CS481: Data Science
Assignment 03

Due Date: Monday, 22th April 2018 **Total Marks: 100**

- Use Table 1, to show the training procedure using AdaBoost. Continue the iterations until there are no mistakes left. The weak learner produces hypotheses of the form: $x < v$, or $x > v$. The threshold v is determined to minimize the probability of error over the entire data (No sampling). Use AdaBoost.pdf and AdaBoostExample.pdf (attachment) for more understanding. [Hint If $(x < v)$ is true, than y = 1 else y = -1, if $(x > v)$ is true than y=1 else y = -1] [**20 Points**]

- Consider a Database D (Table 2) consists of 9 transactions. Find strong association rules using Apriori Algorithm. Suppose min. support count = 2 and min. confidence = 70%. At the end show association rules with frequent itemset with maximum item only. [**20 Points**].

- Using Internet Advertisement dataset and Python, run Bagging with 10 Fold CV; use Decision Tree as base classifier. Compare and analyse the performance using accuracy by varying bagPercentSize (10, 20, 30, ... , 100). Plot graphs to illustrate (x-axis; bagpercentsize; y-axis: accuracy). Also, discuss the importance of bagPercentSize in Bagging [**20 Points**]

- Repeat above exercise using AdaBoost and Random Forest. Plot bar graph for Bagging (best bagpercentsize), AdaBoost and Random Forest. Discuss the importance of key parameters in both AdaBoost and Random Forest. You can prove it using simulation or by using references from key papers [**20 Points**]

- Read the following paper and answer the following [**20 Points**]

  ”Inverse random undersampling for class imbalance problem and its application to multi-label classification” Author(s): M A Tahir, J Kittler and F. Yan Page(s): 3738–3750, Pattern Recognition, 2012 (a) What is class imbalance problem. What are the main research problems in class imbalance problems [5 Points] (b) List five most popular techniques for class imbalance problem [5 Points] (c) What is the main novelty in this paper [10 Points]

Index	0	1	2	3	4	5	6	7	8	9
x value	0	1	2	3	4	5	6	7	8	9
y value	1	1	-1	-1	-1	1	1	1	-1	-1

Table 1: Training Data for AdaBoost.

TID	Items
1	I_1, I_2, I_7
2	I_1, I_2, I_3
3	I_2, I_4, I_5
4	I_1, I_2, I_5
5	I_5, I_6, I_7
6	I_I, I_2, I_7, I_8
7	I_I, I_2, I_3, I_8
8	I_1, I_3, I_5, I_6, I_7, I_8
9	I_2, I_3, I_4, I_6, I_7, I_8

Table 2: Database D for transactions. I_1 = Bread, I_2 = Milk, I_3 = Egg, I_4 = Toy, I_5 = Coke, I_6 = Chicken, I_7 = Cheese, I_8 = Juice.