# EE6227 Assignment 2 - Handling Missing Values and Outliers in Binary Classification Tree

1st Nurfitri Anbarsanti - G2104045K
*School of Electrical and Electronic Engineering*
*Nanyang Technological University*
Singapore
nurfitri006@e.ntu.edu.sg

***Abstract*—In this paper, we present a binary classification tree model, leveraging the Interquartile Range (IQR) method to detect and handle outliers inside the updated training data. We have performed sets of experiments that have used those techniques and provide the result in this paper. The provided code repository further enhances the accessibility of our work, promoting open and collaborative research practices.**

***Index Terms*—Bayes Decision Rule, Naive Bayes, and Linear Discriminant Analysis**

## I. Introduction

To achieve better prediction accuracy, the performance of machine learning highly depends on the quality of data. At least two problems often come in the data preprocessing stage: missing elements in the data set and the dataset's outliers.

### *A. Handling Outliers*

Handling outliers in our data is a vital step in the data preprocessing stage. Some characteristics of the data outliers are its abnormally high or low value or missing value (Not a Number (NaN)) values. There are several approaches to handle them [?]:

- Removal. This approach will be taken if the number of outliers is too many.
- Transformation. A transformation can be applied to our data to reduce the impact of outliers. Common transformations include log, square root, or cube root transformation.
- Imputation. Another approach is to change them by the mean, median, or mode value. The value of the data that is most similar can be used to change the outliers.
- Capping. We could cap them by using the Interquartile Range (IQR) measure.
- Use a robust model. Because some models are more sensitive to outliers than others, a more robust model can handle the outliers.

### *B. Interquartile Range (IQR*

Statistical spread in a dataset could be measured by The Interquartile Range (IQR). IQR is the difference between the first quartile (dataset's 25th percentile) and the third quartile (dataset's 75th percentile). The IQR gives the range within which the central 50% of the data values lie.

$$IQR = Q3 - Q1 \tag{1}$$

To identify outliers using the IQR, we can calculate the ”fences” that define the acceptable range for data values. The lower and upper fences are calculated as follows:

- Lower Fence: $Q1 - k \times IQR$
- Upper Fence: $Q3 + k \times IQR$

The value of $k$ determines how stringent the criterion is for identifying outliers. Common choices for $k$ are $1.5$ and $3.0$. Value $k = 1.5$ is often used for mild outlier identification, while value $k = 3.0$ is used for recognizing extreme outliers.

The calculation and interpretation of the IQR are provided in some statistical textbooks such as [?]

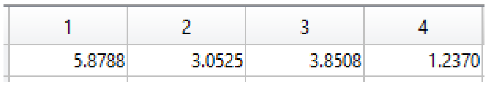
## II. Experiment

We are given the training data with dimension $120 \times 4$, its label with the extent $120 \times 1$, and the test data with size $30 \times 4$. We are inquired to decide whether there are missing values and outliers in the training data. The handling of missing values and outliers is required. Other than that, a Binary Classification Tree will be implemented in this study.

### *A. Handling Missing Values*

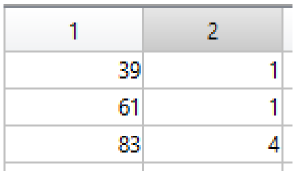
The dataset is assumed to have missing values since it was impossible to get its mean and other statistical measure. Then, we deal with the missing values with these approaches:

- Loading the data from the Excel files.
- Separating the training data and its labels from the raw Excel files.
- Calculating the mean by ignoring missing values (NaN) shown in Figure 1.
- Scan the indexes of missing values and build a list that consists of indexes of missing values that are shown in Figure 2.
- Impute the missing values with the mean we already attain in step 3.
- Updating new train data after imputation.



| 1 | 2 | 3 | 4 |
|---|---|---|---|
| 5.8788 | 3.0525 | 3.8508 | 1.2370 |

Fig. 1. Calculated Mean Ignoring Missing Values



| 1 | 2 |
|---|---|
| 39 | 1 |
| 61 | 1 |
| 83 | 4 |

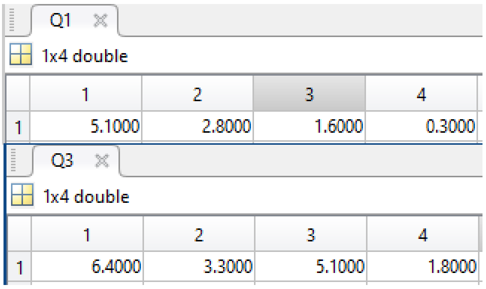
Fig. 2. List of Indexes of Missing Values

The code of this study can be accessed in [3].

### B. Handling Data's Outliers

We implemented the IQR measure to find and impute the outliers inside the updated training data. We deal with the data's outliers with these approaches:

- Defining the percentage for $Q1$ and $Q3$. In this study, we choose $Q1$ as the $25^th$ percentile and $Q3$ as the $75^th$ percentile, shown in Figure 3.
- Defining the value of $k$ to calculate the fences that characterize the acceptable range for data values. In this study, value $k = 1.5$ is chosen.
- Scan the indexes of the outliers and build a list that consists of indexes of outliers that are shown in Figure 4.
- Impute the outliers with the median of the updated training data.
- Updating new train data after imputation of outliers by the median.



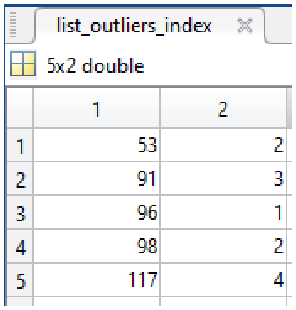
1x4 double

|  | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 5.1000 | 2.8000 | 1.6000 | 0.3000 |

1x4 double

|  | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 6.4000 | 3.3000 | 5.1000 | 1.8000 |

Fig. 3. Q1 and Q3 od the dataset



5x2 double

|  | 1 | 2 |
|---|---|---|
| 1 | 53 | 2 |
| 2 | 91 | 3 |
| 3 | 96 | 1 |
| 4 | 98 | 2 |
| 5 | 117 | 4 |

Fig. 4. List of Indexes of Outliers

The code of this study can be accessed in [3].

### C. Binary Classification Tree Implementation

After updating new training data with the imputation of its missing values by its mean and the imputation of its outliers by the dataset's median, the binary classification tree is implemented by these stages:

- Create and Train the Classification Tree in Matlab
- Visualize the tree that is shown in Figure 5.
- Predict the class or label of the test data as shown in Figure 6.

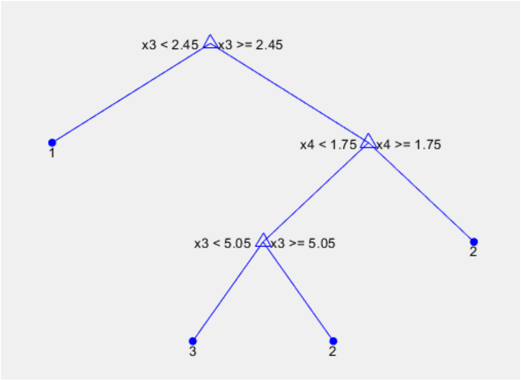
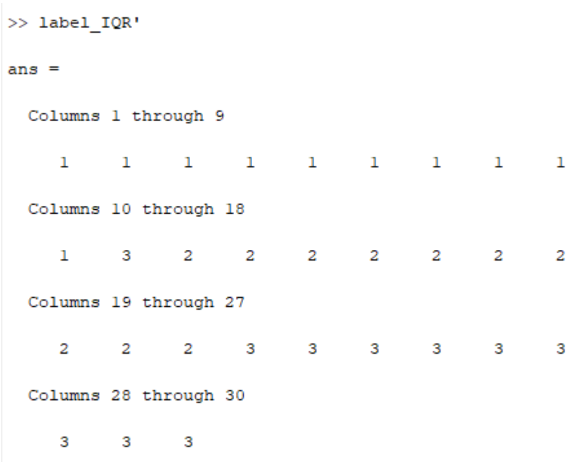


Fig. 5. Q1 and Q3 of the dataset



```
>> label_IQR'

ans =

  Columns 1 through 9

     1     1     1     1     1     1     1     1     1

  Columns 10 through 18

     1     3     2     2     2     2     2     2     2

  Columns 19 through 27

     2     2     2     3     3     3     3     3     3

  Columns 28 through 30

     3     3     3
```

Fig. 6. List of Indexes of Outliers

The code of this study can be accessed in [3].

## III. Conclusion

This study embarked on an exploration of data preprocessing techniques, specifically targeting missing values and outliers, and their subsequent impact on a Binary Classification Tree model. We systematically addressed missing values through imputation, ensuring that our dataset was complete and primed for further analysis. Our approach to outlier detection and handling, leveraging the Interquartile Range (IQR) method.

We employed a Binary Classification Tree, evaluating its performance in the context of our preprocessed dataset. The visualizations and detailed steps provided throughout the paper offer a transparent view of our methodology. The provided code repository further enhances the accessibility of our work, promoting open and collaborative research practices.

## References

[1] C.C Aggarwal, Outlier Analysis.

[2] D.S. Moore, G.P, McCabe, B.A. Craig. Introduction to the Practice of Statistics.

[3] N. Anbarsanti. ”Handling Missing Values and Outliers in Binary Classification Tree” Available: Github. [Online]. Available: https://github.com/anbarsanti/datahandling_binary. [Accessed: October 29, 2023]