BITS F464 : Machine Learning
Assignment - 2
Comprehensive Comparison

Omkar Pitale	Aditya Chopra	Anuradha Pandey
2019A7PS0083H	2019A7PS0178H	2019A7PS0265H

1 Introduction

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

2 Implementation and Model

The given problem statement is a binary classification problem.

1. The dataset given is a CSV file with 10 geometric properties, and 2 classes - jasmine and gonen.
2. The ID column is dropped, and is not used for further calculations.
3. The training data consists of the 10 attribues and the testing data consists of the class.

The models imported from sklearn were -

1. Logistic Regression - LogisticRegression() from sklearn
2. Fisher Linear Discriminant - LinearDiscriminantAnalysis()
3. Naive Bayes - GaussianNB()
4. Linear Perceptron - Perceptron()
5. Support Vector Machines - SVC()
6. Artificial Neural Networks - MLPClassifier()

The models were trained over 7 fold cross-validation. Hence, for each model, we have 7 training accuracies, which are averaged to find out the final training accuracies.

Training Accuracies

Model	LogReg	LDA	NB	SVM	ANN	Perc
fold1	0.98	0.98	0.97	0.92	0.98	0.92
fold2	0.98	0.98	0.97	0.92	0.96	0.94
fold3	0.98	0.98	0.97	0.92	0.93	0.94
fold4	0.99	0.98	0.97	0.92	0.98	0.94
fold5	0.98	0.98	0.97	0.92	0.76	0.97
fold6	0.98	0.98	0.97	0.92	0.97	0.97
fold7	0.98	0.98	0.98	0.93	0.98	0.76
avg	0.988	0.981	0.976	0.926	0.942	0.927

Models were trained and for testing, 7 folds cross validation was used. 7 accuracies were recorded for the 6 models. They were averaged out to get the final testing accuracy.

Testing Accuracies

Model	LogReg	LDA	NB	SVM	ANN	Perc
fold1	0.98	0.98	0.97	0.92	0.97	0.88
fold2	0.98	0.98	0.97	0.92	0.98	0.94
fold3	0.98	0.98	0.97	0.93	0.97	0.98
fold4	0.98	0.98	0.98	0.92	0.97	0.98
fold5	0.99	0.98	0.97	0.92	0.97	0.96
fold6	0.98	0.98	0.97	0.92	0.97	0.97
fold7	0.99	0.98	0.97	0.92	0.96	0.93
avg	0.988	0.986	0.977	0.927	0.975	0.954

3 Results

Based on testing accuracies, Logistic Regression has the best performance whereas SVM has the worst performance. Certain observations about training and testing accuraies -

1. Good performance of logistic regression over training and testing data shows that our dataset is linealy separable. This can also be inferred from the high accuracy of Fisher's Linear Discriminant.
2. 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.
3. Logistic Regression has a better accuracy as compared to Naive Bayes while testing. Logistic Regression belongs to the discriminative class of algorithms, and these algorithms usually make less assumptions as compared to Generative class of algorithms, to which Naive Bayes belongs.
4. SVM tries to maximize the margin between the closest support vectors whereas logistic regression maximize the posterior class probability.

Compilation of boxplots for the six models -