CSC6515 – Machine Learning for Big Data

Assignment-1

Professor : Dr. Stan Matwin

TA (Professor) : Dr. Amilcar Soares Jr.

Date of Submission: Oct 5, 2017

Submitted by:

Hardik Galiawala

B00777450

Faculty of Computer Science

Dalhousie University

Halifax, Nova Scotia

A) Split the data randomly 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?

Ans:-

The accuracy of randomly split data into training and test set in a 7:3 ratio per class of animal (since we have to consider Logistic regression, which only considers 2 outcomes) i.e., DEER, ELK and CATTLE

1. Accuracy :-

- Accuracy of Class – **DEER** :-

	Logistic Regression	Naïve Bayes	Decision Tree	Random Forest
Training Data	72.50 %	35.47 %	100.00 %	98.41 %
Test Data	73.30 %	35.89 %	69.75 %	77.74 %

Figure A.1.1 – Accuracy [Deer]

- Accuracy of Class – **ELK** :-

	Logistic Regression	Naïve Bayes	Decision Tree	Random Forest
Training Data	68.72 %	54.75 %	100.00 %	98.91 %
Test Data	67.72 %	52.36 %	67.03 %	71.96 %

Figure A.1.2 – Accuracy [Elk]

- Accuracy of Class – **CATTLE** :-

	Logistic Regression	Naïve Bayes	Decision Tree	Random Forest
Training Data	81.16 %	79.10 %	100.00 %	98.66 %
Test Data	79.68 %	78.32 %	77.67 %	82.02 %

Figure A.1.3 – Accuracy [Cattle]

In order to summarize above data, there is a huge difference between training set and test set if we observe Decision Tree and Random Forest classifiers. It is probably due to over-fitting of the models. Pruning is one of the ways to overcome this problem.

2. Confusion Matrix :-

- Confusion Matrix of Class – **DEER** :-

	Logistic Regression	Naïve Bayes	Decision Tree	Random Forest
Training Data	[[2468 116] [872 138]]	[[332 2252] [67 943]]	[[2584 0] [0 1010]]	[[2578 6] [51 959]]
Test Data	[[1087 41] [370 43]]	[[165 963] [25 388]]	[[889 239] [227 186]]	[[1042 86] [257 156]]

Figure A.2.1 – Confusion Matrix [Deer]

- Confusion Matrix of Class – **ELK** :-

	Logistic Regression	Naïve Bayes	Decision Tree	Random Forest
Training Data	[[1106 594] [530 1364]]	[[142 1558] [68 1826]]	[[1700 0] [0 1894]]	[[1688 12] [27 1867]]
Test Data	[[504 263] [242 532]]	[[66 701] [33 741]]	[[510 257] [251 523]]	[[565 202] [230 544]]

Figure A.2.2 – Confusion Matrix [Elk]

- Confusion Matrix of Class – **CATTLE** :-

	Logistic Regression	Naïve Bayes	Decision Tree	Random Forest
Training Data	[[2826 55] [622 91]]	[[2752 129] [622 91]]	[[2881 0] [0 713]]	[[2880 1] [47 666]]
Test Data	[[1188 22] [291 40]]	[[1161 49] [285 46]]	[[1047 163] [181 150]]	[[1167 43] [234 97]]

Figure A.2.3 – Confusion Matrix [Cattle]

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.

Ans :-

Cross Validation :-

1. Accuracy (Mean and Standard Deviation) per Class :

	Accuracy	Logistic Regression	Naïve Bayes	Decision Tree	Random Forest
Class – **DEER**	Mean	71.57 %	35.01 %	67.03 %	74.90 %
	Standard Deviation	0.0341	0.0351	0.0362	0.0378
Class – **ELK**	Mean	67.29 %	53.87 %	66.41 %	72.64 %
	Standard Deviation	0.0681	0.0263	0.0277	0.039
Class – **CATTLE**	Mean	80.33 %	78.85 %	77.21 %	82.90 %
	Standard Deviation	0.0142	0.119	0.0289	0.0211

2. Accuracy (Aggregated) :

Accuracy	**Logistic Regression**	**Naïve Bayes**	**Decision Tree**	**Random Forest**
Mean	73.06 %	55.91 %	70.21 %	76.81 %
Standard Deviation	0.0703	0.1814	0.0585	0.0555

3. **Statistical significance test [Student's t-test]** :-

As we can see from the above data, overall (aggregated) accuracy of Random Forest seems to be highest. Hence, we will take it as our reference sample while performing student's t-test.

v/s	P-Values		
	Logistic Regression	**Naïve Bayes**	**Decision Tree**
Random Forest	0.0475694072673	1.06536e-06	0.000133142685251

Figure B.3 – P-values [Student's t-test]

From figure-B.3, we can see that all the values are less than given significance threshold (0.05). The P-values are low and thus they are good. It also indicates that data is not occurring by chance for any of the classifiers.

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.

Ans :-

Training a random Forest using a 10-fold cross validation with 10, 20, 50 and 100 trees (estimators)

1. **Mean accuracies of Random Forests (n- Estimators) :-**

Accuracy	**10 - Estimators**	**20 - Estimators**	**50 - Estimators**	**100 - Estimators**
Mean	76.87 %	77.30 %	78.22 %	78.25 %

Figure C.1 – Accuracy n-Estimators

From figure C.1, it is observed that we get highest accuracy when we consider 100 estimators. But if we consider time and resources consumed during the computation along with the amount of difference in the mean accuracies with n-estimators, it is concluded that 20-Estimators is best case scenario for us.

2. **Statistical significance test [Student's t-test] :-**

v/s	P-Values		
	Logistic Regression	**Naïve Bayes**	**Decision Tree**
Random Forest(20-Estimators)	9.917334743e-07	2.33438052505e-08	4.49148564373e-14

Figure C.2 – P- Values [Student's t-test]

From figure C.2, we can easily conclude that all the data is occurring normally and not by chance since, all p-values are less than our threshold. We have assumed threshold to be 0.05 as we are not given any exact value to be taken into consideration for this question.

References :-

1. http://dataaspirant.com/
2. https://machinelearningmastery.com/
3. http://scikit-learn.org
4. https://stats.stackexchange.com/
5. http://www.statisticshowto.com/probability-and-statistics/t-test/