Project 1: Classification

1. Iris data set (http://archive.ics.uci.edu/ml/datasets/Iris)

The data set contains 3 classes of 150 records and 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other. The Predicted attribute is class of iris plant. The raw data files can be downloaded from http://archive.ics.uci.edu/ml/machine-learning-databases/iris/

Attribute Information:
1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class: -- Iris Setosa , -- Iris Versicolour , -- Iris Virginica

2. Task Description

1.1 Selecting 70% data records as the training data set and next predicting the remaining 30% data records. The training data set is randomly selected from the Iris data set, and you need to set the random seed number equal to 12345.

1.2 Using C50 decision tree to build the classification model based on the training data set. Next, to improve the accuracy of decision trees, you need to add adaptive boosting by adding an additional parameters trails = 10.

1.3 Using the knn function in the "class" package to train the data and next to predict the test data set. The number of k is set to be 11. Next, by varying a different number of k including 1, 3, 5, 11, 17 and 21, you need to repeat the classification process. After that, by normalizing the first 4 attributes (sepal length, sepal width, petal length and petal width), you need to compare the knn classification model with the one without normalization. Finally, with the help of z-score standardization (using the scale() function), you need to compare the knn classification model with the one without standardization.

1.4 For the steps 2.2-2.3, you need to compute the accuracy of the classification model by using the CrossTable() function to compute the percentages of false negatives, false positives and Percent classified Incorrectly.

3. What to submit

3.1 a pdf document file to list the accuracy of all classification models built by different approaches

and varying parameters (such as the boosting, different k numbers, normalization, and standardization). In addition, you need to compare these results and discuss the reasons that lead to such results. The document should contain the name, student Number, and email address of each group member.

3.2 the r source code, the output file and saved R workspace file.

4. When and where to submit

4.1 Submission Deadline: 2014-10-25. If n (>=1) days are late to submit the file, the final score of this project is reduced by n*15 %.

4.2 All files are zipped with the naming conversion like proj1-xxx-xxx-xxx.zip, where each xxx indicates the student num. of a group member.