Report on

Comprehensive Comparison of various ML algorithms

Team

1. Sai Satvik Vuppala (2017B4A71449H)
2. Shreya Kotagiri (2017B3AA0296H)
3. Dhanush Karupakula(2017B3A71011H)

Introduction

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

- Fisher Linear Discriminant
- Linear Perceptron
- Naive Bayes
- Logistic Regression
- Artificial Neural Networks and,
- 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 in built methods.

Libraries Used

1. Numpy
2. Pandas
3. Matplotlib
4. Sklearn

Comparative analysis

The accuracies of the different ML models over each of the 7 folds of the cross validation are:

Logistic Regression (LR):

Logistic regression is a machine learning algorithm for classification. In this algorithm, the probabilities describing the possible outcomes of a single trial are modelled using a logistic function.

[0.8537336412625096, 1.0, 1.0, 1.0, 1.0, 1.0, 0.9988448209472468]

Support Vector Machine (SVM):

Support vector machines (SVMs) are a set of supervised learning methods used for classification. They are effective for high dimensional spaces and cases where the number of dimensions is greater than the number of samples.

[0.8929946112394149, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0]

Linear Discriminant Analysis (LDA):

The Fisher Linear Discriminant Analysis is a classification technique that identifies the linear combination of features that characterizes or separates two or more classes. The model fits a Gaussian density to each class, assuming that all classes share the same covariance matrix.

[0.9318706697459584, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0]

Linear Perceptron (LP):

A Perceptron is a linear classification algorithm that makes its predictions based on a linear predictor function combining a set of weights with the feature vector. It uses the Gradient Descent Algorithm to update the weights based on misclassified points on each iteration.

[0.6874518860662048, 1.0, 1.0, 1.0, 1.0, 1.0, 0.9996149403157489]

Naive Bayes (NB):

Naive Bayes algorithms are a set of supervised statistical classification machine learning algorithms based on the Bayes probability theorem.

Bayes theorem states that:

$$P(A|B) = \frac{P(B/A)*P(A)}{P(B/A)*P(A)+P(C/A)*P(A)}$$

An important assumption made by Bayes theorem is that the value of a particular feature is independent from the value of any other feature for a given class.

[0.9491916859122402, 0.9865280985373364, 0.9911470361816782, 0.9923017705927637, 0.9888375673595073, 0.9888375673595073, 0.9626492106276473]

Artificial Neural Networks (ANN):

Artificial neural networks is a machine learning technique used for classification problems. ANN is a set of connected input output networks in which weight is associated with each connection. It consists of one input layer, one or more intermediate layer and one output layer. A Multi Layer Perceptron is a supervised learning technique with a feed forward artificial neural network through back-propagation that can classify non-linearly separable data.

[0.783679753656659, 1.0, 0.9992301770592764, 0.9892224788298691, 0.9988452655889145, 0.9838337182448037, 0.9834424335772045]

Model Performance and reasons

The average accuracy for LR is: 97.89397803156795%

The average accuracy for SVM is: 98.47135158913449%

The average accuracy for LDA is: 99.02672385351369%

The average accuracy for LP is: 95.52952609117077%

The average accuracy for NB is: 97.992756236724%

The average accuracy for ANN is: 95.00700649401273%

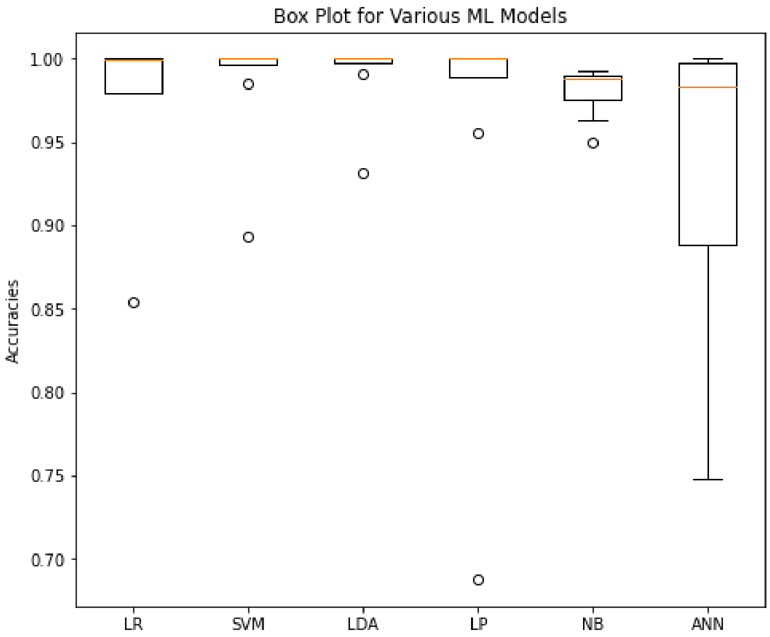
On ranking the models in the descending order of their average accuracies are as follows:

Fisher's Linear Discriminant, Support Vector Machines, Naive Bayes, Logistic Regression, Perceptron and Artificial Neural Network.

The best performing model is the **Fisher's Linear Discriminant Algorithm** and the worst performing model is the **Artificial Neural Network**.

- Fisher's linear discriminant has the highest accuracy implying that the data might be linearly separable
- Average accuracy for naive bayes algorithm is high due to outliers present in the first fold in all of the models
- All models except naive bayes and ANN have a 100% median accuracy
- ANN and linear perceptron have higher variances(accuracies ranging from 78% and 68% in ANN and LP respectively to 100%), due to random predictions of weights done by the models

Box Plots


Box Plot for Various ML Models
Accuracies
LR
SVM
LDA
LP
NB
ANN