CIS419 Machine Learning

Assignment I

Part I

James Wang
Sept. 23, 2015

PennKey: jamwang
PennID: 46576241

1. Decision Tree Learning

(a) At the root node for the decision tree in this domain, the information gains for the attributes Outlook and Humidity are calculated by subtracting the weighted average entropy of the children splits to the parent node.

$Parent\ Entropy: A = Outlook$

$$-\left(\frac{5}{14} \times log_2\ \frac{5}{14}\right) - \left(\frac{9}{14} \times log_2\ \frac{9}{14}\right) \approx 0.940286$$

$Child\ Entropy_{Overcast}$:

$$-\left(\frac{5}{14} \times log_2\ \frac{5}{14}\right) - \left(\frac{9}{14} \times log_2\ \frac{9}{14}\right) = 0$$

$Child\ Entropy_{Rain}$:

$$-\left(\frac{5}{14} \times log_2\ \frac{5}{14}\right) - \left(\frac{9}{14} \times log_2\ \frac{9}{14}\right) \approx 0.970951$$

$Child\ Entropy_{Sunny}$:

$$-\left(\frac{5}{14} \times log_2\ \frac{5}{14}\right) - \left(\frac{9}{14} \times log_2\ \frac{9}{14}\right) \approx 0.970951$$

$Weighted\ Average\ Entropy\ of\ Children_{Outlook}$:

$$\frac{4}{14}(0) + \frac{5}{14}(0.970951) + \frac{5}{14}(0.970951) \approx 0.693536$$

$\boldsymbol{Information\ Gain_{Outlook}}$:

$$Gain(S, A) = I_s(A, Y) = H_s(Y) - H_s(Y|A) = 0.940286 - 0.693536 = \mathbf{0.24675}$$

$Parent\ Entropy: A = Humidity$

$$-\left(\frac{5}{14} \times log_2\ \frac{5}{14}\right) - \left(\frac{9}{14} \times log_2\ \frac{9}{14}\right) \approx 0.940286$$

$Child\ Entropy_{>75\%}$:

$$-\left(\frac{4}{9} \times log_2\ \frac{4}{9}\right) - \left(\frac{5}{9} \times log_2\ \frac{5}{9}\right) = 0.991076$$

$Child\ Entropy_{\leq 75\%}$:

$$-\left(\frac{1}{5} \times log_2\ \frac{1}{5}\right) - \left(\frac{4}{5} \times log_2\ \frac{4}{5}\right) \approx 0.721928$$

$Weighted\ Average\ Entropy\ of\ Children_{Humidity}$:

$$\frac{9}{14}(0.991076)\ + \frac{5}{14}(0.721928) \approx 0.894952$$

$\boldsymbol{Information\ Gain_{Humidty}}$:

$$Gain(S, A) = I_s(A, Y) = H_s(Y) - H_s(Y|A) = 0.940286 - 0.894952 = \mathbf{0.045334}$$

(b) Calculating gain ratios gives us another method for judging the quality of a split. Gain ratios are used to compensate for information gain's affinity for attributes that have a large number of distinct values. Gain ratio is calculated by dividing the information gain (*Gain*) by the information due to the split (*SplitInfo)*.

$SplitInfo\ (\boldsymbol{Outlook}, X)$:

$$-\left(\frac{4}{14} \times log_2\ \frac{4}{14}\right) - \left(\frac{5}{14} \times log_2\ \frac{5}{14}\right) - \left(\frac{5}{14}\ log_2 \frac{5}{14}\right) \approx 1.577406$$

$SplitInfo\ (\boldsymbol{Humidity}, X)$:

$$-\left(\frac{9}{14} \times log_2\ \frac{9}{14}\right) - \left(\frac{5}{14} \times log_2\ \frac{5}{14}\right) \approx 0.940286$$

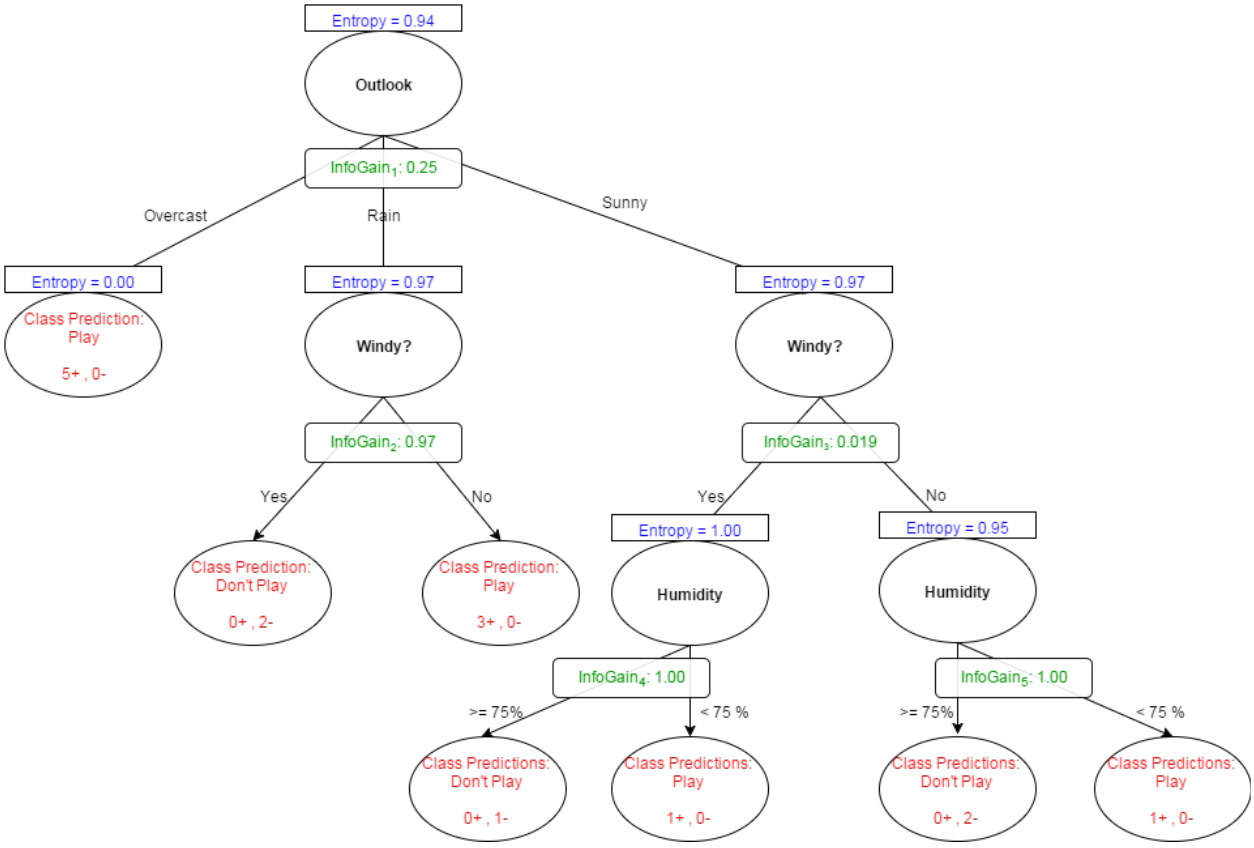
$\boldsymbol{Gain\ Ratio}\ (\boldsymbol{Outlook}, \boldsymbol{X})$:

$$\frac{0.24675}{1.577406} \approx \mathbf{0.156428}$$

$\boldsymbol{Gain\ Ratio}\ (\boldsymbol{Humidity}, \boldsymbol{X})$:

$$\frac{0.045334}{0.940286} \approx \mathbf{0.048213}$$

(c) The complete, unpruned decision tree. Class predictions at the leaves. The methodology to choosing which attributes to split on was simple: choose the attribute which would result in the largest zero-entropy (homogenous) leaf-node. Information Gain was used to calculate the splits and decide on attributes because most splits were binary, and thus, the true advantage of Gain Ratio (less biased towards attributes with distinct values than Information Gain) would not come into play in this situation as much.



InfoGain₁ and *InfoGain₂ calculated previously*.

InfoGain₃: $0.97 - 0.950978 = 0.019022$

$$Child\ Entropy_{Windy-Yes}: -\left(\frac{2}{3}log_2\frac{2}{3}\right) - \left(\frac{1}{3}log_2\frac{1}{3}\right) \approx 0.918296$$

$$Child\ Entropy_{Windy-No}: \frac{2}{5}\ (1) + \frac{3}{5}\ (0.918296) \approx 0.950978$$

InfoGain₄ and InfoGain₅ calculated intuitively.