Assignment 3

Q.1] Read pruning, entropy and decision gain in decision tree. Compute entropy manually from slide, put your steps and result in a word file.

Definitions –

1. Pruning – Pruning can be defined as the technique in modern world computing that reduces the size of the decision trees by withdrawing those sections of the tree that gives less force or ability to classify the instances.

2. Entropy – Entropy can be defined as the sum of the probability of each label times the log of probability of that same label.

Calculation of Entropy –

Consider a task of learning to classify the first names of male and female groups. The given data is as follows: -

Name Gender

Rohan M

Rosalie F

Brian M

Shruti F

Given the above data we want to predict the gender of the “Pratik”

The first step is to decide what features of the data are relevant in predicting the gender of the name. Some of the features include: first/last letter of the name, length of the name, number of vowels, name ending in vowel or not. Considering all these features, the data can be evaluated as –

Name Ending vowel Number of vowels Length of name Gender

Rohan 0 2 5 M

Rosalie 1 4 7 F

Brian 0 2 5 M

Shruti 1 2 6 F

Now building a decision tree for predicting the gender –

One of the example of decision tree can be –

Length < 7

Number of vowels < 3 : Male

Number of vowels >=3

Ending vowel = 0 : Male

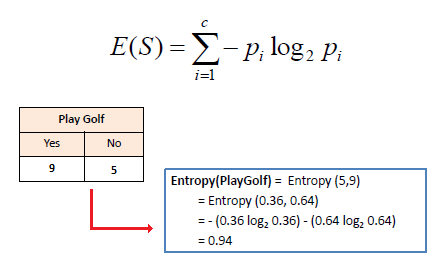
Ending vowel = 1: Female

Thus if we run the name “Pratik” down this tree, first it tests is the length < 7, and in the case of “Pratik” the length is 6 so it is less than 7, it goes down the branch of the decision tree. Then it evaluates number of vowels less than 3 or not. It is less than 3, thus the gender of the person named “Pratik” evaluates to be a Male.

Thus if we want to find the entropy we will calculate as shown below –

The formula of Entropy is: -

Entropy =



Where pi is the proportion of S belonging to class i

Imagine the following scenario where the process of constructing a tree leads to this situation :

Ending vowels

[9 Male, 5 Female]

=1 =0

Which corresponds to [3 Male, 4 Females] where ending vowel = 1 and [6 Males, 1 Female] where ending vowel = 0

Now p(Male) = 9/14 and p(Female) = 5/14

Entropy before the decision = - (5/14)\*log2(5/14) – (9/14)\*log2(9/14) = 0.9403

Entropy where ending vowel = 1 🡺 -(3/7)\*log2(3/7) – (4/7)\*log2(4/7) = 0.9852

Entropy where ending vowel = 0 🡺 -(6/7)\*log2(6/7) – (1/7)\*log2(1/7) = 0.5917

Combining the entropies we get

Entropy = 7/14\*0.9852 + 7/14\*0.5917 = 0.7885

Information gain = Entropy before the decision – Combined Entropy = 0.1518