BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI
HYDERABAD CAMPUS
Second Semester 2020-21
BITS F464 - Machine Learning
Assignment - 2C

Group Members

Danish Mohammad 2018A7PS0103H

Mridul Kumar Rai 2018AAPS0359H

Ketan Goyal 2018A8PS0900H

Comprehensive Comparison

Problem Statement

1) As for this assignment, you need to do a comparative study and analysis of the ML models you have studied till now, i.e. Fisher Linear Discriminant, Linear Perceptron, Naive Bayes, Logistic Regression, Artificial Neural Networks and Support Vector Machines. Note, only for this assignment you can use the Sklearn library to directly import models/methods and use them.
2) For all the models, use 7-fold cross-validation and generate a box plot over the test set accuracy over each fold. Visualize all box plots in a single image. That is, the image must contain six box plots, one for each model with the box plot denoting the variation of test set accuracy over each fold.
3) Try to vectorize your code as much as possible to make your computations faster and efficient. Do not hard code any parts of the implementation unless it is absolutely necessary.

What needs to be documented

1) A comparative analysis of the models and their accuracies (train and test).

Model	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7	Median
GaussianNB	0.983064	0.989222	0.993072	0.989607	0.986528	0.981909	0.912977	0.986528
SVC	0.655504	0.998845	1.0	0.999615	0.998845	0.996151	0.933	0.998845
LDA	0.859892	0.99923	0.999615	0.99846	0.998075	0.993841	0.906816	0.998075
Perceptron	0.824095	0.974981	0.992302	0.986528	0.985373	0.795227	0.90181	0.974981
LogisticRegression	0.677829	0.997306	1.0	0.999615	0.998845	0.997306	0.93377	0.997306
MLPClassifier	0.916089	0.981524	0.996151	0.998075	0.994611	0.991532	0.549095	0.991532

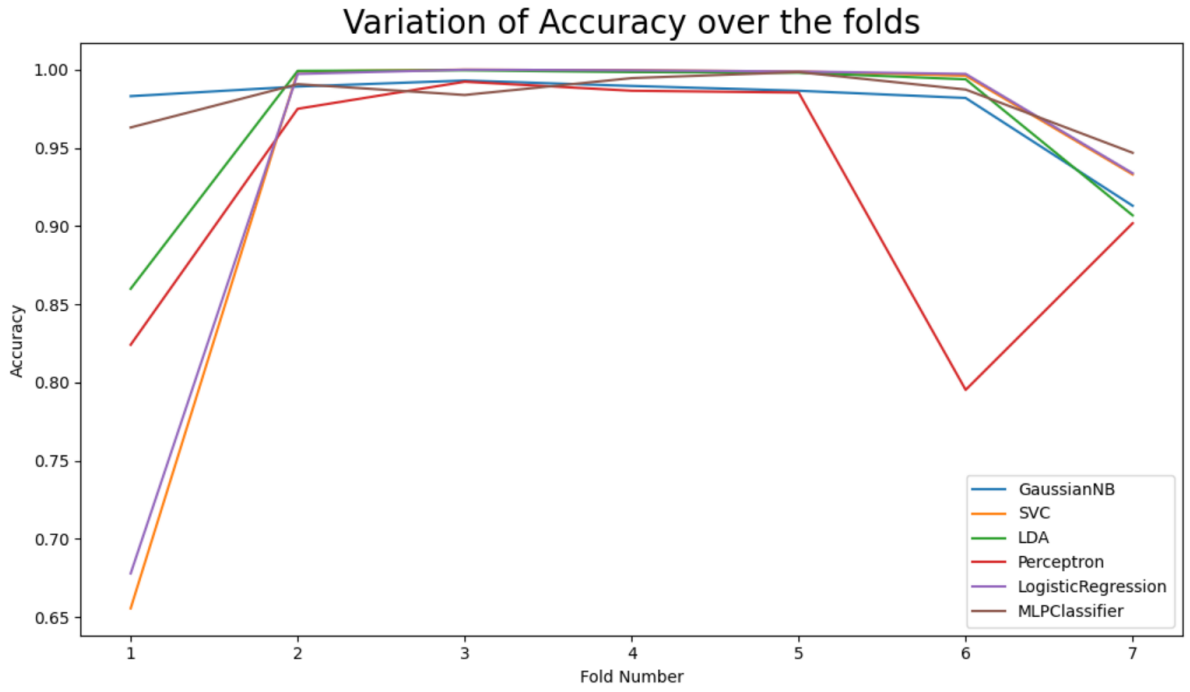
Maximum Accuracy: 1.0

In the figure attached above, we can see the accuracy of all the different models over the 7 folds. Some general observation points are that the SVC, using the linear kernel, seems to be the best performing model, this can be due to the fact that there is a clear separation between the

data classes. Then we see the Logistic Regression model which again uses 200 iterations for convergence although during the test run we did not achieve convergence as the training data was not scaled properly. Following this is the ANN which has a single hidden layer of 10 neurons using the relu activation function and it runs for a total of 200 epochs. It uses the Stochastic Average Gradient as the solver method. Then we see the NaiveBayes classifier which uses the Gaussian Naive Bayes algorithm for classification which basically means that the likelihood of all the features is assumed to be gaussian in nature. The worst performance is shown by the perceptron algorithm. In the case of the Perceptron model not only do we see a comparatively low accuracy but also comparatively low precision as clear from the above image.

Model	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7	Median
Gaussian NB	0.983064	0.989222	0.993072	0.989607	0.986528	0.981909	0.912977	0.986528
SVC	0.655504	0.998845	1.0	0.999615	0.998845	0.996151	0.933	**0.998845**
LDA	0.859892	0.99923	0.999615	0.99846	0.998075	0.993841	0.906816	0.998075
Perceptron	0.824095	0.974981	0.992302	0.986528	0.985373	0.795227	0.90181	**0.974981**
Logistic Regression	0.677829	0.997306	1.0	0.999615	0.998845	0.997306	0.93377	0.997306
MLP Classifier	0.963048	0.990762	0.983834	0.994611	0.99846	0.987298	0.946862	0.987298

Maximum Accuracy: 1.0



2) **The model that performed best and one that performed worst. (Do mention reasons why that certain model may have given the best or worst results.)**

Support Vector Classifier is the best performing model whereas Perceptron is the worst performing model. The SVM model works quite well when there is a clear margin of separation

between the classes and hence we get the highest accuracy in this case. Perceptron produces unsatisfactory results due to the data not being normalized and due to a slow learning rate. SVM also works better since it projects the data into higher dimensional spaces and then tries to separate them, where it might be easier. Perceptron relies on separability at lower dimensions, which might not be present.

3) The image containing box plots for each model.