2)H(\boldsymbol{T}_2) \\
&= -(\frac{6}{10}\log_2(\frac{6}{10}) + \frac{4}{10}\log_2(\frac{4}{10})) + \frac{4}{10} * (\frac{2}{4}\log_2(\frac{2}{4}) \\
&\quad + \frac{2}{4}\log_2(\frac{2}{4})) + \frac{6}{10} * (\frac{2}{6}\log_2(\frac{2}{6}) + \frac{4}{6}\log_2(\frac{4}{6})) \\
&= 0.971 - 0.4 - 0.551 \\
&= \mathbf{0.020}
\end{aligned}
$$

Therefore,

$$IG(feature = \boldsymbol{Snow}, \boldsymbol{S}) > IG(feature = \boldsymbol{Season}, \boldsymbol{S}) > IG(feature = \boldsymbol{Weather}, \boldsymbol{S})$$

The feature ***Snow*** should be the first split condition for root node as it will creat a highest information gain after split.