CS 6375- ASSIGNMENT 4

Please read the instructions below before starting the assignment.

- This assignment consists of implementing 5 new classifiers that you have learned on the same dataset that you used for assignment 2.

- It is very important to include a report summarizing your experiments and your results.

- You need to demonstrate that you made an effort to find the best set of parameters for each classifier. For this, you should maintain and submit a log file indicating the parameter selection and the output in terms of accuracy.

- In the code folder, please include a README file indicating how to compile and run your code. Also, mention clearly which language and packages you have used.

- You should use a cover sheet, which can be downloaded at: http://www.utdallas.edu/~axn112530/cs6375/CS6375_CoverPage.docx

- You are allowed to work in pairs i.e. a group of two students is allowed. Please write the names of the group members on the cover page.

- The deadline for this assignment is Friday November 11 at 11:59 PM. No extensions are allowed.

- You have a total of 4 free late days for the entire semester. You can use at most 2 days for any one assignment. After that, there will be a penalty of 10% for each late day. The submission for this assignment will be closed 2 days after the due date.

- Please ask all questions through Piazza, and not through email.

ASSIGNMENT 4

In assignment 2, you compared the performance of five classifiers on your chosen different dataset. Since then, you have learned several more classifiers. Below is the list of the new classifiers learned:

- Logistic Regression
- k-Nearest Neighbors
- Bagging
- Random Forests
- AdaBoost

In this assignment, you will test the performance of these classifiers on the same dataset that you used in assignment 2. The idea of this assignment is to compare classifier performance and accuracy.

You are free to use any language and any packages. If you decide to do this assignment in R, the list of suggested packages is shown below.

Classifier	Package	Training Function
Logistic Regression	stats	glm with family = binomial
k-Nearest Neighbors	class	knn
Bagging	adabag	bagging
Random Forest	randomForest	randomForest
Boosting	adabag	boosting

Please make an effort to learn these packages or any other software tool that you use. You should definitely vary the parameters to suit your dataset and its needs.

Finding best classifier parameters

Each classifier requires a number of parameters. For example, a bagging model requires you to specify number of models (generally trees) to use, and the parameters of individual trees, such as maximum depth. Similarly, a Random Forest model requires you to specify the number of variables randomly sampled as candidates at each split. In order to get the best performance from each classifier, **you have to train the classifier using different set of parameters and then select the best set of parameters**. This is a critical part of the assignment that is overlooked by many students. You need to perform n-fold cross validation, with a suitable choice of n (>= 10).

In order for us to see that you really did try a number of different parameters, **you have to maintain a log of your experiments.**
A tabular form as shown below would be great:

Experiment #	Classifier	Cross-Validation fold	Parameter1	Parameter2	…	Average Accuracy
1	LogisticRegression	10	parameter 1 = x1	parameter 2 = x2	…	
2	k-NN	10	parameter 1 = x1	parameter 1 = x1		
…	…		..	..	.	.

It is up to you how many times you run the experiments or how you many permutations of the parameters you try. However, you need to convince us that you made a sincere effort for finding the best set of parameters.

Training and Testing Methodology for both options

To test the performance of the classifiers, you will use n-fold cross validation. It is up to you to choose a suitable value of n, but it should be at least 10. You should then report the average value of accuracy of the n-fold cross validation. Please include pseudocode showing your training and testing methodology. It should indicate whether you manually created n-folds or you used the package for this.

Besides accuracy, you should evaluate your models on at least one more parameter discussed in class. It could be precision, recall, ROC or area under ROC curve. An excellent source for model evaluation is the pROC package in R. Other languages also have evaluation methodologies. Please make an effort to learn these and implement them in your code.

Reporting your results:

You should report your results as follows:

Number of instances in dataset:
Number of attributes in dataset:
How many fold cross validation performed:

Classifier	Technique	Accuracy	Another evaluation metric
Logistic Regression	Example: ___-fold cross validation average results		
k-NN			
Bagging	..	..	
Random Forest	…	..	..
Boosting	…	…	..

Analysis:

In a few paragraphs, explain your results. Analyze which methods performed best and why do you think they performed best. Analyze which were weak methods and why. Do some attributes influence the output more than others? Is accuracy a good evaluation metric or did you find another metric that is better than accuracy. Any other details you wish to include.

What to submit:

- Project report including pseudocode, validation techniques used, results table, and analysis.
- Log file showing how you found the best set of parameters for each model.
- Code file.
- README file indicating which languages and packages you used and how to compile your code.