YILDIZ TECHNICAL UNIVERSITY
COMPUTER ENGINEERING DEPARTMENT
0114850 DATA MINING MIDTERM EXAM

Instructor: Assistant Prof. Songül ALBAYRAK 20th April 07

Important Note: You have exactly 90 minutes for the exam and any type of cheating will be dealt with seriously.

Name and Last Name: **Student ID:**

	Question 1	Question 2	Question 3	Question 4	Question 5	Question 6
Points	26P	10P	12P	18P	16P	18P

QUESTIONS

1.) Table 1 shows attributes of two classes of planes: civil and military. With using each attribute,

Table 1

Plane Weight(*10^3kg)	Plane Velocity (km/h)	Plane Type
26	1460	War
33	1450	War
36	1550	War
37	1350	War
72	1564	War
505	595	Civil
477	825	Civil
590	600	Civil

a.) Find mean, median, variance and standard deviation of datasets.
b.) Realize the calculations and draw boxplot analysis in regular form.
c.) Use min-max normalization to transform the value 400 (*10^3kg) for plane weight onto range [0.0 , 1.0]
d.) Use z-score normalization to transform the value value 400 (*10^3kg) for plane weight.

2.) Suppose a group of 12 sales prices records has been sorted as follows:

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215

Partition them into three bins by each of the following methods:

a.) Equal-frequency partitioning

b.) Equal-width partitioning

3.) Given 5-dimensional numeric samples A= (1, 0, 2, 5, 3) and B= (2, 1, 0, 3, -1), find

a.) The Euclidean distance between points

b.) The city block distance

c.) The Minkowski distance for p=3

4.) Suppose that the data mining task is to cluster the following eight points (with (x,y) representing location) into three clusters:

$A_1(2,10)$, $A_2(2,5)$, $A_3(8,4)$, $A_4(5,8)$, $A_5(7,5)$, $A_6(6,4)$, $A_7(1,2)$, $A_8(4,9)$

The distance function is Euclidean distance. Suppose initially we assign A_1, A_4 and A_7 as the center of each cluster, respectively. Use the k-means algorithm to show only the three cluster centers after the first round execution.

5- Given a set of 5-dimensional categorical samples

A= (1, 0, 1, 1, 0)
B= (1, 1, 0, 1, 0)
C= (0, 0, 1, 1, 0)
D= (0, 1, 0, 1, 0)
E= (1, 0, 1, 0, 1)
F= (0, 1, 1, 0, 0)

Suppose that the samples above are distributed into two clusters:

C1={A,B,F} C2={C,D,E}

Using K-nearest neighbor algorithm with Simple Matching Coefficient(SMC) , find the classification for the following samples:

a. Y= (0, 1, 0, 1, 1) using K=1
b. Y= (0, 1, 0, 1, 1) using K=3
c. Z= (1, 1, 0, 0, 0) using K=1
d. Z= (1, 1, 0, 0, 0) using K=5

6- For the training set given below, predict the classification of the following sample using Simple Bayesian Classifier.

(Red Domestic SUV) stolen or not?

Example No:	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes