DATA ANALYTICS PROJECT REPORT

Group-19

- A. Sarat Chandra -S20180010011
- M. Venkateswarlu - S20180010107
- P. Ragan Murali - S20180010130
- T. Vishruth -S20180010175
- V. Maurya Babu -S20180010194

Topic-8: Decision Tree Based Classifier Model using ID3 Algorithm

Goals:

To Build a Classifier model based on ID3 Algorithm and divide the Dataset randomly in 2:1 ratio using Bootstrap sampling method and K Fold Cross Validation methods to train the model accordingly for predicting test samples.

Decision Tree ID3:

Decision Tree is a Data Structure which helps us to classify computational data by dividing it into nodes and terminals based

on the factors like Entropy, Information gain etc. ID3 is an algorithm for Decision Tree implementation which uses Entropy and Information Gain.

I. Entropy

The entropy of a set of m distinct values is the minimum number of yes/no questions needed to determine unknown values from these m possibilities.

Measure of randomness or Degree of uncertainty.

If pi denotes the frequencies of occurrences of m distinct objects, then

the entropy E is

$$E = \sum_{i=1}^{m} p_i \log(1/p_i) \text{ and } \sum_{i=1}^{m} p_i = 1$$

II. Information Gain

ID3 algorithm defines a measurement of a splitting called Information Gain to determine the goodness of a split.

- The attribute with the largest value of information gain is chosen as the splitting attribute and
- It partitions into a number of smaller training sets based on the distinct values of attributes under split.

Information Gain, α(A,D) of the training set D splitting on the attribute A is given by

α(A,D) = E(D) - E

Information Gain

$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N} I(D_{left}) - \frac{N_{right}}{N} I(D_{right})$$

f: feature split on
D_p: dataset of the parent node
D_{left}: dataset of the left child node
D_{right}: dataset of the right child node
I: impurity criterion (Gini Index or Entropy)
N: total number of samples
N_{left}: number of samples at left child node
N_{right}: number of samples at right child node

Information gain = entropy (parent) – [weightes average] ★ entropy (children)

Bootstrap Sampling:

Bootstrap Sampling is a method which mainly involves picking up sample data repeatedly from the training data by replacing it from a data set to estimate parameters.

K-Fold Cross Validation:

K-Fold Cross Validation is a technique in which the dataset is divided into k number of groups which are divided into training and validation datasets. Where for one iteration we consider k-1 number of groups as training set and the remaining as validation set and this process is repeated for the further iterations by

changing the groups. And then we find the mean of all the scores that represents our model accuracy.

Best value of k:

The k value will be argmin {e1,e2,.....,em}, where e1, e2,... are mean of error rates on validation for each iteration respectively.

$$k \text{ value} = \underset{j}{\text{argmin}} \{e_1, e_2, \ldots, e_j, \ldots, e_m\}$$

Model:

1. Dataset Details:

   There are a total of 649 data samples in the dataset, which consists of 33 attributes. Using the first 32 attributes we try to predict the final grade('G3') output.

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3	4	1	1	3	4	0	11	11
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1	1	3	2	9	11	11
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3	2	2	3	3	6	12	13	12
3	GP	F	15	U	GT3	T	4	2	health	services	...	3	2	2	1	1	5	0	14	14	14
4	GP	F	16	U	GT3	T	3	3	other	other	...	4	3	2	1	2	5	0	11	13	13
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
644	MS	F	19	R	GT3	T	2	3	services	other	...	5	4	2	1	2	5	4	10	11	10
645	MS	F	18	U	LE3	T	3	1	teacher	services	...	4	3	4	1	1	1	4	15	15	16
646	MS	F	18	U	GT3	T	1	1	other	other	...	1	1	1	1	1	5	6	11	12	9
647	MS	M	17	U	LE3	T	3	1	services	services	...	2	4	5	3	4	2	6	10	10	10
648	MS	M	18	R	LE3	T	3	2	services	other	...	4	4	1	3	4	5	4	10	11	11

649 rows × 33 columns

2. Implementation:

- We use get_bootstrap_samples() function to divide our dataset randomly into 2:1 ratio by using bootstrap sampling method
- Then by _get_entropy() we will find the Entropy for overall dataset.
- To identify the root node we find the entropy for each attribute by using _get_entropy(), and then we find the information gain for each attribute by _get_information_gain() method
- We choose the maximum information gain attribute as a root node.
- Then we repeat the above 3 steps for finding the child nodes for the corresponding parent node which in whole are the part of _id3_recv() method.
- The edge determines the value of an attribute.

Accuracy:

The Accuracy of our model is defined by the total number of correctly predicted samples divided by the total number of samples.

F1-Score:

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

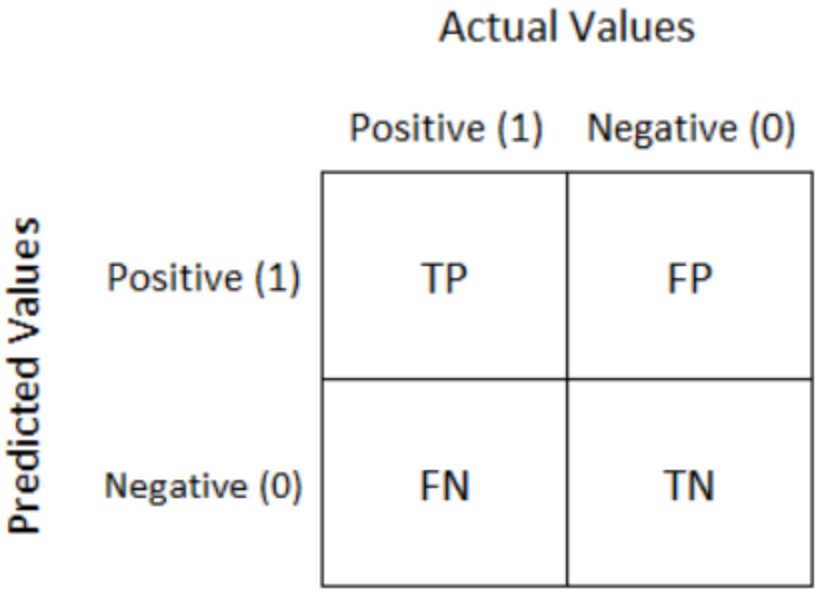
Precision:

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall:

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Confusion Matrix:



Results:

By K-Fold Cross Validation

```
Mean Accuracy on overall Validation sets: 65.0 %
Mean Accuracy on test set: 57.0 %
```

By Bootstrap Sampling Method

```
Confusion Matrix :  [[0, 1, 2, 0, 3], [0, 1, 61, 0, 62], [0, 5, 121, 0, 126], [0, 1, 24, 0, 25], [0, 8, 208, 0, 0]]
Accuracy% :  56.481481481481474
Precision% :  1.1111111111111112
Recall% :  16.666666666666664
F Measure :  2.0833333333333335
```