CIS 419/519: Homework 1

{Yupeng Li}

Although the solutions are entirely my own, I consulted with the following people and sources while working on this homework: {http://www.cs.utep.edu/vladik/cs5315.13/cs5315_13kader.pdf}

1 Decision Tree Learning

a. Show your work:

$$InfoGain(PainLocation) = Info(X) - Info(X|PainLocation)$$

$$Info(X) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.9402$$

$$InfoGain(PainLocation) = Info(X) - \frac{5}{14}(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}) - \frac{5}{14}(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}) - 0 = 0.2467$$

$$InfoGain(Temperature) = Info(X) - Info(X|Temperature)$$

$$= Info(X) - \frac{4}{14}(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}) - \frac{10}{14}(-\frac{7}{10}\log_2\frac{7}{10} - \frac{3}{10}\log_2\frac{3}{10}) = 0.02499$$

b. Show your work:

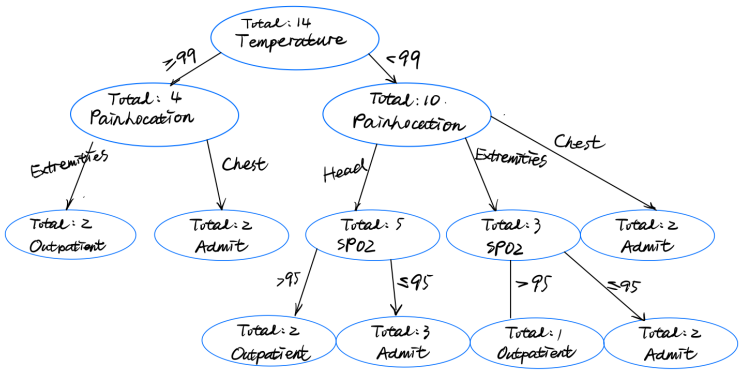
$$SplitInformation(PainLocation) = -\frac{5}{14}\log_2\frac{5}{14} - \frac{5}{14}\log_2\frac{5}{14} - \frac{4}{14}\log_2\frac{4}{14} = 1.5774$$

$$GainRatio(PainLocation) = \frac{InfoGain(PainLocation)}{SplitInformation(PainLocation)} = \frac{0.2467}{1.5774} = 0.156397$$

$$SplitInformation(Temperature) = -\frac{4}{14}\log_2\frac{4}{14} - \frac{10}{14}\log_2\frac{10}{14} = 0.86312$$

$$GainRatio(Temperature) = \frac{InfoGain(Temperature)}{SplitInformation(Temperature)} = \frac{0.02499}{0.86312} = 0.028953$$

c.



d. No because finding the global optimal decision tree is NP hard which has been proved in May, 1976. The advantage of ID3 is that it uses greedy approach which turns out to be really fast. ID3 will not produce the optimal decision tree, however, it gives good enough approximation. For more that 50% of the datasets, ID3 can give the optimal solution.

2 Decision Trees & Linear Discriminants [CIS 519 ONLY]

A decision tree can include oblique splits by...

3 Programming Exercises

Features: What features did you choose and how did you preprocess them?

Parameters: What parameters did you use to train your best decision tree
Performance Table:

Feature Set	Accuracy	Conf. Interval [519 ONLY]
DT 1	a	b
DT 2	a	b
DT 3	a	b

Conclusion: What can you conclude from your experience?