HOMEWORK 1

Data Mining

Deadline ： 2016/04/01

Name : _____ Student ID : _____

1. Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.
Example: Age in years. **Answer**: Discrete, quantitative, ratio

(a) Angles as measured in degrees between 0◦ and 360◦.
(b) Number of patients in a hospital.
(c) ISBN numbers for books.
(d) Time in terms of AM or PM.

2. Given an array, divide the numbers into 3 groups in various ways :
(0, 1, 1, 2, 3, 5, 7, 9, 10, 11, 13, 19, 20, 21, 26, 27, 29, 30)
(a) Equal interval width.
(b) Equal frequency.

Q : Do you have any other ideas to divide it?

3. You are given a set of m objects that is divided into K groups, where the ith group is of size mi. If the goal is to obtain a sample of size n < m, what is the difference between the following two sampling schemes? (Assume sampling with replacement.)

(a) We randomly select n ∗ mi/m elements from each group.
(b) We randomly select n elements from the data set, without regard for the group to which an object belongs.

4. Explore the cosine and correlation measures.
(a) What is the range of values that are possible for the cosine measure?
(b) If two objects have a cosine measure of 1, are they identical? Explain.

5. Consider a game, the winning probabilities of A, B, C, D are listed below. Calculate the entropy of "Who win the game" respectively and answer the question.

(a)

A	B	C	D
100%	0%	0%	0%

(b)

A	B	C	D
50%	25%	25%	0%

(c)

A	B	C	D
25%	25%	25%	25%

Q : We can see, (a/b/c) is the maximum, (a/b/c) is the minimum. Try to analyse the reason.

6. For the following vectors, x and y, calculate the indicated similarity or distance measures. (**Computing process is required !**)

Q1 : $x = (1, 1, 0, 0); y = (0, 0, 1, 1)$: Correlation, Jaccard.

Q2 : $x = (1, 2, 4, 8); y = (2, 4, 8, 16)$: Cosine, Euclidean.

Q3 : $x = (10, 20, 40, 80); y = (1.1, 1.2, 1.4, 1.8)$: Cosine, Correlation.

Q4 : According to the result of Q3, tell the differences between cosine and correlation.

7. Given a similarity measure with values in the interval [0,1] describe two ways to transform this similarity value into a dissimilarity value in the interval $[0, \infty]$.

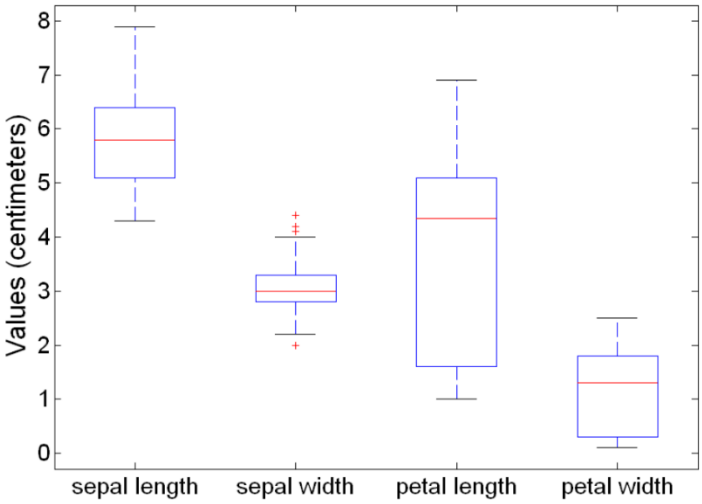
8. Construct a data cube from Table 1. Is this a dense or sparse data cube? If it is sparse, identify the cells that are empty.

Table 1.

Product ID	Location ID	Number Sold
1	1	21
1	2	5
2	3	3
2	1	8

9. List some information obtained from Figure 1. (At Least 3)

Figure 1.



10. Talk about your comprehension of *Data Mining*.

Any questions, send e-mail to ***ypub@msn.com***