SMAI (CSE 471)
Spring-2019
Assignment-1

Name: Karnati Venkata Kartheek

Roll Number 2018801010

PART1

Algorithm implementation

In the categorical data each attribute have some limited set of values. From the given data the tree is constructed such that at each node a decision is made (Like if a categorical attribute is equal to a particular value? or if a numerical attribute is less than particular value). By taking a decision samples are divided into two parts. If decision attribute is categorical attribute then one part is with categorical attribute having particular value and other with categorical attribute not equal to that that particular value. If decision attribute is numerical attribute then one part is with that numerical attribute having values less than decision attribute value and other part is with values having more than that decision attribute value.

Here impurity measuring functions such as Gini Index, Misclassification rate and entropy, gives us which decision is best so that mixing of positive and negative labels is minimized. If for a particular decision the weighted average of impurity of two parts of data is less than those of all others(decisions) then that particular decision is the best decision.

Here In the code a recursive function takes in all the samples of data and calls the function that calculates the Best Attribute at that node which returns the Best Attribute Along with its value. On obtaining the Best Attribute a node is created to represent that attribute and function is called recursively to create two child nodes for this node, giving input the two groups of samples that are formed at this node. Recursive function is terminated when all the samples have labels positive or all have labels negative or less than a minimum number of samples in each nodes or impurity of current samples becomes less than a particular value.

Given data is split into two groups. Training data and validation data. A random number generator is used to assign random numbers between 0 and 1 to each row of given data. A row a selected for training data if its less than 0.8. As random number generator is mostly uniform distribution, the proportion of labels is preserved in both training and validation data.

Results on validation data for part 1:

Using Entropy as impurity criteria

_____

Precision is 0.32556270096463025

Recall is 0.7758620689655172

F1_Score is 8.720987654320988

Accuracy is 0.5749221876389506

_____

Here there is noise in data, many observations are rows with all same attributes but different labels because data considered here only has categorical dimensions and numerical dimensions are omitted. Hence the accuracy is low.

PART2

Algorithm implementation

Code is mostly similar to categorical case except here numerical part is involved.

Results for part2 On Validation Data:

Using Entropy as impurity measure

_____

Precision is 0.9959016393442623

Recall is 0.9310344827586207

Accuracy is 0.9831036016007114

F1_Score is 4.168724279835391

_____

Here the tree depth is maintained at level 8 as it is observed that validation error increases after than value.

PART3

Results for part2 On Validation Data:

Using Gini Coefficient as impurity measure

Precsion is 0.9859437751004017

Recall is 0.9406130268199234

Accuracy is 0.9831036016007114

Using Entropy as impurity measure

Precision is 0.9959016393442623

Recall is 0.9310344827586207

Accuracy is 0.9831036016007114

Using Misclassification Rate as impurity measure

Precsion is 0.9643044619422572

Recall is 0.8532280538783094

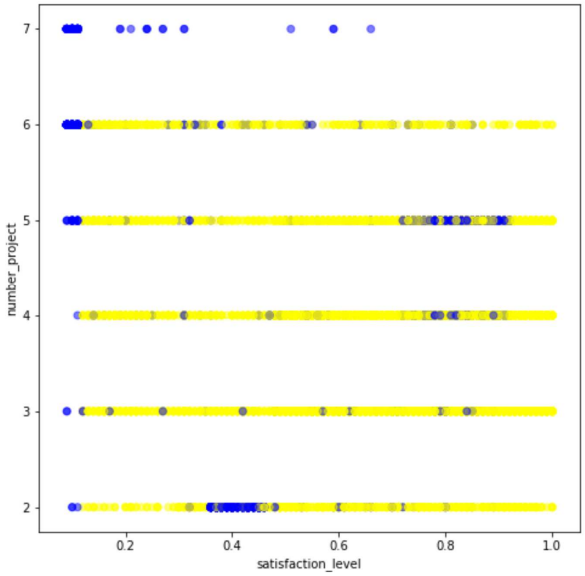
Accuracy is 0.957281121370564

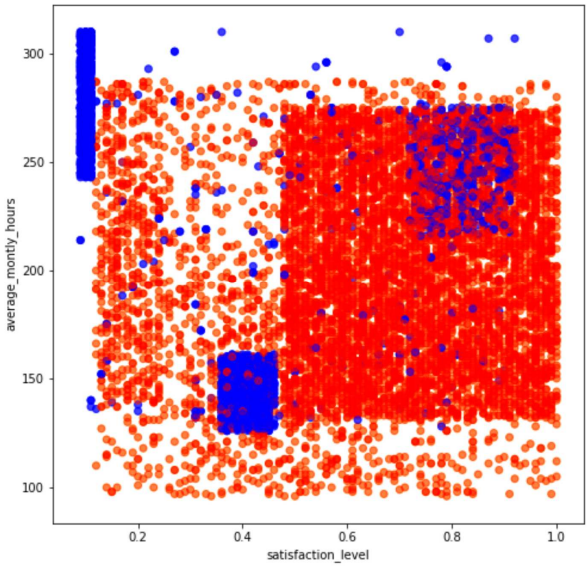
Here Entropy is Giving Highest Precision, followed by Gini Coefficient and MisClassification Rate

For Accuracy Entropy is as good as Gini Coefficient followed by MisClassification Rate

While Both Gini Coefficient and Entropy giving high recall value MisClassification Rate has less recall Value

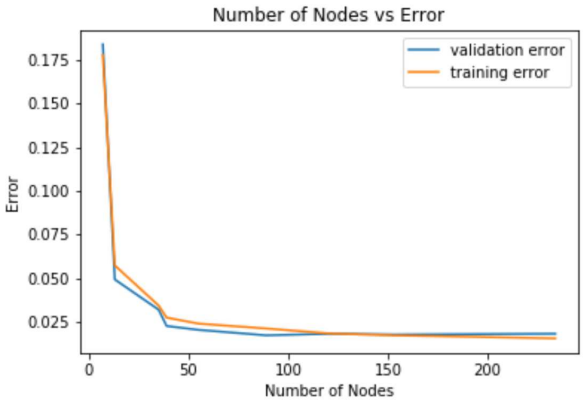
PART4


number_project
satisfaction_level


average_montly_hours
satisfaction_level

PART5


Depth of Decison Tree vs Error
validation error
training error
Error
Depth


Number of Nodes vs Error
validation error
training error
Error
Number of Nodes

PART6

1)Those test samples where the attributes values are missing can be ignored or removed if test samples are huge and its fine even if some samples are not predicted

2)Missing Values can be averaged

3)If we are at a node and the decision attribute is missing then go along both the children nodes using the remaining attributes, when two leaf nodes are reached check which leaf node has the minimum impurtiy and go along that path.

4)If multiple attributes are missing, then as above go through both nodes at missing attributes and in all possible paths find the path where we end up at a leaf node having minimum impurity or minimum mixing take that path.