Name: Zhengai Dong

CSE 3521 Artificial Intelligence SU'19

Homework Assignment #6 (19 points)

Due: Friday, June 21

1. After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive (+) for a serious disease (known as disease "X"). The accuracy of the test is as follows:

   The probability of testing positive (+) given that you have disease X is 0.98

   The probability of testing negative (-) given that you don't have disease X is 0.85.

   The good news is that disease X is rare, striking only one in 8,000 people.

   Using Bayes Rule, what is the chance that you actually have the disease (i.e., what is P(X | +))? SHOW YOUR WORK! (5 pts)

$(x_2 - x_1)^2 + (y_2 - y_1)^2$

2. Manually perform K-Means clustering for 4 iterations on the following 1-D dataset (see that K=3 below). Report the updated cluster assignment for each datapoint and the new means at each iteration. (7 pts)

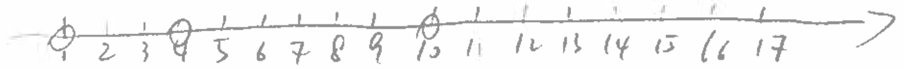
non-parametric

x_1 x_2 x_{11}

Data = [1 3 5 6 8 9 10 12 13 16 17]

μ_1 μ_2 μ_3

Initial means = [1 4 10] randomly cluster centroids



$\|x^{(i)} - \mu_k\|$

μ_1	0	2	4	5	7	8	9	11	12	15	16
μ_2	3	1	1	2	4	5	6	8	9	12	13
μ_3	9	7	5	4	2	1	0	2	3	6	7

Avg

$avg1 = \frac{\mu_1.sum}{\mu_1.size.to2f} = 8.0909$

avg2 = 5.8181

avg3 = 4.818

(cont.)

In Cartesian coordinate, if $p = (p_1, p_2 \cdots p_n)$ and $q = (q_1, \text{—} q_n)$ are two points in Euclidean n-space, then the distance (d) from p to q is given by Pothagoream formula:

$$d(p,q) = d(q,p) = \sqrt{(q_1-p_1)^2 + (q_2-p_2)^2 + \cdots + (q_n-p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

3. Use **K-Nearest Neighbors** to classify the 2-D point (x,y)=(2,3).

Use the following training data to make your determination:

x	1	1	0	2	3	3
y	2	4	3	5	5	3
Class labels	red	red	red	blue	blue	green
	1	2	3	4	5	6

3.1. Calculate the (Euclidean) distance from the point to each training data point. (3pts)

$\sqrt{(x_i - x)^2 + (y_i - y)^2}$

3.2. Find the 3 closest data points (i.e., K=3). From the class label of those 3, how should you classify the point? (1pt)

Kth nearest neighbor around given point

3.3. How would you classify the point for K=1? K=5? (2pts)

3.4. Plot the point together with the training data points. Do your answers to the previous two questions agree with this plot? Explain. (1pt)

	(1,2)	(1,4)	(0,3)	(2,5)	(3,5)	(3,3)
	red	red	red	blue	blue	green
3.1 $\sqrt{(x_i - x)^2 + (y_i - y)^2} =$	1.414	1.414	2	2	2.36	1

3.2 when k=3 : 3 closer pots (3,3) green, (1,2) red, (1,4) red ⇒ So red.

3.3

k=1: green

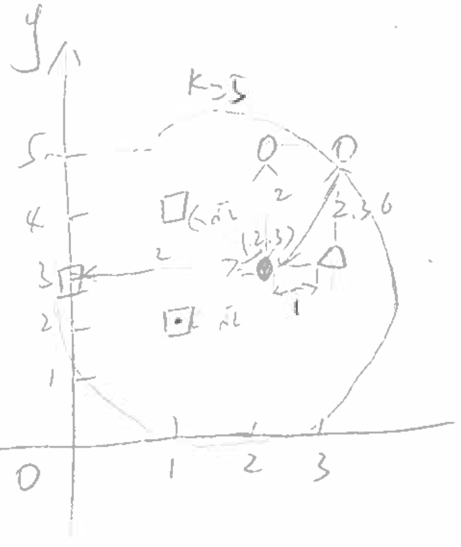
k=5: blue

red: □

blue: o

green: △

3.4



1.

let X = has a serious disease

X' = doesn't has a serious disease

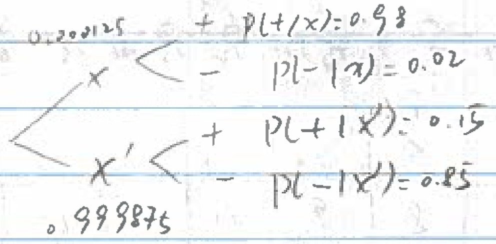
+ = tested positive

− = tested negative

$$P(+|X) = 0.98$$

$$P(-|X') = 0.85$$

$$P(X) = 1/8000 = 0.000125$$



$$P(X|+) = \frac{P(X\cap+)}{P(+)} = \frac{P(+|X)P(X)}{P(+|X)P(X)+P(+|X')P(X')} = \frac{0.98 \times 0.000125}{0.98(0.000125)+0.15(0.9998}$$

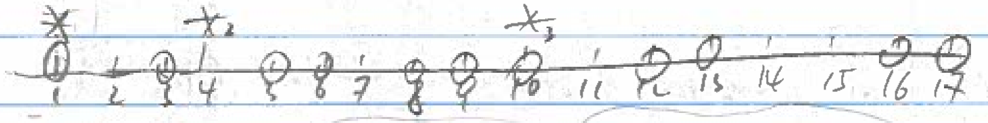
$$= \frac{0.0001225}{0.1501} = 0.0008161$$

2.

let Data = $x_1, x_2 - x_{11}$ = [1, 3, 3, 6, 8, 9, 10, 12, 13, 16, 17]

Initial means = μ_1, μ_2, μ_3 = [1, 4, 10]

Distance = $\|x^{(i)} - \mu_k\|$



Loop1

Data :	1	3	5	6	8	9	10	12	13	16	17
D1 :	0	2	4	5	7	8	9	11	12	15	16
D2 :	3	1	1	2	4	5	6	8	9	12	13
D3	9	7	5	4	2	1	0	2	3	6	7

$C_1(1) = [1]$

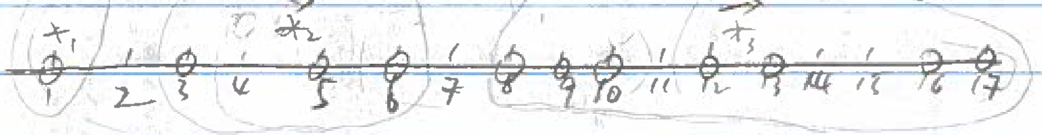
$\mu_1 = 1$

$C_2(4) = [3, 5, 6]$

$\mu_2 = 4.67$

$C_3(10) = [8, 9, 10, 12, 13, 16, 17]$

$\mu_3 = 12.14$



Loop2

	1	3	5	6	8	9	10	12	13	16	17
D1 :	0	2	4	5	7	8	9	11	12	15	16
D2	4.5	2.5	0.5	0.5	2.5	3.5	4.5	6.5	7.5	10.5	11.5
D3	11.14	9.14	7.14	6.14	4.14	3.14	2.14	0.14	0.86	3.85	4.86

$C_1(1) = [1\ ]$

Mean 2 = [1, 4.67, 12.14]

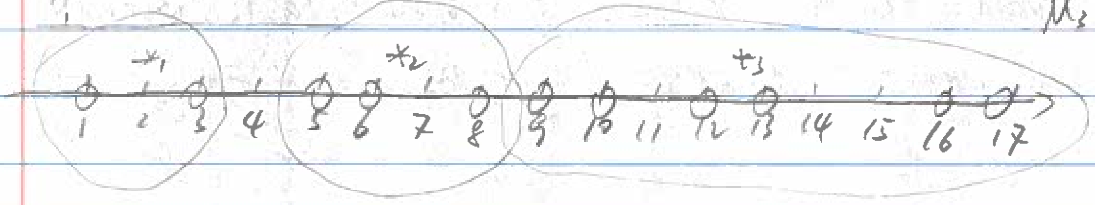
$\mu_1 = 2$

$C_2(4.67) = [3, 5, 6, 8]$

$\mu_2 = 5.5$

$C_3(12.14) = [9, 10, 12, 13, 16, 17]$

$\mu_3 = 12.83$



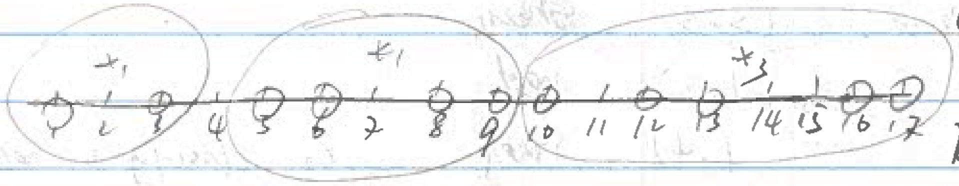
$\frac{19}{19}$

Mean3=[2, 5.5, 12.83]

Loop3	1	3	5	6	8	9	10	12	13	16	17
D_1 :	1	1	3	4	6	7	8	10	11	14	15
D_2 :	4.5 ~~5.33~~	2.5 ~~3.33~~	0.5 ~~1.33~~	0.5 ~~0.33~~	2.5 ~~1.67~~	3.5 ~~2.67~~	4.5 ~~3.67~~	6.5 ~~5.67~~	7.5 ~~6.67~~	10.5 ~~9.67~~	11.5 ~~10.67~~
D_3 :	11.83	9.83	7.83	6.83	4.83	3.83	2.83	0.83	0.17	3.17	4.17

$C_1(2)=[1,3]$

$\mu_1 = 2$



$C_2(6.33)=[5,6,8,9]$

$\mu_2 = 7$

$C_3(12.83)=[10,12,13,16,17]$

$\mu_3 = 13.6$

Mean 4 =[2, 7, 13.6]

Loop4	1	3	5	6	8	9	10	12	13	16	17
D_1	1	1	3	4	6	7	8	10	11	14	15
D_2	6	4	2	1	1	2	3	5	6	9	10
D_3	12.6	10.6	8.6	7.6	5.6	4.6	3.6	1.6	0.6	2.4	3.4

$C_1(2) = [1,3]$

$\mu_1 = 2$

$C_2(7) = [5,6,8,9,10]$

$\mu_2 = 7.6$

$C_3(13.6) = [12,13,16,17]$

$\mu_3 = 14.5$

3.

3.1

x	1	1	0	2	3	3
y	2	4	3	5	5	3
label	red	red	red	blue	blue	green
Euc Dis	1.414	1.414	2	2	2.36	1

Point (2.3)

3.2

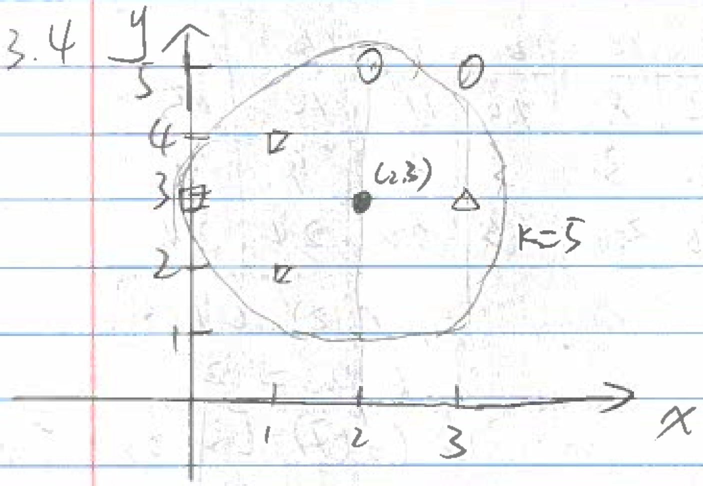
3 closest point : (3,3) green
(1,2) red
(1,4) red , so classify to red.

3.3

k=1 : green

k=5 : red

3.4



The previou answe do agree with this plot, when we choose 1, the nearest neighbor to point is (3,3), so it's green, When we use k=5, there are 5 point, but 3 of them are red, and it's the fact from graph, red has the most points in k=5.