YILDIZ TECHNICAL UNIVERSITY
COMPUTER ENGINEERING DEPARTMENT
0114850 DATA MINING MIDTERM EXAM

Instructor: Assistant Prof. Songül ALBAYRAK 7^{th} April 08

Important Note: You have exactly 100 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
Points	20P	25P	25P	30P

QUESTIONS

1.) Classification using KNN (K-nearest neighbor)
Consider the one-dimensional data set shown in the table below.

value	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
class	-	-	+	+	+	-	-	+	-	-

(a) Classify the data point x = 5.0 according to its 1-, 3-, 5-, and 9-nearest neighbors (using majority vote).
(b) For the same sets of neighbors, classify the data point x = 5.0 using a distance-weighted voting approach (rather than using a simple majority vote) with weight $w_i = \frac{1}{d(x, x_i)^2}$

where x_i is the value of the i-th nearest neighbor and d(·) computes the distance.

instance	Value	X=5.0 distance	class	Order	weight
1	0.5	4.5	(−)	10	0.049
2	3.0	2	(−)	8	0.25
3	4.5	0.5	(+)	5	4
4	4.6	0.4	(+)	4	6.25
5	4.9	0.1	(+)	1	100
6	5.2	0.2	(−)	2	25
7	5.3	0.3	(−)	3	11
8	5.5	0.5	(+)	6	4
9	7.0	2	(−)	7	0.25
10	9.5	4.5	(−)	9	0.049

10P

a-) 1-NN ⇒ X=5 (1+) ⇒ Class=(+)
3NN ⇒ X=5 (2−, 1+) ⇒ Class=(−)
5NN ⇒ X=5 (3+, 2−) ⇒ Class=(+)
9NN ⇒ X=5 (5− 4+) ⇒ Class=(−)

10P

b-) 1-NN ⇒ X=5 (100+, 0−) ⇒ Class=+
3-NN ⇒ X=5 (100+, 36−) ⇒ Class=(+)
5NN ⇒ X=5 (110.25+, 36−) ⇒ Class=(+)
9NN ⇒ X=5 (114.25+, 36.549−) ⇒ Class=(+)

2.) Naïve Bayes Classification

Use Bayes' model to predict the new instance (35, medium, yes, fair)?

Age	Income	Married	Credit_rating	Loan (Class information)
21	low	no	excellent	no
25	low	no	fair	no
27	medium	no	excellent	no
28	medium	no	fair	no
29	medium	yes	fair	yes
31	medium	no	excellent	yes
32	high	yes	fair	no
36	high	yes	fair	yes
41	medium	yes	fair	yes
45	low	yes	fair	no
45	low	yes	excellent	no
47	medium	yes	fair	yes

For the numerical data use normal distribution (Gauss) to compute the probability:

$$P(x) = N(x \mid \sigma, \mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$\boxed{\text{AGE}}$

NO

$\mu = \frac{21+25+27+28+32+45+45}{7}$

1P $\mu_{\substack{AGE \\ C=NO}} = 31.8$

1P $\sigma_N^2 = 72.71$

$(\sigma_{N-1}^2 = 91.5)$

YES

$\mu = \frac{29+31+36+41+47}{5}$

1P $\mu_{\substack{AGE \\ C=YES}} = 36.8$

1P $\sigma^2 = 43.36$

$\sigma_{N-1}^2 = 54.2$

$\underbrace{P(35|No)}_{0{,}041} \times \underbrace{P(Medium|No)}_{\frac{2}{7}} \times \underbrace{P(Yes|No)}_{\frac{3}{7}} \times \underbrace{P(Fair|No)}_{\frac{4}{7}} \times \underbrace{P(N}_{\frac{7}{12}}$

(σ_{N-1} ile) 1P 2P 2P 2P 2P

$= 0{,}006$

$\underbrace{P(35|Yes)}_{0{,}052} \times \underbrace{P(Medium|Yes)}_{\frac{4}{5}} \times \underbrace{P(Yes|Yes)}_{\frac{4}{5}} \times \underbrace{P(Fair|Yes)}_{\frac{4}{5}} \times \underbrace{P(Ye}_{\frac{5}{12}}$

(σ_{N-1} ile) 1P 2P 2P 2P 2P

$= 0{,}011$

(3P)

THE RESULT FOR instace (35, medium, yes, fair) = YES

3.) Table shows a small bioinformatical dataset of 5 points in 3-dimensional space (e.g. 5 genes over 3 arrays). Start with initial points p1 and p2 as cluster centers .Then find out the clusters and label them with A and B by k-means clustering. Show every step of your calculations clearly.(Use Euclidean Distance…)

Data Point	A1	A2	A3
1	11	10	12
2	10	11	13
3	9	12	10
4	1	3	2
5	4	2	3

$C_1 = P_1$ $\quad C_2 = P_2$

$C_1 = (11, 10, 12)$ $\quad C_2 = (10, 11, 13)$

$P_3 \Rightarrow d_2(P_3, C_1) = \sqrt{12}$ — $d_2(P_3, C_2) = \sqrt{11} \Rightarrow P_3$ is member of C_2 4P

$P_4 \Rightarrow d_2(P_4, C_1) = \sqrt{249}$ — $d_2(P_4, C_2) = \sqrt{266} \Rightarrow P_4$ is member of C_1 4P

$P_5 \Rightarrow d_2(P_5, C_1) = \sqrt{194}$ — $d_2(P_5, C_2) = \sqrt{217} \Rightarrow P_5$ is member of C_1 4P

New cluster centers with their members;

$C_1 = \{P_1, P_4, P_5\}$ $\quad C_2 = \{P_2, P_3\}$

$C_1 = \left(\frac{16}{3}, \frac{15}{3}, \frac{17}{3}\right)$ 4P $\quad C_2 = \left(\frac{19}{2}, \frac{23}{2}, \frac{23}{2}\right)$ 4P

$(5.3, 5, 5.6)$ $\quad (9.5, 11.5, 11.5)$

$P_1 \Rightarrow d_2(P_1, C_1) = \sqrt{5.7^2 + 5^2 + 6.4^2} = \sqrt{100.45}$ $\quad d_2(P_1, C_2) = \sqrt{1.5^2 + 1.5^2 + 0.5^2} = \quad \Rightarrow P_1 \Rightarrow C_2$

$P_2 \Rightarrow d_2(P_2, C_1) = \sqrt{4.7^2 + 6^2 + 7.4^2} =$ $\quad d_2(P_2, C_2) = \sqrt{0.5^2 + 0.5^2 + 1.5^2} = \quad \Rightarrow P_2 \Rightarrow C_2$

$P_3 \Rightarrow d_2(P_3, C_1) = \sqrt{3.7^2 + 7^2 + 4.4^2}$ $\quad d_2(P_3, C_2) = \sqrt{0.5^2 + 0.5^2 + 1.5^2} = \quad \Rightarrow P_3 \Rightarrow C_2$

$P_4 \Rightarrow d_2(P_4, C_1) = \sqrt{4.3^2 + 2^2 + 3.6^2}$ $\quad d_2(P_4, C_2) = \sqrt{8.5^2 + 8.5^2 + 9.5^2} = \quad \Rightarrow P_4 \Rightarrow C_1$

$P_5 \Rightarrow d_2(P_5, C_1) = \sqrt{1.3^2 + 3^2 + 2.6^2}$ $\quad d_2(P_5, C_2) = \sqrt{5.5^2 + 9.5^2 + 8.5^2} = \quad \Rightarrow P_5 \Rightarrow C_1$

$C_1 = \{P_4, P_5\}$ $\quad C_2 = \{P_1, P_2, P_3\}$ 5P.

$C_1 = \left(\frac{5}{2}, \frac{5}{2}, \frac{5}{2}\right)$ $\quad C_2 = \left(\frac{30}{3}, \frac{33}{3}, \frac{35}{3}\right)$

4.) Given a training data set Y:

	A	B	C	Class
1	15	1	A	C_1
2	20	3	B	C_2
3	25	2	A	C_1
4	30	4	A	C_1
5	35	2	B	C_2
6	25	4	A	C_1
7	15	2	B	C_2
8	20	3	B	C_2

10P a)Find the best threshold (for the maximal gain) for attribute A.
10P b) Find the best threshold (for the maximal gain) for attribute B.
5P c)Find a decision tree for data set Y.
5P d)If the testing set is

A	B	C	D
10	2	A	C_2
20	1	B	C_1
30	3	A	C_2
40	2	B	C_2
15	1	B	C_1

What is the percentage of correct classification using the decision tree developed in c.

a-)

Instance	Ordered A	Class
1	15	C_1
7	15	C_2
2	20	C_2
8	20	C_2
3	25	C_1
6	25	C_1
4	30	C_1
5	35	C_2

$\text{Info}(T) = -\frac{4}{8}\log_2\frac{4}{8} - \frac{4}{8}\log_2\frac{4}{8} = 1 \text{ bit}$

Threshold can be $= \{15, 20, 25, 30\}$

For $Th=15 \Rightarrow E = \frac{2}{8}\left(I(1,1)\right) + \frac{6}{8}\left(I(3,3)\right) = 1$
(≤ 15)

Gain(15) = 0

for $Th=20 \Rightarrow E = \frac{4}{8}\left(I(1,3)\right) + \frac{4}{8}\left(I(3,1)\right) = 0.41$
(≤ 20)

$\text{Gain}(20) = 1 - 0.418 = 0.582$

for $Th=25 \Rightarrow E = \frac{6}{8}\left(I(3,3)\right) + \frac{2}{8}\left(I(1,1)\right) = 1$
(≤ 25)

Gain(25) = 0

for $Th=30 \Rightarrow E = \frac{7}{8}\left(I(4,3)\right) + \frac{1}{8}\left(I(1,0)\right)$
(≤ 30)
$= 0.862$

Good Luck…

Gain(30) = 0,138

The maximal gain is Gain(20) = 0.582, so the threshold is 20

Instance	Ordered B	Class
1	1	C_1
3	2	C_1
5	2	C_2
7	2	C_2
2	3	C_2
8	3	C_2
4	4	C_1
6	4	C_1

Threshold can be $= \{1, 2, 3\}$

For $Th = 1 \Rightarrow E = \frac{1}{8} I(1,0) + \frac{7}{8} I(4,3) = 0{,}862$
(≤ 1)

$Gain(1) = 1 - 0{,}862 = 0{,}138$

For $Th = 2 \Rightarrow E = \frac{4}{8} I(2,2) + \frac{4}{8} I(2,2) = 1$
(≤ 2)

$Gain(2) = 1 - 1 = 0$

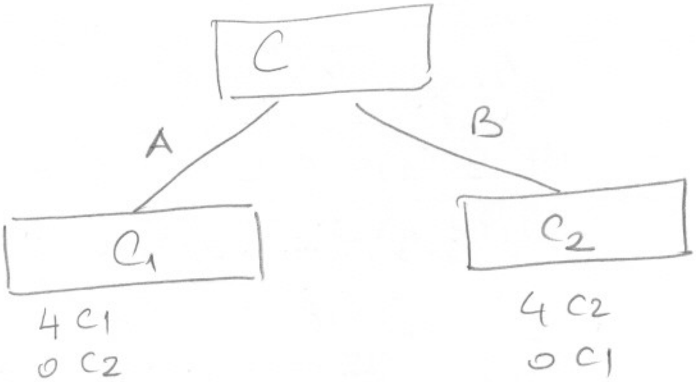
For $Th = 3 \Rightarrow E = \frac{6}{8} I(2,4) + \frac{2}{8} I(2,0) = 0.688$
(≤ 3)

$Gain(3) = 1 - 0.688 = 0.312$

The maximal gain is $G(3) = 0{,}312$, so the theshold is 3

c.) $Info_C(T) = \frac{4}{8} \underbrace{I(4,0)}_{0} + \frac{4}{8} \underbrace{I(4,0)}_{0} = 0$

$Gain(C) = 1 - 0 = 1$



d.) for the test set

Instances	Classified as	Actual class	
1	C_1	C_2	False
2	C_2	C_1	false
3	C_1	C_2	false
4	C_2	C_2	True
5	C_2	C_1	false

The percentace of correct classificatio is %20

(1/5)