Project 2: Clustering

1. SMS Spam Collection Data Set

1) SMS_Spam_Data-1:
The collection (http://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection) is composed by just one text file, where each line has the correct class followed by the raw message. We offer some examples bellow:

ham What you doing?how are you?
ham Ok lar... Joking wif u oni...
ham dun say so early hor... U c already then say...
ham MY NO. IN LUTON 0125698789 RING ME IF UR AROUND! H*
ham Siva is in hostel aha:-.
ham Cos i was out shopping wif darren jus now n i called him 2 ask wat present he wan lor. Then he started guessing who i was wif n he finally guessed darren lor.
spam FreeMsg: Txt: CALL to No: 86888 & claim your reward of 3 hours talk time to use from your phone now! ubscribe6GBP/ mnth inc 3hrs 16 stop?txtStop
spam Sunshine Quiz! Win a super Sony DVD recorder if you canname the capital of Australia? Text MQUIZ to 82277. B
spam URGENT! Your Mobile No 07808726822 was awarded a L2,000 Bonus Caller Prize on 02/09/03! This is our 2nd attempt to contact YOU! Call 0871-872-9758 BOX95QU

Note: the messages are not chronologically sorted.

The raw data file can be downloaded from http://archive.ics.uci.edu/ml/machine-learning-databases/00228/.

2) SMS_Spam_Data-2 (NUS SMS Corpus): The file can be downloaded from http://wing.comp.nus.edu.sg:8080/SMSCorpus/data/corpus/smsCorpus_en_sql_2014.09.06_all.zip
The NUS data set is to use SQL statements to insert the spam text to a RDBMS. We will use the data in the `content` attribute and no others are useless for this project.

2. Task Description

1.1 Use k-means and k-medoids to cluster the ***SMS_Spam_Data-1*** by setting the number of k equal to 2, 4, 8 and 16. Compare and discuss the clustering results between the k-means and

k-medoids approaches. For the k = 2, you need to compare the clustered SMS messages with the classified messages (ham and spam), and discuss the relations of the clustered SMS messages and classified messages.

1.2 We need to you use the ***SMS_Spam_Data-1*** to train the classification model (we use the naïve bayes) and classify the ***SMS_Spam_Data-2***.

1.3 Use k-means and k-medoids to cluster the ***SMS_Spam_Data-2*** by setting the number of k equal to 2, 4, 8 and 16. Compare and discuss the clustering results between the k-means and k-medoids approaches. For the k = 2, you need to compare the clustered SMS messages with the classified messages (ham and spam), and discuss the relations of the clustered SMS messages and classified messages.

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 cøntain 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: 2014-12-14. 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 proj2-xxx-xxx-xxx.zip, where each xxx indicates the student num. of a group member.