Machine Learning Assignment 2 - Part I

Sayan Sinha
16CS10048

INTRODUCTION

A decision tree is a decision support tool that uses a tree-like model of conditional statements and their possible consequences, such that the conditions match certain criteria of the concerned data. It is one way to display an algorithm that only contains conditional control statements.

MODEL

A dataset is already provided which is divided into train and test sets. The dataset are presented as comma separated values is saved in the root directory.

A decision tree classifier is trained using Information Gain and Gini Split impurity scores. Information gain is given by:

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

where entropy is defined as the minimum number of bits required to represent a set of data, i.e.

$$Entropy(t) = -\sum_{j} p(j \mid t) \log_2 p(j \mid t)$$

Higher the information gain, better classification results are obtained by choosing that feature as the next node of the tree.

Gini Split is given by:

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

where the Gini index is given by

$$GINI(t) = 1 - \sum_{j} [p(j \mid t)]^2$$

Gini is a measure of impurity. Hence, lower the Gini split, better classification results are obtained by choosing that feature as the next node of the tree.

Scikit-learn is an open source machine learning library for the Python programming language, which was born out of a GSoC project in 2017. It supports regression, clustering and classification based machine learning algorithms.

The model works with a recursive approach, in which the tree creator is a called by the possible child nodes again and again.

Interestingly, for the given data, the decision tree classifier trained using Information Gain and Gini Split are the same. Both give 100% accuracy on training and test sets.

```
maintenance = high
| capacity = 2 : no
capacity = 4 : no
capacity = 5 : yes
maintenance = low : yes
maintenance = med
| price     = high : yes
price = low : no
price = med
      | airbag  = no : no
airbag  = yes : yes
```

price	maintenance	capacity	airbag	profitable	label from model	label from scikit
med	high	5	no	yes	yes	yes
low	low	4	no	yes	yes	yes

Table 1: Labels on test set using information gain

price	maintenance	capacity	airbag	profitable	label from model	label from scikit
med	high	5	no	yes	yes	yes
low	low	4	no	yes	yes	no

Table 2: Labels on test set using Gini split

Therefore test accuracy using the generated model was 100% but using scikit learn it was at times 50% while sometimes it was 100%. This is because scikit learn uses a non-deterministic (randomised) approach to choosing the next node when they have same impurity.

Max Information Gain obtained using the model created was 0.186 whereas the minimum value of Gini Split was 0.384. 0.417 was the root node Gini index split using Scikit and the information gain is -0.396.

CONCLUSION

We conclude that Decision Tree classifier is indeed a suitable method for making predictions on data of the given nature.

REFERENCES

- Safavian, S. Rasoul, and David Landgrebe. "A survey of decision tree classifier methodology." IEEE transactions on systems, man, and cybernetics 21.3 (1991): 660-674.
- Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2.3 (2002): 18-22.