CSCI 6515

Assignment-1

Instructor: Dr. Stan Matwin
TA: Dr. Amilcar Soares

Date of Submission: Oct 5, 2017

Submitted by:

Bhargav Dalal

B00785773

Faculty of Computer Science

Dalhousie University

Halifax, Nova Scotia

Today, we will be working on Animals classification dataset which was obtained from starkey project.

The target variable is ‘class’ field which is a categorical field. Following are some statistics about the data:

- Number of instances: 5135
- Number of features: 25
- Number of classes: 3

The packages used for the source code are:

- Numpy: multidimensional arrays, vector and matrix operations
- Pandas: data manipulation and analysis
- Scikit-learn: machine learning li brary for classification, regression, clustering, feature selection and much more

Firstly model was selected, then training and done predictions of it. We have use following models:

- Linear Regression
- Naïve Baiyes – Bernoulli Biaiyes
- Decision Tree Classification
- RandomForest Classifier

Then, to measure the performance we have calculate the accuracy and standard deviation later on for each model.

Task A) Split the data randomly into a training set and a testing set (e.g. 70%-30%). Train all classifiers (Logistic Regression, Naïve Bayes, Decision Tree and Random Forests) using the default parameters using the train data. Report the confusion matrix and accuracy for both train and test data. Compare the train and test accuracy. Is there a big difference between train and test accuracy? Why?

Solution:

After, splitting the data as per requirements, training and predicting it on both test data and training data we have generated following confusion matrix and accuracy:

CLASS	LOGISTIC REGRESSION	BERNOULLI BAIYES	DECISION TREE CLASSIFICATION	RANDOMFOREST CLASSIFICATION
CATTLE	[[1198 25] [280 38]]	[[1216 7] [315 3]]	[[1058 165] [164 154]]	[[1195 28] [220 98]]
DEER	[[1068 49] [368 56]]	[[1105 12] [418 6]]	[[854 263] [220 204]]	[[1014 103] [233 191]]
ELK	[[485 257] [255 544]]	[[607 135] [586 213]]	[[461 281] [228 571]]	[[1014 103] [233 191]]

Table A.1: Confusion Matrix of Test Data

CLASS	LOGISTIC REGRESSION	BERNOULLI BAIYES	DECISION TREE CLASSIFICATION	RANDOMFOREST CLASSIFICATION
CATTLE	[[2811 57] [635 91]]	[[2854 14] [716 10]]	[[2868 0] [0 726]]	[[2866 2] [56 670]]
DEER	[[2460 135] [851 148]]	[[2578 17] [982 17]]	[[2595 0] [0 999]]	[[2592 3] [57 942]]
ELK	[[1162 563] [557 1312]]	[[1471 254] [1375 494]]	[[1725 0] [0 1869]]	[[1716 9] [41 1828]]

Table A.2: Confusion Matrix of Train Data

CLASS	ACCURACY TYPE	LOGISTIC REGRESSION (%)	BERNOULLI BAIYES (%)	DECISION TREE CLASSIFICATION (%)	RANDOMFOREST CLASSIFICATION (%)
CATTLE	**TEST**	80.21	79.10	78.65	83.91
	TRAIN	80.75	79.69	100	98.39
DEER	**TEST**	72.93	72.10	68.66	78.20
	TRAIN	72.56	72.20	100	98.33
ELK	**TEST**	66.77	53.21	66.97	78.19
	TRAIN	68.84	54.67	100	98.61

Table A.3: Accuracy of both Test & Train Data

One notable thing from "Table A.3: Accuracy of both Test & Train Data" is accuracy of train data for Decision Tree Classification is 100% and RandomForest Classification is near to 100%. To understand it, let us define 2 terms: Confusion Matrix and Accuracy:

Confusion Matrix: It is a table which often used to describe the performance of a classification model on a set of test data for which the true values are already known.

It's structure can be explain as

Total Observation (instances)	Predicted No	Predicted Yes
Actual(Real) No	**True Negative(TN)**	**False Positive(FP)**
Actual(Real) Yes	**False Negative(FN)**	**True Positive(TP)**
Total		

Accuracy: It is the ratio of sum of TN and TP and total no. of instances.

Thus, in our example if we consider deer class for decision tree model, the 2 confusion matrix along with accuracy are:

Deer	**DECISION TREE CLASSIFICATION**	**ACCURACY (%)**
TEST	[[**TN**=854 **FP**=263] [**FN**=220 **TP**=204]]	68.66
TRAIN	[[**TN**=2595 **FP**=0] [**FN**=0 **TP**=999]]	100

As we can see, **for training confusion matrix, False Positive and False Negative value is 0 while on other hand there are values for both of them for test data confusion matrix and hence due to that, there is difference of accuracy on test data and train data. This is also known as over-fitting. Similar situation of RandomForest Classification for all classes.**

NOTE:

I have chosen Bernoulli's Baiyes classification as it requires input in binary value and as we have divided target class in binary variable (i.e 0,1 for each class), hence selected it.

Task B) Using 10-fold cross-validation, train and evaluate all classifiers. Compare the accuracy of the methods in terms of mean (μ) and standard deviation (σ) of accuracy in 10 folds. Eventually use a statistical significance test (e.g. student's ttest) and determine whether the methods are significantly different or not. Use α = 0.05 as the significance threshold. For applying the significance test, select the classifier with the best average performance, and compare it to all the remaining classifiers.

Solution:

There are numerous methods of cross-validation. One of the method which we are implementing here is 10-fold validation. In 10-fold cross validation method, our data will split in 10 different folds (set or frame) and then it will take 9 folds for training and 1 fold for testing the model. Thus, the model will be trained 10 times and hence the accuracy for it will be high as compared to splitting the data with just test and train data set.

Following is the table containing accuracy and standard deviation of each model which was trained with 10-fold cross validation method.

Aggregate Function	Logistic Regression	Bernoulli Baiyes	Decision Tree Classification	RandomForest Classification
ACCURACY	73.06%	67.82%	70.33%	76.47%
STANDARD DEVIATION	0.07	0.119	0.058	0.07

Table B.1: Accuracy and Standard Deviation for all classifiers

From the table, we can clearly see that best classifier would be the RandomForest Classification.

Now, we will use Student t-test method to prove the above result.

Q. What is Paired Student's t-test?

Student t-test is a test which compares two means of the example and tells how different they actually are from each other. Similarly a paired t-test is a test where we choose 2 different measurements (Mean & Standard Deviation in our scenario) on 2 different samples (2 classifiers in our case) to tell us if they are statistically significant from each other or not. For example, you might test two different groups of customer service associates on a business-related test or testing students from two universities on their English skills.

When we will apply the t-test, we will get 2 values, which is called as **t-values** and **p-values.**

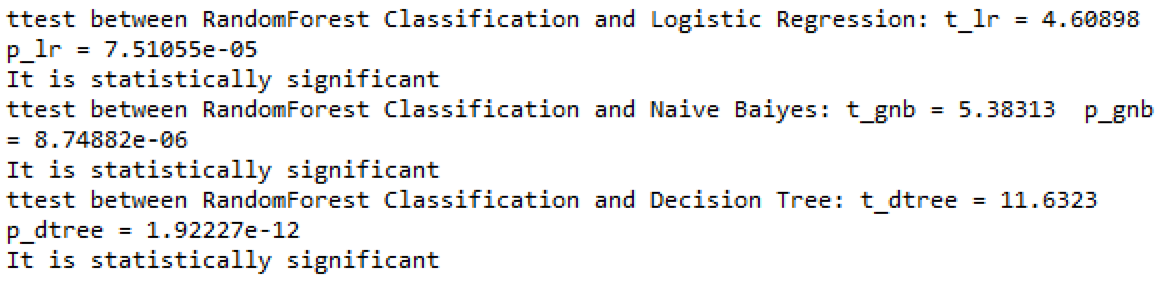
t-values: It is ratio between difference between 2 groups (in our example 2 classification model) and the squared difference between 2 groups. So if t-value is larger, more the difference from each other and if smaller, then it will be similar.

p-value: A p-value is the probability that the results from your sample data occurred by chance. Thus the lower the p-value the less chance of occurring the data by chace.

After applying the test, we obtain the results as:

t-test between Random Forest v/s Logistic Regression, Bernoulli Baiyes Classification and Decision Tree:

A paired-samples t-test was conducted to compare the accuracy of prediction for the classifiers - Random forest classification (RFC) with Logistic Regression Classification (LRC), Bernouli Baiyes Classification (GNB) and Decision Tree (dt), one at a time. There was a significant differences in the scores of RFC (i.e. Mean and Standard deviation) with LRC, GNB and dt. The corresponding t-values and p-values for test with each model is shown in below figure:



```
ttest between RandomForest Classification and Logistic Regression: t_lr = 4.60898
p_lr = 7.51055e-05
It is statistically significant
ttest between RandomForest Classification and Naive Baiyes: t_gnb = 5.38313  p_gnb
= 8.74882e-06
It is statistically significant
ttest between RandomForest Classification and Decision Tree: t_dtree = 11.6323
p_dtree = 1.92227e-12
It is statistically significant
```

Figure B.1: T-values and p-values for parity t-test with each model

Now, as p-value in each comparison is less than given threshold value ($\alpha = 0.05$) we can conclude that **Random Forest classifier is predicting more accurately then the rest three of the classifiers.**

Task C) Train a Random Forest using a 10-fold cross-validation with the 10, 20, 50 and 100 trees (e.g. number of estimators in the scikit package) and report the mean accuracies. Choose one of the solutions, justify why you chose it, and compare it again with your results for Logistic Regression, Naïve Bayes, and Decision Tree using the student's t-test.

Solution:

The following are the accuracies for the classifier with 10, 20, 50 and 100 no. of trees

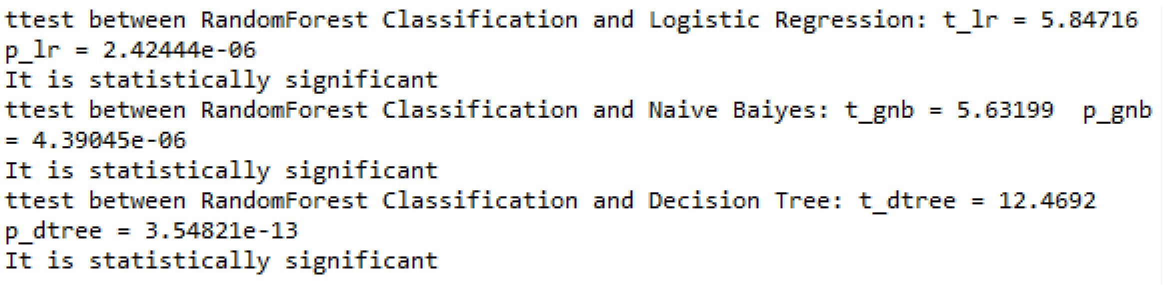
No. of Trees	Accuracy (%)
10	76.88
20	77.30
50	78.03
100	78.29

Here, random forest classification with 50 trees seems to be appropriate selection for the model as with 20 trees it will not be most accurate as per the definition that number of trees would likely be more than the number of features for better performance of the model. As we have 25 features, so it would be more accurate to consider results of classification with 50 trees rather than with 20 trees. We are not choosing a classification with 100 trees as it is clearly indicating that there is no amount of significant increase in accuracy of prediction as compare to classification with 50 trees.

Applying Student t-test for Random Forest Classification with 50 trees v/s all other 3 classifier.

t-test between Random Forest v/s Logistic Regression, Bernoulli Baiyes Classification and Decision Tree:

A paired-samples t-test was conducted to compare the accuracy of prediction for the classifiers - Random forest classification-50trees (RFC) with Logistic Regression Classification (LRC), Bernoulli Baiyes Classification (GNB) and Decision Tree (dt), one at a time. There was a significant differences in the scores of RFC (i.e. Mean and Standard deviation) with LRC, GNB and dt. The corresponding t-values and p-values for test with each model is shown in below figure:



```
ttest between RandomForest Classification and Logistic Regression: t_lr = 5.84716
p_lr = 2.42444e-06
It is statistically significant
ttest between RandomForest Classification and Naive Baiyes: t_gnb = 5.63199  p_gnb
= 4.39045e-06
It is statistically significant
ttest between RandomForest Classification and Decision Tree: t_dtree = 12.4692
p_dtree = 3.54821e-13
It is statistically significant
```

Figure C.1: T-values and p-values for parity t-test with each model

Thus, from above test it can be concluded that Random Forest classifier with 50 trees is predicting more accurately then the rest three of the classifiers.

REFERENCES:

http://scikit-learn.org/

https://www.scipy.org/

https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.ttest_rel.html

http://www.statisticshowto.com/probability-and-statistics/t-test/

https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm