Assignment 1: Classification using Modified Decision Trees

DMG (CSE-506)
Group 25
Bhavya Narang, Yash Aggarwal, Robin Garg, Bhagesh Gaur, Darsh Parikh

Objective

The objective of this assignment is to perform classification on datasets using decision trees but with modifications. Instead of using the traditional methods, we use logistic regression as a function to split the nodes. These trees have been shown to produce good results. The advantage of using logistic regression is that explicit class probability estimates are produced rather than just a classification.

Question 1

For single attribute, given a simple decision tree classifier, we have modified the split part. Instead of simply finding the information gain for every feature we iterate over all features and apply logistic regression at each decision tree node. We split the training dataset again into two smaller parts, first of which is used to fit the decision tree classifier and the second part is used to predict on the classifier. We choose the feature for split which gives the highest accuracy and then calculate the threshold for each split.

Question 1

For multiple attributes, we have chosen a pair of attributes and then fitted using similar process as in case of single attributes. Then we find the threshold using information gain again and use this to split the nodes. As these contains two nested for loops the time complexity has increased and thus multi attribute algorithm is taking sufficient amount of time for being computed.

Question 2

In the second question we have interpreted the rules output from the modified decision tree(logistic regression split) and the normal decision tree from sklearn. Since the modified tree could not be visualized in the form of a plot, we have printed the details of node splitting in a level order traversal. We have also created a visualization function for the same, but due to the large size of datasets, we preferred to print the values instead. Visualizing helped to compare and observe the rules output of both the trees.

Question 3

In the third question, we perform k-fold cross validation(k=5) to get the best performance of the decision trees. We have performed statistical tests to get a different insight, compare similarities and differences, and check which of the two hypotheses,i.e, single attribute or multiple attribute, is better. We have done the same using the scipy tool in python.

Results

```
Single Attribute Accuracy for dataset fetal_health is: 0.8943661971830986
Single Attribute Recall for dataset fetal_health is: 0.8943661971830986
Single Attribute Precision for dataset fetal_health is: 0.899548463628636
Single Attribute F1-Score for dataset fetal_health is: 0.8965688705495971

Single Attribute Accuracy for dataset banking_dataset is: 0.9021607186210245
Single Attribute Recall for dataset banking_dataset is: 0.9021607186210245
Single Attribute Precision for dataset banking_dataset is: 0.8978487749725944
Single Attribute F1-Score for dataset banking_dataset is: 0.8997595299016761

Single Attribute Accuracy for dataset cervical_cancer is: 0.9534883720930233
Single Attribute Recall for dataset cervical_cancer is: 0.9534883720930233
Single Attribute Precision for dataset cervical_cancer is: 0.9503588860178007
Single Attribute F1-Score for dataset cervical_cancer is: 0.9515186532375612
```

Results

```
Multi Attribute Accuracy for dataset fetal_health is: 0.9272300469483568
Multi Attribute Recall for dataset fetal_health is: 0.9272300469483568
Multi Attribute Precision for dataset fetal_health is: 0.9246593066790796
Multi Attribute F1-Score for dataset fetal_health is: 0.9246760902616017

Multi Attribute Accuracy for dataset banking_dataset is: 0.9047098810390871
Multi Attribute Recall for dataset banking_dataset is: 0.9047098810390871
Multi Attribute Precision for dataset banking_dataset is: 0.8945751315570651
Multi Attribute F1-Score for dataset banking_dataset is: 0.8976631563563743

Multi Attribute Accuracy for dataset cervical_cancer is: 0.9767441860465116
Multi Attribute Recall for dataset cervical_cancer is: 0.9767441860465116
Multi Attribute Precision for dataset cervical_cancer is: 0.9757967269595177
Multi Attribute F1-Score for dataset cervical_cancer is: 0.9757593266187806
```

References

- https://link.springer.com/content/pdf/10.1007/s10994-005-0466-3.pdf
- https://github.com/python-engineer/MLfromscratch/blob/master/mlfromscratch/decision_tree.py