CIS 519: Introduction to Machine Learning
Assignment 1

Archith Shivanagere Muralinath(PennID: 82629708, PennKey: archith)

Problem 1a:

Information gain associated with *Outlook* attribute at the root node:

Parent Entropy $= -\left(\frac{9}{14}.log_2\frac{9}{14}\right) - \left(\frac{5}{14}.log_2\frac{5}{14}\right)$

Parent Entropy $= 0.94029$

Entropy of child *Sunny* $= -\left(\frac{2}{5}.log_2\frac{2}{5}\right) - \left(\frac{3}{5}.log_2\frac{3}{5}\right)$

Entropy of child *Sunny* $= 0.97$

Entropy of child *Overcast* $= 0$ (because it is homogeneous)

Entropy of child *Rain* $= -\left(\frac{2}{5}.log_2\frac{2}{5}\right) - \left(\frac{3}{5}.log_2\frac{3}{5}\right)$

Entropy of child *Rain* $= 0.97$

Average entropy of children $= \left(\frac{5}{14} \times 0.97\right) + 0 + \left(\frac{5}{14} \times 0.97\right) = 0.6928$

Information Gain $= 0.94029 - 0.6928$
Information Gain $=$ **0.2472**

Information gain associated with *Humidity* attribute at the root node:

Parent Entropy $= -\left(\frac{9}{14}.log_2\frac{9}{14}\right) - \left(\frac{5}{14}.log_2\frac{5}{14}\right)$

Parent Entropy $= 0.94029$

Entropy of child *with humidity less than or equal to 75* $= -\left(\frac{1}{5}.log_2\frac{1}{5}\right) - \left(\frac{4}{5}.log_2\frac{4}{5}\right)$

Entropy of child *with humidity less than or equal to 75* $= 0.7219$

Entropy of child *with humidity greater than 75* $= -\left(\frac{4}{9}.log_2\frac{4}{9}\right) - \left(\frac{5}{9}.log_2\frac{5}{9}\right)$

Entropy of child *with humidity greater than 75* $= 0.9910$

Average entropy of children $= \left(\frac{5}{14} \times 0.7219\right) + \left(\frac{9}{14} \times 0.9910\right) = 0.8948$

Information Gain $= 0.94029 - 0.8948$
Information Gain $=$ **0.0455**

Problem 1b:

Gain Ratio associated with *Outlook* attribute at the root node:

We know from above calculation for *Outlook* that information gain $= 0.247$

Splitinfo $= -\left(\frac{5}{14}.log_2\frac{5}{14}\right) - \left(\frac{4}{14}.log_2\frac{4}{14}\right) - \left(\frac{5}{14}.log_2\frac{5}{14}\right)$

Splitinfo $= 1.577$

Gain Ratio $= \frac{0.247}{1.577} =$ **0.1569**

Gain Ratio associated with *Humidity* attribute at the root node:

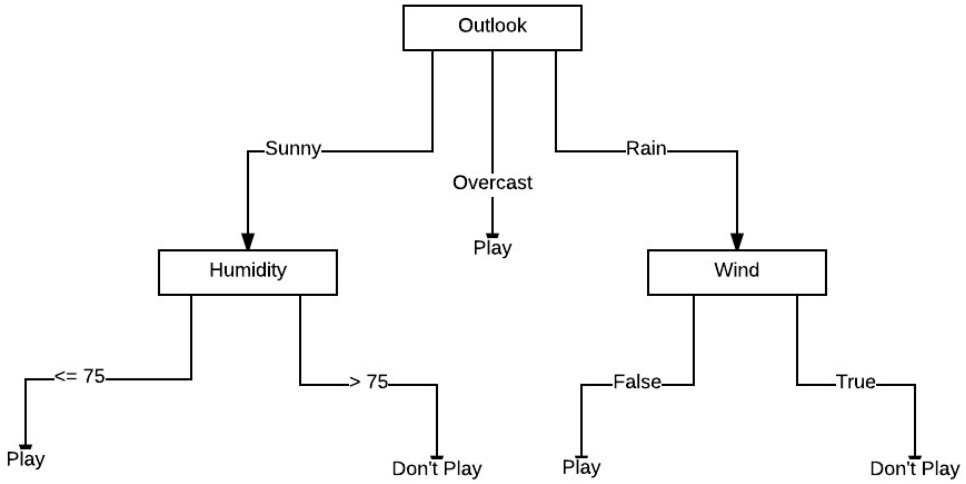
We know from above calculation for *Humidity* that information gain $= 0.0455$

Splitinfo $= -\left(\frac{5}{14}.log_2\frac{5}{14}\right) - \left(\frac{9}{14}.log_2\frac{9}{14}\right)$

Splitinfo $= 0.94029$

Gain Ratio $= \frac{0.0455}{0.94029} =$ **0.0484**

Problem 1c:


Outlook
Sunny
Overcast
Rain
Play
Humidity
Wind
<= 75
> 75
False
True
Play
Don't Play
Play
Don't Play

1.3 Comparing Decision Trees

Decision Tree Accuracy = 0.739175925926 (0.0839000478025)
Decision Stump Accuracy = 0.793064814815 (0.0780263328936)
3-level Decision Tree = 0.759819444444 (0.0814928389338)

This is because, as the depth increases, the decision tree will try to overfit on the training data.

Problem 2:

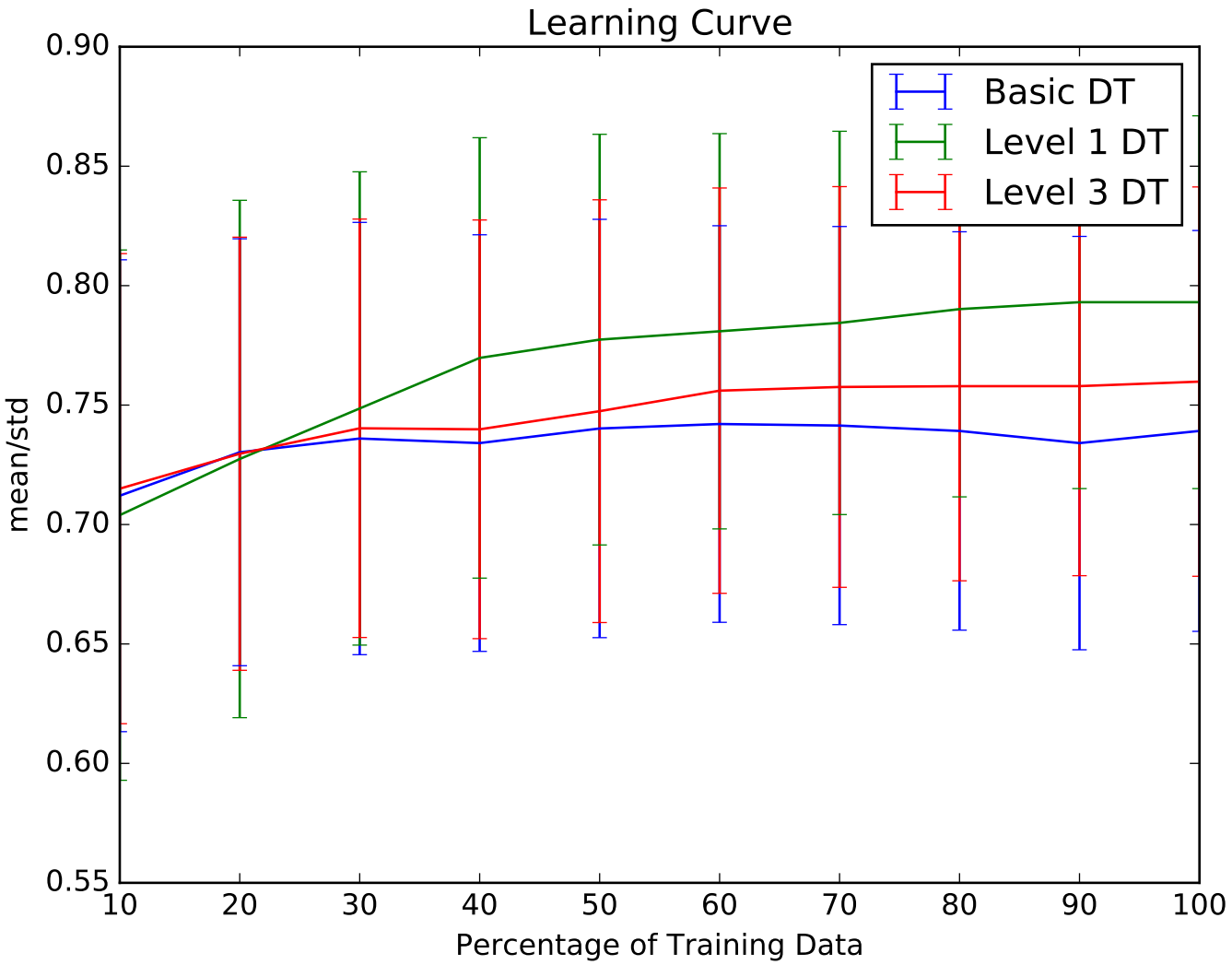
By making use of perceptrons concepts or using N-dimensional hyperplanes methods, we can modify a classic decision tree algorithm (ID3/C4.5) to obtain oblique splits (i.e, splits that are not parallel to an axis). Instead of splitting the data using just information gain values, we can adapt these techniques to split the data at intermediate levels.

Consider for instance, we need to predict if a person is healthy or malnutrition. We can split the data based on attributes like age, weight, height. The algorithm can help us define a better split of data according to some ratios (height/weight or age:weight), for a person, if these ratios are above or below some previously defined values. Traversing through a decision tree, we can go ahead and learn hyperplanes at all levels.

Part 2:

1.4 Generating a Learning Curve:

Please find below the learning curve generated by running the program.


Learning Curve
Basic DT
Level 1 DT
Level 3 DT
mean/std
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
10
20
30
40
50
60
70
80
90
100
Percentage of Training Data