Applied Machine Learning

HW 1 - Q3

Irfan Ali - CS16BTECH11019

Problem Statement : Create a Decision Tree Model to predict the quality of wine based on 11 different attributes

Language Used : Python

Solution :
DecisionTree.py **:** This program implements the basic decision tree without any optimizations like pruning. This program divides the dataset completely. It uses mean to get the splitting point of the attribute . This tree overfits the training data because we did not specify any threshold for entropy

Accuracy achieved : *83.65 %*

DecisionTree-Pruned.py **:** This program is an optimized version of the above program. To prevent overfitting of data, we give a threshold to entropy. If entropy is over the threshold, only then we split the dataset further else we just classify based on the maximum output in the dataset. The best threshold for which maximum accuracy was achieved after cross-validation was selected to be the threshold. This program uses mean to get the splitting point of an attribute.

Threshold = 0.32
Accuracy achieved : *84.28 %*

***DecisionTree-Pruned-2.py* :** This program is almost similar to the above pruned one. Except that the function used for getting the splitting point is different. To get the splitting point of the attribute, the dataset is sorted based on the attribute value. Then for each pair of consecutive values of the attribute, mean is taken and gain is calculated. Then the best splitting point is chosen as the one which gives the highest gain. This did not lead to great improvement in accuracy but it did reduce the size of the decision tree from ~800 to ~350,

Threshold = 0.32
Accuracy achieved : *84.26 %*

***DecisionTree-GiniIndex.py* :** This program uses Gini Index instead of entropy as a measure to get the splitting point of an attribute. There was no improvement in accuracy if only gini index was used. But when pruned with gini index there was an improvement.

Threshold = 0.11
Accuracy achieved : *84.20 %*

Overall, the best performance was when pruning was used with entropy. This was because decision trees overfit the data if we let them go till the end. So we put a threshold to entropy which reduces the size of the tree and also overfitting.