CS554 Introduction to Machine Learning Multivariate Classification Methods Report

Mert Erkol

April 12, 2023

1 Introduction

In this study Nearest Mean and Nearest Neighbour Classifiers fitted to the given iris dataset and their results evaluated with accuracy, errors and confusion matrix

2 Background

The goal is to fit a model to given data that generalizes the problem and predicts next outcomes when new input is given. In this section Nearest Mean and Nearest Neighbour Classifier topics explained in detail.

2.1 Nearest Mean Classifier

The nearest mean algorithm is a classification algorithm that assigns the given point to the closest class mean in the training data. To be able to fit the model correctly means of each feature in specific class has been calculated according to the formula [1]. Given a set of k classes $C_1, C_2, ..., C_k$ with $n_1, n_2, ..., n_k$ observations, where $\boldsymbol{x_{ij}}$ denotes the j-th observation from class i, the mean vector for class i is defined as

$$m_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{i_j} \tag{1}$$

Then, given a new data point $\boldsymbol{x}$, the nearest mean classifier assigns $\boldsymbol{x}$ to the class with the closest mean

$$\text{argmin}_{i=1}^{k} ||x - m_i|| \tag{2}$$

where $| \cdot |$ denotes the Euclidean distance.

2.2 Nearest Neighbour Classifier

The nearest neighbour algorithm is a simple classification algorithm that assigns the given point to the closest neighbour in the training data the formula defined as

$$\text{argmin}_i ||x - x_{i_j}|| \tag{3}$$

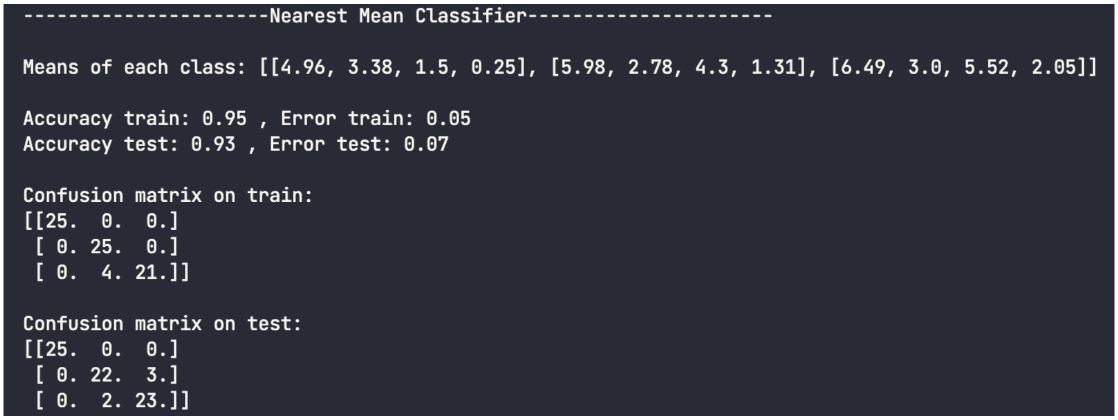
Given a set of k classes $C_1, C_2, ..., C_k$ with $n_1, n_2, ..., n_k$ observations, $\boldsymbol{x_{ij}}$ denotes the j-th observation from class i, then given a new data point $\boldsymbol{x}$, the nearest neighbor classifier assigns $\boldsymbol{x}$ to the class of its closest neighbor

3 Methodology

To complete the project Python@3.9 has been used as a programming language. No external libraries have been used during calculating the mean or the distance between points The structure is follows, Read the train and test data, fit the nearest mean and nearest neighbour models, evaluate them on test data by calculating accuracy, error and confusion matrix. Also the four dimensional points of class means have been shown.

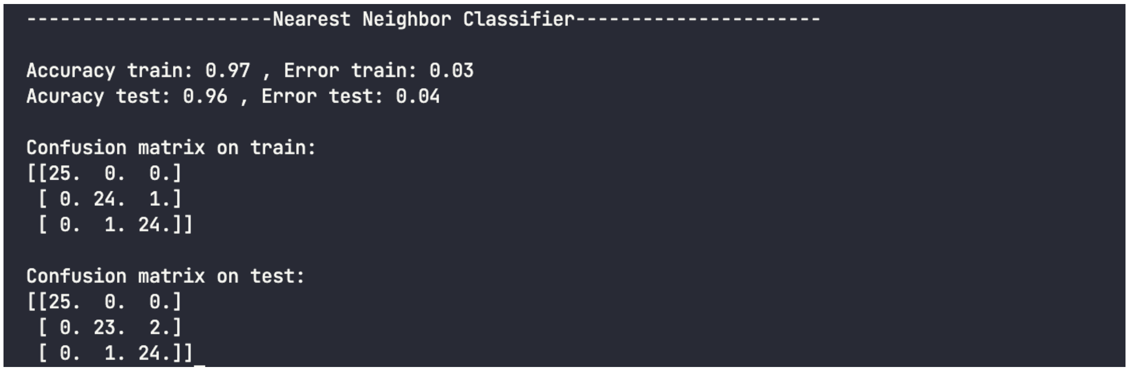
4 Results

In this section evaluation results of both models have been shown



```
----------------------Nearest Mean Classifier----------------------

Means of each class: [[4.96, 3.38, 1.5, 0.25], [5.98, 2.78, 4.3, 1.31], [6.49, 3.0, 5.52, 2.05]]

Accuracy train: 0.95 , Error train: 0.05
Accuracy test: 0.93 , Error test: 0.07

Confusion matrix on train:
[[25.  0.  0.]
 [ 0. 25.  0.]
 [ 0.  4. 21.]]

Confusion matrix on test:
[[25.  0.  0.]
 [ 0. 22.  3.]
 [ 0.  2. 23.]]
```

Figure 1: Evaluation results of Nearest Mean Classifier



```
----------------------Nearest Neighbor Classifier----------------------

Accuracy train: 0.97 , Error train: 0.03
Acuracy test: 0.96 , Error test: 0.04

Confusion matrix on train:
[[25.  0.  0.]
 [ 0. 24.  1.]
 [ 0.  1. 24.]]

Confusion matrix on test:
[[25.  0.  0.]
 [ 0. 23.  2.]
 [ 0.  1. 24.]]
```

Figure 2: Evaluation results of Nearest Neighbour Classifier

5 Discussion

As seen in the figure 1 and figure 2 the nearest neighbour classifier performed slightly better than nearest mean classifier. The performance difference between the nearest mean algorithm and the nearest neighbour algorithm could be due to the fact that the nearest mean algorithm only uses the class means to process a given point, as depicted in Figure 1. This means that the algorithm calculates the distance between the given point and the mean of each class in the training set, as given by the equation 2. In contrast, the nearest neighbour algorithm finds the closest training instance to the given point and assigns the same class to the point, as given by equation 3, which may lead to better results in some cases. However, the simplicity of the nearest mean algorithm could be an advantage in scenarios where there are fewer training instances or when there is a limited number of features. Therefore, the choice of algorithm depends on the specific problem and the characteristics of the dataset. The class wise performance in the confusion matrices shows that the zero class decomposes from other classes both algorithms achieved 100 percent accuracy. Only mistakes that both algorithms made is between at class 1 or class 2, also class 0 decomposes so well that it has zero false positives on class 1 and class 2 on both confusion matrices.

6 Conclusion

In this study basic multivariate classification models implemented from scratch and evaluated with classification metrics. According to the results and the discussion section even these basic models can understand and generalize multi-sample feature problems. To advance this study dataset size, feature count can be increased and model performance change observed when more features and more samples are in hand.

References

[1] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2014.