CS 4375

ASSIGNMENT 5

Names of students in your group:

Sriraam Ramakrishnan

Jithin Paul

Number of free late days used: 0

Note: You are allowed a total of 4 free late days for the entire semester. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

http://scikit-learn.org/stable/

http://machinelearningmastery.com

Dataset

pima-indians-diabetes

https://archive.ics.uci.edu/ml/machine-learning-databases/pima-indians-diabetes/pima-indians-diabetes.data

Preprocessing

The data was scaled as the range of the values of the attributes differed significantly. Data was comprised of attributes with varying scales . Hence the dataset was scaled using the **sklearn.preprocessing.MinMaxScaler** library.

Pseudo Code

Read the data from the URL into a dataFrame.

Pre -process the data.

Store all classifier functions in a list

for each classifier in the list

Split the data into 10 folds.

Run the classifier algorithm on the data.

Output the Average Accuracy and Average Precision

Evaluation Metric Used

We have used Average Accuracy and Average Precision as our evaluation metrics where

Accuracy = (Correctly Predicted Test Instances / Total Test Instances) * 100

Precision = (True Positive / (True Positive + False Positive)) * 100

Results

Sl No	Classifier	Best Parameters Used	Accuracy(in %)	Precision(in %)
1	Decision Tree	max_depth=5	74.79	61.45
2	Perceptron	n_iter=10	72.57	70.96
3	Neural Net	hidden_layer_sizes=(4 , 2) learning_rate_init=.01	73.19	69.99
4	Deep Learning	hidden_layer_sizes=(15 , 6) learning_rate_init=.01	76.82	72.43
5	SVM	kernel=linear	76.69	72.23
6	Naive Bayes	fit_prior=True	65.10	54.48
7	Logistic Regression	Tolerance=0.00001	76.05	71.00
8	KNN	n_neighbors=5 weights=distance	74.62	66.69
9	Bagging	n_estimators=100	76.44	69.37
10	Random Forests	n_estimators=100 max_features=2	77.21	72.90
11	Adaboost	n_estimators=60	75.54	66.64

12	Gradient Boosting	n_estimator=50	76.56	70.83

Analysis

From the above results, it can be noted that Random Forests learner has a slight advantage over the others in terms of accuracy and Precision. Random forest algorithm is an improvement over the bagged decision tree approach. Although Bagging minimizes the variance, the base decision tree models can have a lot of structural similarities and in turn have high correlation in their predictions. Combining predictions from multiple models in ensembles works better if the predictions from the sub-models are uncorrelated or at best weakly correlated. Random forest does exactly this and hence it performed better than others.

Naive Bayes classifier fared worst with low accuracy and low precision. It makes sense since Naive Bayes is a probabilistic approach that typically assumes the independence of all the features of a dataset.

There are a few attributes which have more influence on attributes than others. In case of random forest, altering the number of decision trees and the number of chosen features greatly influenced the accuracy and precision. It could be observed that altering the attributes only improved the accuracy upto a point beyond which the bias-variance tradeoff became more and more evident. Hence, it can be concluded that finding the right balance between bias and variance is the key to finding an optimum model for any learning algorithm.

Even though accuracy is the best evaluation metric of all, we found that precision taken along with accuracy was a better option. This is because precision gives the number of times that the model correctly classifies the data. When combined with accuracy, we get the number of times the values are identified as 1 or TRUE, along with the percentage of correct classification.