CSCI3230 / ESTR3108 2021-22 First Term Assignment 2

I declare that the assignment here submitted is original except for source material explicitly acknowledged, and that the same or closely related material has not been previously submitted for another course. I also acknowledge that I am aware of University policy and regulations on honesty in academic work, and of the disciplinary guidelines and procedures applicable to breaches of such policy and regulations, as contained in the following websites.

University Guideline on Academic Honesty:
`http://www.cuhk.edu.hk/policy/academichonesty/`
Faculty of Engineering Guidelines to Academic Honesty:
`http://www.erg.cuhk.edu.hk/erg-intra/upload/documents/ENGG_Discipline.pdf`
Student Name: ***Lai Man Hin***
Student ID : ***1155136167***

1a) The optimization problem is

$\max_\alpha \left[\sum_{i=1}^{8} \alpha_i - \frac{1}{2} \sum_{i=1}^{8} \sum_{j=1}^{8} \alpha_i \alpha_j y_i y_j x_i^T x_j \right]$

$s.t. \sum_{i=1}^{8} a_i y_i = 0, a_i \geq 0$

By using scikit-learn library, the optimal α is

$$\alpha_1 = 0.03172284, \alpha_2 = 0, \alpha_3 = 0.1360743, \alpha_4 = 0,$$
$$\alpha_5 = 0, \alpha_6 = 0, \alpha_7 = 0, \alpha_8 = -0.16779714$$
$$\text{where } \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5 + \alpha_6 + \alpha_7 + \alpha_8 = 0$$

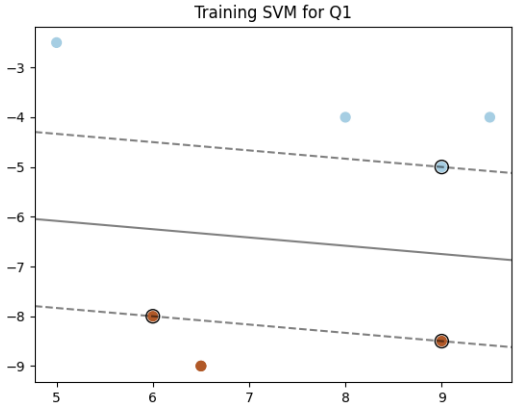
(The library calculated $\alpha_i \leftarrow y_i \alpha_i$,
α_8 should be 0.16779714 if following $\alpha_i \geq 0$)

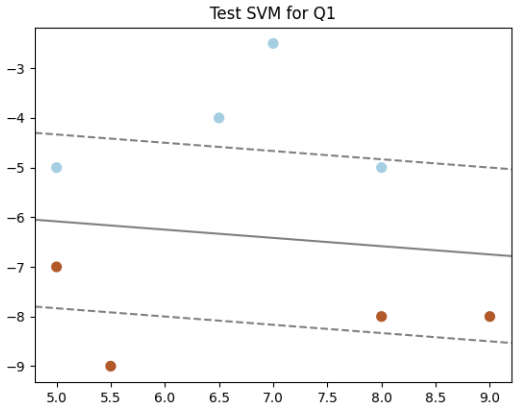
The support vector is $\left\{ \begin{bmatrix} 6 \\ -8 \end{bmatrix}, \begin{bmatrix} 9 \\ -8.5 \end{bmatrix}, \begin{bmatrix} 9 \\ -5 \end{bmatrix} \right\}$

$\mathbf{w}$ is $\begin{bmatrix} -0.09516853 \\ -0.57142857 \end{bmatrix}$

$$b = \frac{1}{|S|} \sum_{s \in S} \left(\frac{1}{y_s} - w^T x_s \right) = -3.000556512$$

We can get the hyperplane:
$\begin{bmatrix} -0.09516853 \\ -0.57142857 \end{bmatrix} x - 3.000556512 = 0$





b) For all test case predicted result y_i indexed by i:

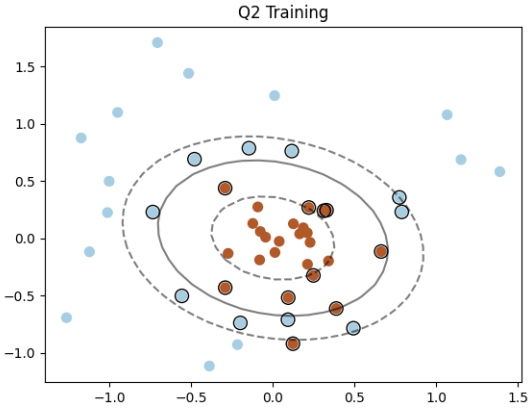
$$\begin{cases} y_i = -1 \text{ if } \begin{bmatrix} -0.09516853 \\ -0.57142857 \end{bmatrix}^T x_i - 3.000556512 < 0 \\ y_i = 1 \text{ if } \begin{bmatrix} -0.09516853 \\ -0.57142857 \end{bmatrix}^T x_i - 3.000556512 > 0 \end{cases}$$

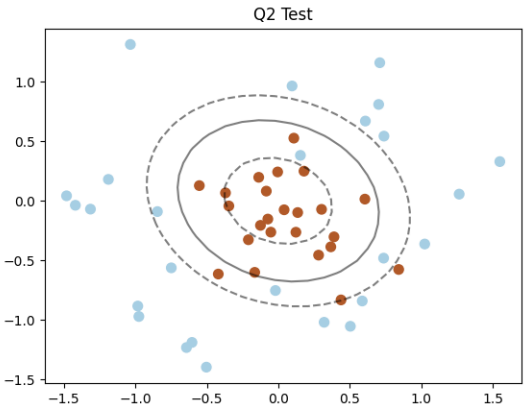
The prediction obtained is:
$y_1 = 1, y_2 = 1, y_3 = 1, y_4 = 1, y_5 = -1, y_6 = -1, y_7 = -1, y_8 = -1$

c) Removing the 2nd data point in training set will **NOT** change the optimal hyperplane, but removing the 8th one **WILL** change it. Because 2nd one is **NOT** the support vectors, but 8th one **IS**.

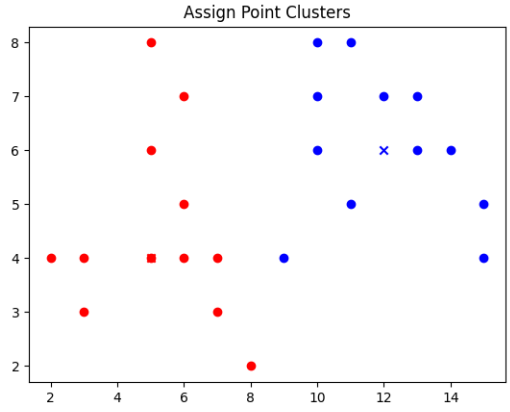
2) Polynomial kernel is used to be the kernel function since the positive cases are centralized like a circle and the negative cases are surrounding the positive result. The degree chosen for the polynomial kernel is 2 but not higher value is to prevent overfitting. (Accuracy of the model is 92%)

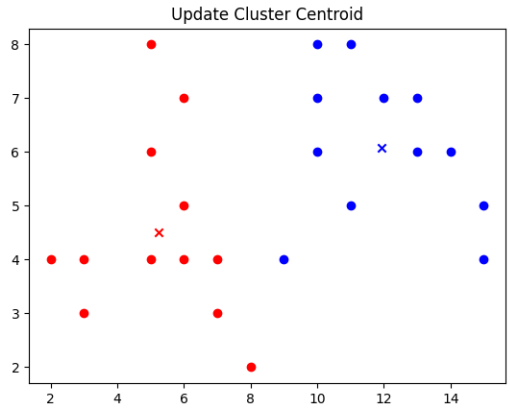
See attached file *CSCI3230Assg2Q2.py* for coding details.





3a) By calculating the distance between the initial cluster points and the data given, the data points will be put into the closest cluster centroid by calculating
$\sqrt{(data_x - centroid_x)^2 + (data_y - centroid_y)^2}$
So from the Assign Point Cluster Figure,
Cluster center (5, 4): a, b, c, d, e, f, g, h, i, j, k, l
Cluster center (12, 6): m, n, o, p, q, r, s, t, u, v, w, x
New centroid for (5, 4)
$= (\frac{2+3+3+5+5+5+6+6+6+7+7+8}{12}, \frac{4+4+3+8+6+4++7+5+4+4+3+2}{12})$
$= (5.25, 4.5)$
New centroid for (12, 6)
$= (\frac{10+10+10+11+11+12+13+13+14+15+15+9}{12},$
$\frac{8+7+6+8+5+7+7+6+6+5+4+4}{12})$
$= (11.916667, 6.083333)$





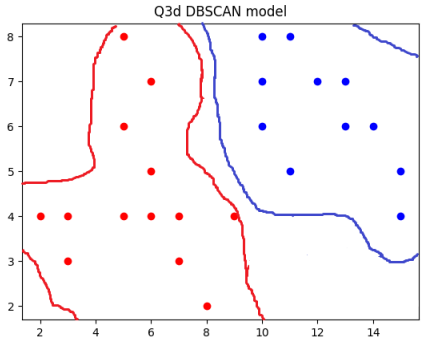
b) Consider the following table, all the point that having neighbourhood relationship will be marked as -1,

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x
a	-1	-1	-1																					
b	-1	-1	-1			-1																		
c	-1	-1	-1																					
d				-1	-1		-1																	
e				-1	-1	-1	-1	-1																
f		-1			-1	-1		-1	-1	-1														
g				-1	-1		-1	-1																
h					-1	-1	-1	-1	-1	-1														
i						-1		-1	-1	-1	-1													
j						-1		-1	-1	-1	-1													-1
k									-1	-1	-1	-1												
l											-1	-1												
m													-1	-1	-1	-1								
n													-1	-1	-1	-1		-1						
o													-1	-1	-1		-1							
p													-1	-1		-1		-1						
q															-1		-1							
r														-1		-1		-1	-1	-1				
s																		-1	-1	-1	-1			
t																		-1	-1	-1	-1			
u																			-1	-1	-1	-1		
v																					-1	-1	-1	
w																						-1	-1	
x										-1														-1

By scanning each rows, if Sum of that row ≤ -3, it will be a core point. Therefore,
Core points: a, b, c, d, e, f, g, h, i, j, k, m, n, o, p, r, s, t, u, v

c) Yes, the intermediate points on the chain is $u \leftarrow t \leftarrow r \leftarrow n \leftarrow o$.

d) The right hand side shown the points processed by DBSCAN model, which:
Red color represents the Cluster 1: a, b, c, d, e, f, g, h, i, j, k, l, x;
Blue color represents the Cluster 2: m, n, o, p, q, r, s, t, u, v, w
Reminded that there is **NO noise points**.



4) For **MIN distance**,

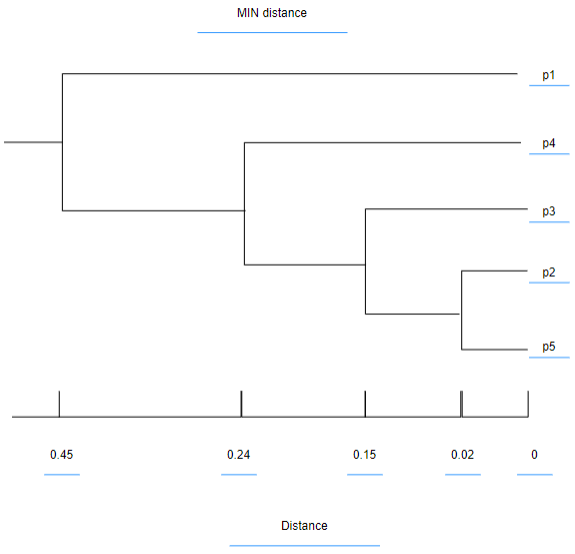
$$M_{min} = \begin{array}{c} \\ p1 \\ p2 \\ p3 \\ p4 \\ p5 \end{array} \begin{array}{c} \begin{array}{ccccc} p1 & p2 & p3 & p4 & p5 \end{array} \\ \begin{pmatrix} 0 & - & - & - & - \\ 0.9 & 0 & - & - & - \\ 0.59 & 0.36 & 0 & - & - \\ 0.45 & 0.53 & 0.56 & 0 & - \\ 0.65 & 0.02 & 0.15 & 0.24 & 0 \end{pmatrix} \end{array} \Rightarrow \text{pick p2, p5} = 0.02 \text{ because they are minimum}$$

$$\Rightarrow \begin{array}{c} \\ p25 \\ p1 \\ p3 \\ p4 \end{array} \begin{array}{c} \begin{array}{cccc} p25 & p1 & p3 & p4 \end{array} \\ \begin{pmatrix} 0 & - & - & - \\ 0.45 & 0 & - & - \\ 0.15 & 0.59 & 0 & - \\ 0.24 & 0.45 & 0.56 & 0 \end{pmatrix} \end{array} \Rightarrow \text{pick p25, p3} = 0.15 \text{ because they are minimum}$$

$$\Rightarrow \begin{array}{c} \\ p253 \\ p1 \\ p4 \end{array} \begin{array}{c} \begin{array}{ccc} p253 & p1 & p4 \end{array} \\ \begin{pmatrix} 0 & - & - \\ 0.45 & 0 & - \\ 0.24 & 0.59 & 0 \end{pmatrix} \end{array} \Rightarrow \text{pick p253, p4} = 0.24 \text{ because they are minimum}$$

$$\Rightarrow \begin{array}{c} \\ p2534 \\ p1 \end{array} \begin{array}{c} \begin{array}{cc} p2534 & p1 \end{array} \\ \begin{pmatrix} 0 & - \\ 0.45 & 0 \end{pmatrix} \end{array} \Rightarrow \text{pick p2534, p1} = 0.45 \text{ because they are minimum}$$

The dendrogram is shown below:



For **MAX distance**,

$$M_{max} = \begin{array}{c} \\ p1 \\ p2 \\ p3 \\ p4 \\ p5 \end{array} \begin{array}{c} \begin{array}{ccccc} p1 & p2 & p3 & p4 & p5 \end{array} \\ \begin{pmatrix} 0 & - & - & - & - \\ 0.9 & 0 & - & - & - \\ 0.59 & 0.36 & 0 & - & - \\ 0.45 & 0.53 & 0.56 & 0 & - \\ 0.65 & 0.02 & 0.15 & 0.24 & 0 \end{pmatrix} \end{array} \Rightarrow \text{pick p2, p5} = 0.02 \text{ because they are minimum}$$

$$\Rightarrow \begin{array}{c} \\ p25 \\ p1 \\ p3 \\ p4 \end{array} \begin{array}{c} \begin{array}{cccc} p25 & p1 & p3 & p4 \end{array} \\ \begin{pmatrix} 0 & - & - & - \\ 0.9 & 0 & - & - \\ 0.36 & 0.59 & 0 & - \\ 0.53 & 0.45 & 0.56 & 0 \end{pmatrix} \end{array} \Rightarrow \text{pick p25, p3} = 0.36 \text{ because they are minimum}$$

$$\Rightarrow \begin{array}{c} \\ p253 \\ p1 \\ p4 \end{array} \begin{array}{c} \begin{array}{ccc} p253 & p1 & p4 \end{array} \\ \begin{pmatrix} 0 & - & - \\ 0.9 & 0 & - \\ 0.56 & 0.45 & 0 \end{pmatrix} \end{array} \Rightarrow \text{pick p1, p4} = 0.45 \text{ because they are minimum}$$

$$\Rightarrow \begin{array}{c} \\ p253 \\ p14 \end{array} \begin{array}{c} \begin{array}{cc} p253 & p14 \end{array} \\ \begin{pmatrix} 0 & - \\ 0.9 & 0 \end{pmatrix} \end{array} \Rightarrow \text{pick p253, p14} = 0.9 \text{ because they are minimum}$$

The dendrogram is shown below: