Comprehensive Comparison

Mudit Chaturvedi 2018A7PS0248H

Ashna Swaika 2018A7PS0027H

Rohan Sachan 2018B3A70992H

Overview

We have used scikit-learn (sklearn) library to import different models - Fisher Linear Discriminant, Linear Perceptron, Naive Bayes, Logistic Regression, Support Vector Machines and Artificial Neural Networks.

Implementation and Model

The given statement is a binary classification problem. The dataset consists of 18185 samples with 10 columns and 2 classes (Jasmine and Gonen).

1. Dataset is loaded into a pandas dataframe and then into a numpy array with the ID column being dropped.
2. The values are normalized according to the rule
   Value_{new} = (Column Max – Value_{old}) / (Column Max – Column Min)

Models imported from sklearn are :

1. Fisher Linear Discriminant - LinearDiscriminantAnalysis()
2. Linear Perceptron - Perceptron()
3. Naïve Bayes – GaussianNB()
4. Logistic Regression - LogisticRegression()
5. SVM – SVC()
6. ANN - MLPClassifier()

- **Results**

1. **Fisher Linear Discriminant**

   a. Train Accuracy

      Over cross folds : 0.983, 0.982, 0.983, 0.981, 0.983, 0.983, 0.982
      Mean : 0.982357

   b. Test Accuracy

      Over cross folds : 0.977, 0.981, 0.983, 0.985, 0.984, 0.980, 0.985
      Mean : 0.982238

2. **Perceptron**

   a. Train Accuracy

      Over cross folds : 0.985, 0.986, 0.980, 0.986, 0.986, 0.967, 0.975
      Mean : 0.980992

   b. Test Accuracy

      Over cross folds : 0.983, 0.985, 0.983, 0.988, 0.989, 0.969, 0.975
      Mean : 0.981798

3. **Naïve Bayes**

   a. Train Accuracy

      Over cross folds : 0.987, 0.986, 0.985, 0.985, 0.984, 0.986, 0.985
      Mean : 0.985464

   b. Test Accuracy

      Over cross folds : 0.980, 0.984, 0.985, 0.987, 0.990, 0.983, 0.988
      Mean : 0.985263

4. **Logistic Regression**

   a. Train Accuracy

      Over cross folds : 0.988, 0.987, 0.987, 0.986, 0.987, 0.987, 0.987
      Mean : 0.987114

b. Test Accuracy

Over cross folds : 0.983, 0.987, 0.985, 0.989, 0.988, 0.987, 0.989
Mean : 0.987077

5. **SVM**

a. Train Accuracy

Over cross folds : 0.989, 0.989, 0.989, 0.988, 0.989, 0.988, 0.988
Mean : 0.988617

b. Test Accuracy

Over cross folds : 0.983, 0.988, 0.987, 0.989, 0.989, 0.989, 0.989
Mean : 0.988177

6. **ANN**

a. Train Accuracy

Over cross folds : 0.988, 0.989, 0.988, 0.987, 0.988, 0.988, 0.987
Mean : 0.988003

b. Test Accuracy

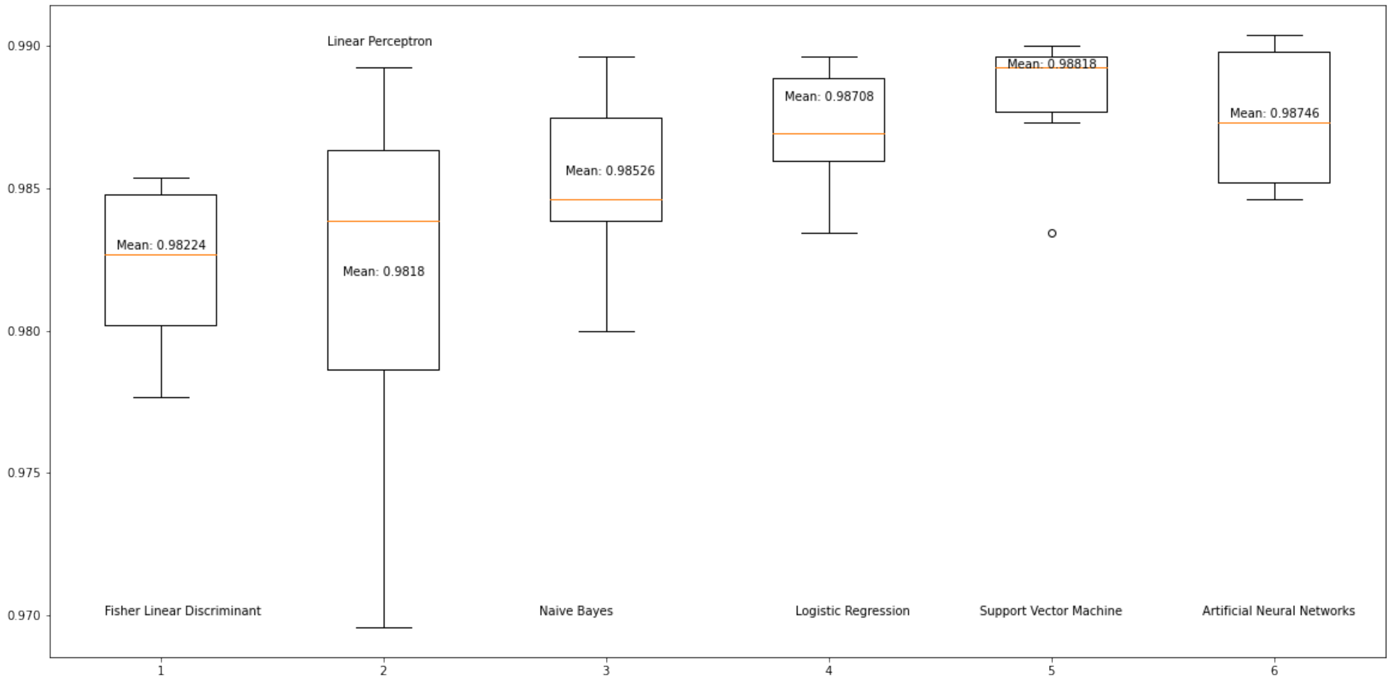
Over cross folds : 0.985, 0.985, 0.985, 0.990, 0.990, 0.987, 0.990
Mean : 0.987462

• Explanation

Based on testing accuracies, SVM gives the best results while LDA gives the worst.

- Higher accuracy of SVM can be attributed to its ability to avoid overfitting.
- ANN performs better than Perceptron model during both training and testing. This might originate from the fact that Perceptron is a single layered Neural Network, hence it is less complex and doesn't fit data as well as ANN, which is multi layered Neural Network.
- Logistic Regression gives better results than Naïve Bayes, because it doesn’t assume the features to be independent of one another which Naïve Bayes does (in the dataset the features might not be independent).
- Perceptron gives the most variance as it is affected a lot by the order in which training examples are encountered and initialization of weights
- Overall Logistic regression tends to perform better than LDA, which is very sensitive to outliers.

- **Box Plot**


0.990
0.985
0.980
0.975
0.970
Linear Perceptron
Mean: 0.98224
Mean: 0.9818
Mean: 0.98526
Mean: 0.98708
Mean: 0.98818
Mean: 0.98746
Fisher Linear Discriminant
Naive Bayes
Logistic Regression
Support Vector Machine
Artificial Neural Networks
1
2
3
4
5
6