Data Mining – Assignment 1 and 2

- Write this assignment in your class note book. Complete by **24/02/2019.** I will start checking randomly from 25/02/2019. If found to be not written marks of 3 assignments will be deducted.
- Don't copy. Solve by yourself and write your own answers. And remember practice reduces your test error not coping.

1. Discuss whether or not each of the following activities is a data mining task.

   (a) Dividing the customers of a company according to their gender.

   (b) Dividing the customers of a company according to their profitability.

   (c) Computing the total sales of a company.

   (d) Sorting a student database based on student identification numbers.

   (e) Predicting the outcomes of tossing a (fair) pair of dice.

   (f) Predicting the future stock price of a company using historical records.

   (g) Monitoring the heart rate of a patient for abnormalities.

   (h) Monitoring seismic waves for earthquake activities.

   (i) Extracting the frequencies of a sound wave

2. Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving specific examples of how techniques, such as clustering, classification, association rule mining, and anomaly detection can be applied.

3. Consider the market basket transactions shown in the following Table

Transaction ID	Items Bought
1	{Milk, Beer, Diapers}
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{Milk, Diapers, Bread, Butter}
7	{Bread, Butter, Diapers}
8	{Beer, Diapers}
9	{Milk, Diapers, Bread, Butter}
10	{Beer, Cookies}

   Find the most interesting association rule(s).

4. Recently, the annual number of driver deaths per 100,000 for the selected age groups was as follows:

Age	Number of Driver Deaths per 100,000
17.5	38
22	36
29.5	24
44.5	20
64.5	18
80	28

a) Using "ages" as the independent variable and "Number of driver deaths per 100,000" as the dependent variable, make a scatter plot of the data.
b) Calculate the least squares (best–fit) line. Put the equation in the form of: $\hat{y} = a + bx$
c) Find the correlation coefficient. Is it significant?
d) Predict the number of deaths for ages 40 and 60.
e) Based on the given data, is there a linear relationship between age of a driver and driver fatality rate?
f) What is the slope of the least squares (best-fit) line? Interpret the slope.

5. The data below are measurements of height h (in m) and diameter d (in cm) of 18 Corsican Pines

Tree	1	2	3	4	5	6	7	8	9
Diameter (cm)	32	31	30	29	29	28	25	23	20
Height (m)	22.7	22.7	22.6	22.6	21.9	21.9	21.8	21.0	20.4
Tree	10	11	12	13	14	15	16	17	18
Diameter (cm)	18	17	17	16	16	15	13	11	11
Height (m)	18.6	19.2	18.9	18.5	18.1	17.7	17.2	16.5	15.5

It is of interest to describe the tree height as a function of the diameter. Explore the relationship between the variables using scatter plot. Which model better fits the data, linear or polynomial regression. If polynomial, what polynomial order seems necessary to describe the data? (You don't need to do regression, just analyze using the scatter plot)

6. Consider the following set of training example.

Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−

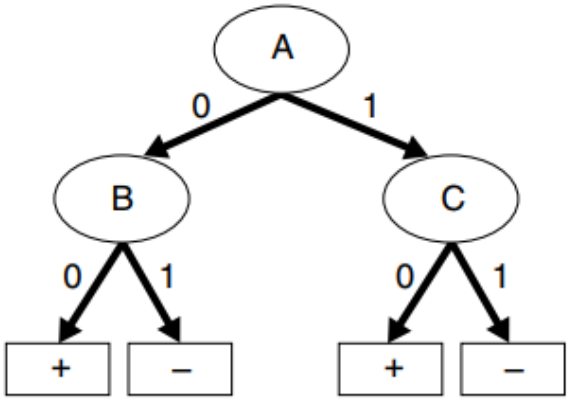
a) What is the entropy of this collection of training examples with respect to the positive class?
b) What are the information gains of a1 and a2 relative to these training examples?
c) For a3, which is a continuous attribute, compute the information gain for every possible split.

7. Consider the following set of training example.

X	Y	Z	No. of Class C1 Examples	No. of Class C2 Examples
0	0	0	5	40
0	0	1	0	15
0	1	0	10	5
0	1	1	45	0
1	0	0	10	5
1	0	1	25	0
1	1	0	5	20
1	1	1	0	15

a) Compute a two-level decision tree using the greedy approach described in this chapter. Use the Gini Index as the criterion for splitting. What is the overall error rate of the induced tree?
b) Repeat part (a) using X as the first splitting attribute and then choose the best remaining attribute for splitting at each of the two successor nodes. What is the error rate of the induced tree?
c) Compare the results of parts (a) and (b). Comment on the suitability of the greedy heuristic used for splitting attribute selection.
d) Evaluate the first model's performance with Precision, Recall and F1 measure.

8. Apply Post Pruning: Cross Validation method in the following example.



Training:

Instance	A	B	C	Class
1	0	0	0	+
2	0	0	1	+
3	0	1	0	+
4	0	1	1	–
5	1	0	0	+
6	1	0	0	+
7	1	1	0	–
8	1	0	1	+
9	1	1	0	–
10	1	1	0	–

Validation:

Instance	A	B	C	Class
11	0	0	0	+
12	0	1	1	+
13	1	1	0	+
14	1	0	1	–
15	1	0	0	+