CS 459 – Introduction to Machine Learning

Assignment No. 2

Wine Quality Classification

Submitted To:

Dr. Ayaz Hussain

Submitted By:

Muhammad Nasir Khurshid - 04071913019

1. Introduction

The objective of this assignment is to conduct experiments on logistic regression and neural network models. These models will be trained using varying numbers of features, training examples, and regularization parameter values. The analysis of graphs will help in identifying problems related to high bias and high variance. Additionally, the factors that contribute to reducing these issues will be discussed.

2. Data Set

The dataset was acquired from UCI Machine Learning website. The dataset consists of almost 4000 examples and 11 attributes:

1. Fixed Acidity
2. Volatile Acidity
3. Citric Acid
4. Residual Sugar
5. Chlorides
6. Free Sulfur Dioxide
7. Total Sulfur Dioxide
8. Density
9. ph
10. Sulphates
11. Alcohol

The ground truth variable is ***'quality'*** which can be any value from 0-10 which indicates the quality of the wine.

3. Data Preprocessing

In data preprocessing, only duplicate values were removed and nothing else was required to be done as the data was already clean.

4. Results & Observations

4.1. Effect of Changing Number of Features

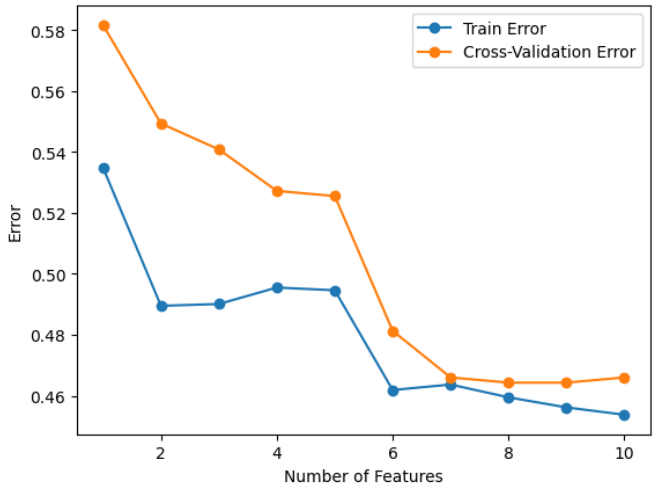

Figure 1. Logistic Regression

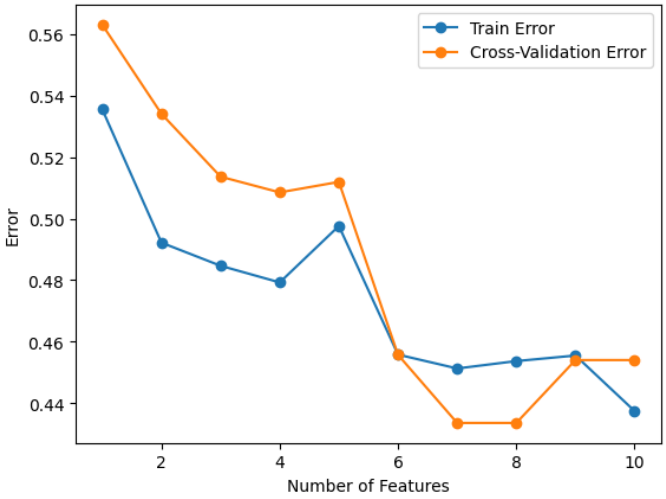

Figure 2. Neural Network

Based on the graphs, we observe that when the model has fewer features, it suffers from high bias. This means that the model oversimplifies the data and may not capture its complexities accurately. However, we find that by increasing the number of features, we can alleviate the problem of high bias i.e., incorporating more features helps the model to better understand and represent the complexities present in the data, resulting in improved performance.

4.2. Effect of Changing Regularization Parameter

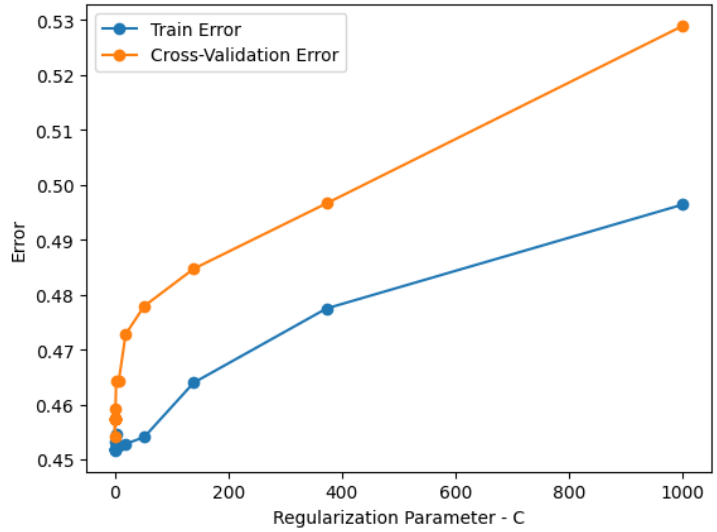

Figure 3. Logistic Regression

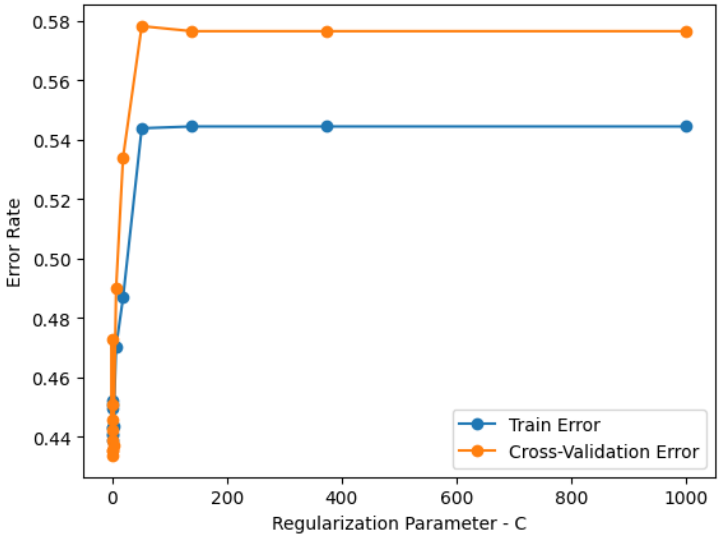

Figure 4. Neural Network

Based on the graphs, it is evident that when the regularization parameter has a large value, the model encounters high bias. This implies that the model makes weights very small by penalizing them with larger regularization parameter and oversimplifies the data and fails to capture its complexities accurately. However, we find that by reducing the value of the regularization parameter, we can overcome the problem of high bias i.e., lowering the regularization parameter allows the model to be more flexible and better adapt to the intricacies of the data, resulting in improved performance and reduced bias.

4.3. Effect of Changing Number of Examples

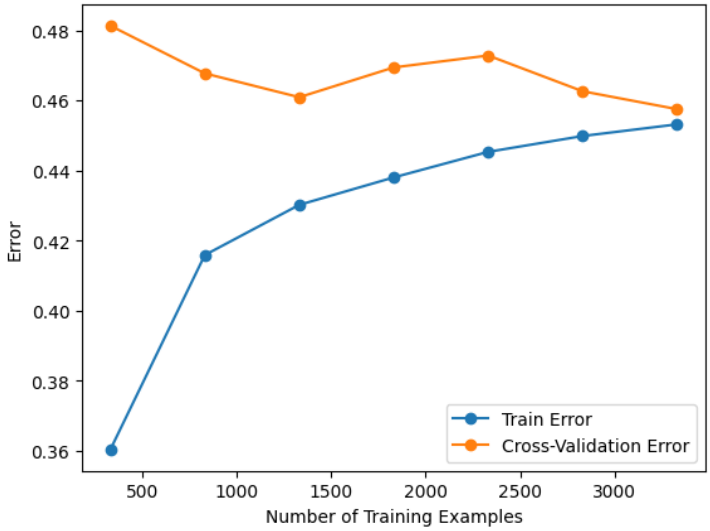

Figure 5. Logistic Regression

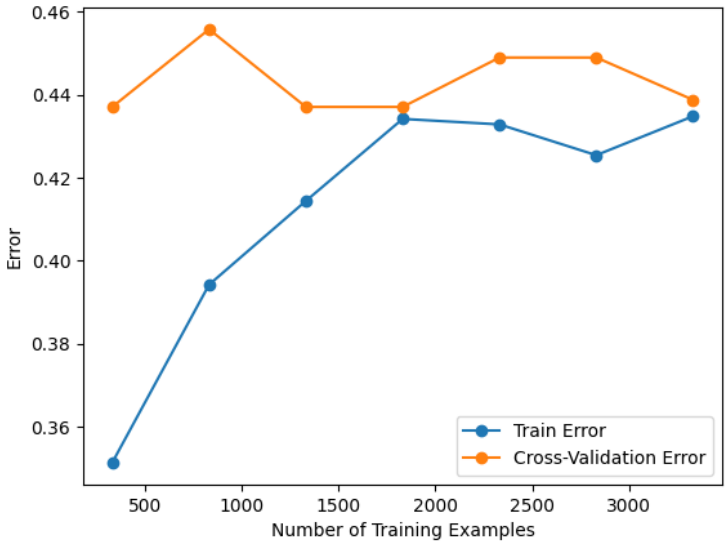

Figure 6. Neural Network

Based on the graphs, we can observe that when there are fewer training examples, the model will suffer from high variance. This means that the model becomes overly sensitive to the specific data points in the training set and struggles to generalize well to new, unseen data. However, when we increase the number of training examples, we are able to reduce the problem of high variance. By exposing the model to more diverse instances, it can better understand the underlying patterns and variations within the data, leading to improved generalization and a reduction in variance.