A Project Report

On

Comprehensive Comparison of Different ML Models

BY

AMAN BADJATE	2017B3A70559H
GARVIT SONI	2017B3A70458H
KSHITIJ VERMA	2017B1A71145H

UNDER THE SUPERVISION OF

DR. BHANU MURTHY

SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS OF

BITS F464: MACHINE LEARNING

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI (RAJASTHAN)

HYDERABAD CAMPUS

APRIL, 2021

Contents

List of Figures

1 Introduction

We do a comparative study and analysis of the following Machine Learning models-

- Fischer Linear Discriminant
- Linear Perceptron
- Naive Bayes
- Logistic Regression
- Artificial Neural Networks
- Support Vector Machines

The models are imported from the sklearn libraries. The data is scaled using the preprocessing sklearn library, StandardScaler and a 7-fold cross validation is done for each model using the sklearn.

2 Fischer Linear Discriminant

Fisher linear discriminant is a classification model that classifies the data by finding a linear discriminant in the data that separates the two classes. The algorithm finds a projection vector in one dimension onto which if the data is projected it gives the best separation in data. The algorithm finds the projection by maximizing the difference in means of the two classes and minimizing their within class variance.
Accuracy of Training Data: 98.68
Accuracy of Testing Data: 98.72

3 Linear Perceptron

The Perceptron has many inputs (often called features) that are fed into a Linear unit that produces one binary output. Therefore, perceptrons can be applied in solving Binary Classification problems where the sample is to be identified as belonging to one of the predefined two classes.

The function f(x) = b + w.x where w.x is a linear combination of weight and feature vectors and b is the bias. The weights signify the effectiveness of each feature xi in x on the model's behavior.

Accuracy of Training Data: 91.92

Accuracy of Testing Data: 91.33

4 Naive Bayes

Naive Bayes is a very famous classifier which has found applications in many areas including Natural Language Processing (NLP). It is a very simple algorithm to understand and gives above average performance in different tasks. This classifier determines the most probable class using Bayes' theorem.

Accuracy of Training Data: 97.67

Accuracy of Testing Data: 97.69

5 Logistic Regression

Logistic Regression is an algorithm mostly used for classification problems. We have a binary classification problem here. This algorithm uses the sigmoid function to calculate the probability of a datapoint to belong to a particular class. Sigmoid function is basically an S-shaped curve which can take any real value lying between 0 and 1. In this assignment, the aim is to detect forged banknotes by classifying a datapoint as 0 or 1. using Bayes' theorem.

Accuracy of Training Data: 98.93

Accuracy of Testing Data: 98.94

6 Artificial Neural Networks

Artificial neural network is a machine learning technique used for classification problems. ANN is a set of connected input output network in which weight is associated with each connection. It consists of one input layer, one or more intermediate layer and one output layer.

Accuracy of Training Data: 97.47

Accuracy of Testing Data: 97.33

7 Support Vector Machines

SVM is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

Accuracy of Training Data: 92.64

Accuracy of Testing Data: 92.65

8 Box Plots

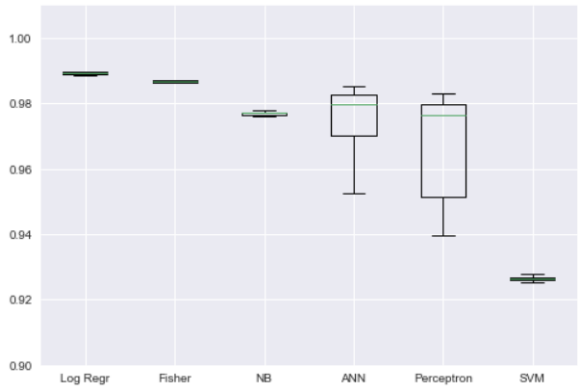

Figure 1: Box Plot for Testing Accuracy's

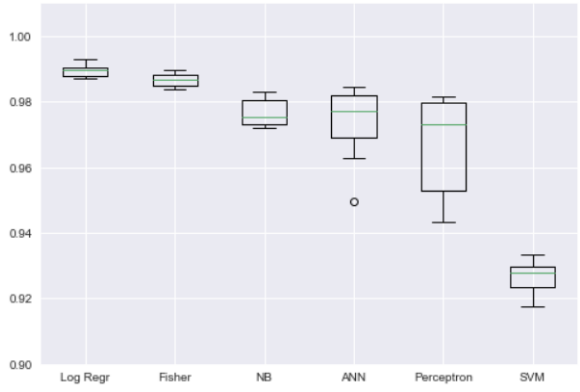

Figure 2: Box Plot for Testing Accuracy's

Above two figure represent the box plots for training and testing accuracy's for all models respectively.

9 Conclusions

- The box plots of the training accuracies show that Logistic Regression and Fischer Linear Discriminant perform the best while training the datasets, while SVM performs the worst.
- Linear Perceptron has the highest variance as shown by its boxplot. This could be due to the fact it makes random predictions of the weights while training.
- Median Accuracies for Logistic Regression and Fischer's LDA are higher than the others'.
- The high accuracy of Fisher LDA and Perceptron could indicate that the the data is in fact linearly seperable.