Mathematical Essay on Decision Tree Classifier

Chaithanya Krishna Moorthy
Dept. of Physics
Indian Institute of Technology, Madras
Chennai, India
ph17b011@smail.iitm.ac.in

***Abstract*—In this assignment, I have used the Decision Tree Classifier on a dataset containing the features of samples of a car and predict its degree of acceptance. The decision tree model works very well and gives an accuracy of 98% on unseen data.**

***Index Terms*—Decision Trees, Classification, Machine Learning, Supervised Learning**

I. Introduction

Classification models can be used in cases when gathering enough data points can be expensive. An example of this is judging if a car is acceptable in terms of safety. Performing multiple crash tests can be very expensive and hence, using a machine learning to classify can he helpful.

Decision Tree Classifier is classification technique that builds a tree based on the decision principles learned from the training model. Decision trees are very popular because of their interpretability and ease of implementation on categorical and numerical data and are used for both regression and classification.

We can use decision trees for the classification task at hand - to build a classifier for quality acceptance of cars based on their features, such as safety degree, price, capacity and maintenance costs. Doing so, we can understand the criterion that decide if a car is of good quality or not and also understand which features have the most weightage in this decision.

In this study, the a Decision Tree is built for the classification of cars based on its features. Section II will be on the principles of Decision Trees, section III will include Data Cleaning, Exploratory Data Analysis and applying the Desicion Trees Classifier on the cleaned data. Section IV will be conclusions and the study will end with references.

II. Principles of the Decision Tree Classifier

The Decision Tree classifier uses a flowchart with a tree-like structure, that contains nodes - conditions on the values that features take, to divide into branches. These conditions can be binary ('yes'/'no') or can have more divisions.

A. Mathematical model of Decision Tree Classifier

The feature space is split into regions that contain only/mostly datapoints belonging to a single class. The conditions on the values of the features are the basis of splitting the feature space. The leaf nodes (nodes with no further branching) generally corresponding to a single class. An example of a decision tree is shown in figure 1.

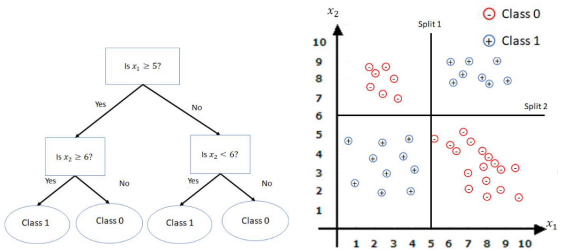

Fig. 1. Example of a decision tree dividing the feature space into regions containing examples of a single class [6]

For classification, the features are split accordingly to maximize the Gini index/Entropy. The splitting is stopped according to the criterion - the leaf nodes contain pure / majority class examples.

B. Advantages of decision trees

- Easy to interpret and explain
- They do not require any domain knowledge for feature selection
- Can easily handle both numeric and categorical data

C. Disadvantages

- Can overfit very easily - giving large trees that memorize the training data

Over-fitting is avoided by restricting tree depth and number of nodes, and tuning hyperparameters.

D. Metrics for Evaluation

Accuracy is defined as:

Accuracy = Number of correct predictions from the model/Total number of predictions

Accuracy is a good measure to see if our model is able to make correct predictions, implying that the input features influence the output features.

III. Using Decision Tree Classifier for the Dataset

The dataset contains example cars with their features specified and are labeled if they are acceptable or not, and how good they are, if accepted (4 classes)[5]. The data has 1728 rows and contains the columns as is table 1. The dataset was divided in 80-20 ratio as training-test set to evaluate the performance of the model on new data.

Feature	Description	Key
buying	buying price	vhigh, high, med, low
maint	Price of the maintenance	vhigh, high, med, low
doors	Number of doors	2, 3, 4, 5, more
persons	Capacity in terms of persons to carry	2, 4, more
lug_boot	The size of luggage boot	small, med, big
safety	Estimated safety of the car	low, med, high
Target	Target variable to predict	unacc, acc, good, vgood

TABLE I
FEATURES IN THE DATASET

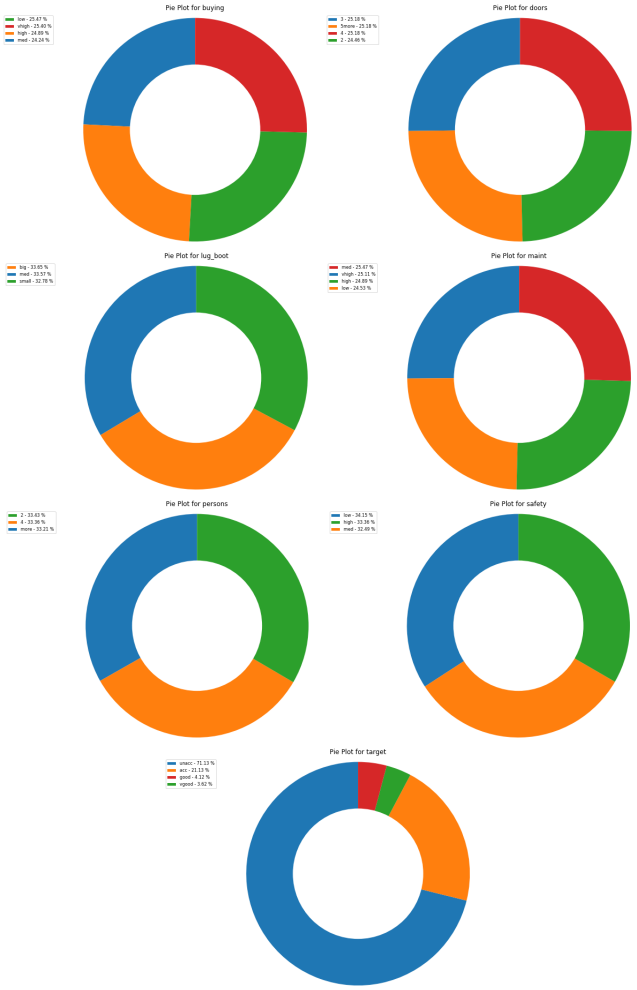

Fig. 2. Pie charts of categorical features and target variable, showing their size in the total dataset

1) Data cleaning and Exploratory Data Analysis: Pie plots of all the features that are categorical is as in figure 2. The data contains ordinal variables which were converted by OrdinalEncoder of scikit-learn module in Python to numerical values which are ordered (0,1,2..).

2) Correlations: The target features for the study is 'target', which indicates the degree of acceptance level of the car. Rest of the columns are taken as features.
The correlation matrix heatmap for every pair of columns is shown in figure 4. As seen in the figure, the correlations between the figures is very low.

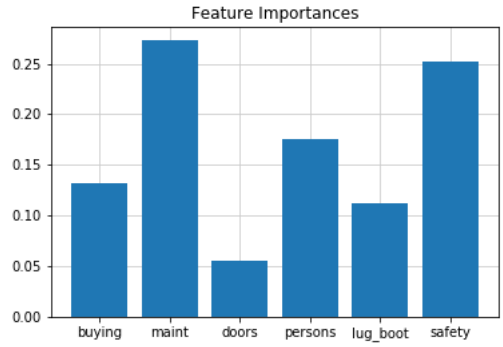

Fig. 3. Feature importance of each figure input to the model, in the final model

3) Applying Decision Tree Classifier: A Decision Tree Classifier model was built using sklearn's tree module. The resultant tree had 14 levels, with 91 tree nodes. The full decision tree is shown here. The weightage of each feature in the final model is shown in figure 3. The accuracy on the training model obtained is 1.00
The model was used on test data and the accuracy 0.98

4) Observations:

- As we can see in figure 2, the fractions of nominal values in the features is almost equal, but there is some imbalance in the target variable's class frequencies.

IV. Conclusions

From the model that was built, we can see that the Decision tree classifier performed exceptionally well for this dataset, giving accuracies of 100% on the the training and 98% on the test dataset. This high accuracy on the training set also indicates slight overfitting, which can happen in decision trees. The overfitting can be avoided by using methods such as

- Restricting the depth (number of levels) of the tree
- Restricting the number of leaf nodes
- Bagging technique (resampling and building an ensemble of small decision trees)
- Random Forests (choosing a subset of features for each small tree in an ensemble of decision trees)

These will be explored in the next assignments.

V. References

[1] Waskom, L.: seaborn: statistical data visualization, *Journal of Open Source Software, v6* [2] Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion et al.: Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research, v12* [3] Hunter J. D: Matplotlib: A 2D graphics environment, *Computing in Science & Engineering, v9*

[4] Andreas C. Müller and Sarah Guido: Introduction to Machine Learning with Python: A Guide for Data Scientists [5]

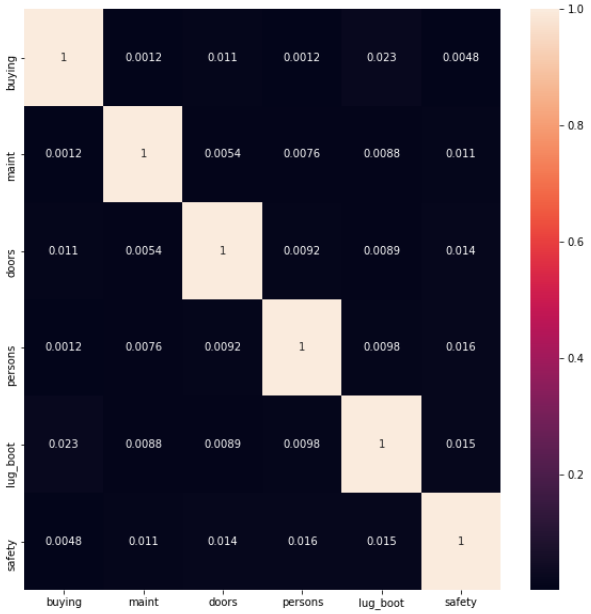

Fig. 4. Correlation between all features

Sources: (a) Creator: Marko Bohanec (b) Donors: Marko Bohanec (marko.bohanec@ijs.si) Blaz Zupan (blaz.zupan@ijs.si)
[6] Desision Trees notes, Data Analytics Laboratory, 2021