1) Dataset

class A : $\{[0.5, 0.5]^T, [0.75, 0.25]^T, [0.25, 0.75]^T, [1, 1]^T, [2, 2]^T\}$

class B : $\{[-2, -2]^T, [-0.25, -0.25]^T, [-1.5, -2.5]^T, [-3, -2]^T, [-2, -3]^T\}$

$q = [0, 0]^T$

distances of q from each of the 10 points. (euclidean distances)

A) 1) $[0.5, 0.5]^T \rightarrow 0.707$

A) 2) $[0.75, 0.25]^T \rightarrow 0.790$

A) 3) $[0.25, 0.75]^T \rightarrow 0.790$

A) 4) $[1, 1] \rightarrow 1.414$

A) 5) $[2, 2] \rightarrow 2.828$

B) 6) $[-2, -2] \rightarrow 2.828$

B) 7) $[-0.25, -0.25]^T \rightarrow 0.354$

B) 8) $[-1.5, -2.5]^T \rightarrow 2.915$

B) 9) $[-3, -2]^T \rightarrow 3.6$

B) 10) $[-2, -3]^T \rightarrow 3.6$

If $K = 1$

Using KNN, point $q [0, 0]^T$ is closest to $[-0.25, -0.25]^T$

distance = 0.354.

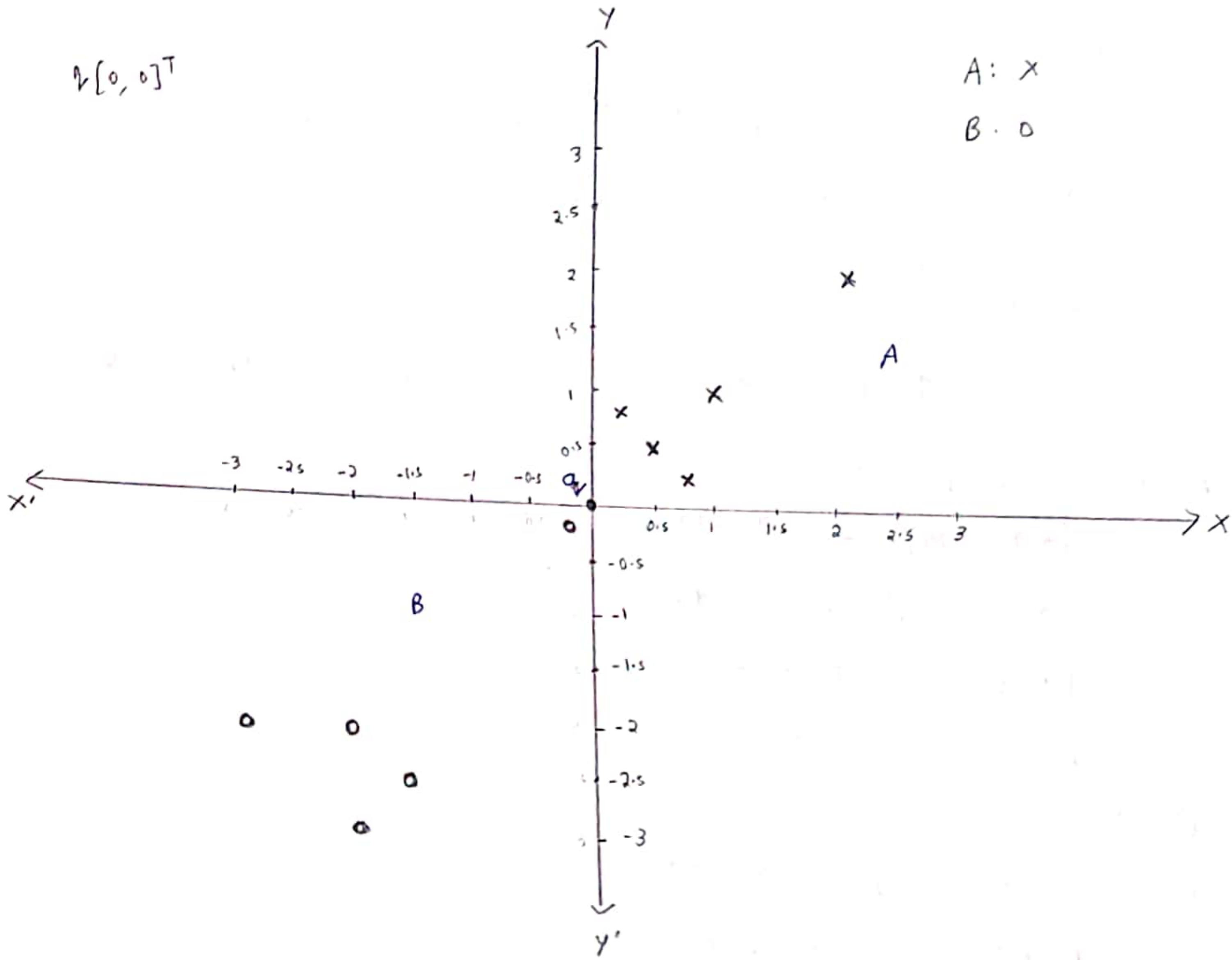
$\therefore$ q belongs to class B

If $K = 3$

Using KNN, point $q [0, 0]^T$ is closest to $[-0.25, -0.25]^T$ (dis = 0.354), $[0.5, 0.5]^T$ (dis = 0.707), $[0.75, 0.25]^T$ (dis = 0.790). Since two points belong to class A, q belongs to class A

NOTE - $[0.25, 0.75]^T$ is another point at dis 0.790 in class A again.

Scanned by CamScanner



Graph for KNN (class A, class B)

Scanned by CamScanner

	Name	Age	Height	Role	Batting avg	Bowling avg	No of matchs
1	Virat Kohli	30	175	1	59.4	166.25	236
2	Rohit Sharma	32	170	1	48.92	64.38	215
3	Mayank Agarwal	28	175	1	NA	NA	NA
4	Kuldeep Yadav	24	168	2	12.62	23.97	51
5	Mohammed Shami	28	178	2	7.39	24.76	67
6	Jasprit Bamrah	25	178	2	3.8	21.88	58
7	Bhuvaneshwar Jumar	29	175	2	14.58	34.98	111
8	Yuzvendra Chalal	29	168	2	7.8	26.36	49
9	Rishab Pant	21	170	3	26.12	160	9
10	Lokesh Rahul	27	180	3	39.11	160	23
11	MS Dhoni	38	175	3	50.58	31	350
12	Dinesh Karthik	34	170	3	30.21	160	94
13	Hardik Pandya	25	183	4	29.91	40.65	54
14	Kedar Jhadav	34	165	4	43.24	35.96	65
15	Ravindra Jadeja	30	173	4	30.61	35.9	153
	Kumar Sangakara	41	178	1	41.99	160	404
	David Warner	32	170	1	45.36	160	116
	AB De Villiers	35	178	3	53.5	28.86	228

2) i) Player similar to Sangakara using KNN

Representing Sangakara in the same feature space and applying KNN we can write distance (euclidian distance) by the formula. (Shown in data set)

dis from other players $= \sqrt{(x_1-x_1')^2+(x_2-x_2')^2+(x_3-x_3')^2+(x_4-x_4')^2+(x_5-x_5')^2+(x_6-x_6')^2} \rightarrow$ 1

($x_1, x_2 \cdots x_6$ is Sankakara's feature, $x_1' x_2' \cdots x_6'$ is the other players features)

distances obtained are :-

✓ Kohli	169.3	Jasprit B	376.8	$\boxed{\text{MS Dhoni} \quad 140.1}$	
Rohit S	212.2	Bhuvaneshwar J	319.4	Dinesh K	310.6
Mayank A	NA (did not play ODI)	Yuzvindra Chahal	381.1	Hardik P	370.3
Kuldeep Y	379.9	Rishab P	395.9	Kedar J	361.2
Mohamed S	365.0	Lokesh R	331.2	Ravindra J	280.5

MS Dhoni with KNN algo seems to be the shortest distance / closest / similar to Sangakara.

NOTE : Same formula of euclidian distance 1 was used in next 2 questions. Bowling avg was considered a high value of 160 if not given.

2) Player similar to David Warner. Representing David in the same feature space and applying the above formula we have respective distances as: -

Kohli	121.0	Jasprit B	155.8	MS Dhoni	267.3
Rohit S	137.6	Bhuvenshwar J	128.9	$\boxed{\text{Dinesh K} \quad 26.8}$	
Mayank A	NA (did not play ODI)	Yuzvendra Chahal	156.1	Hardik P	136.2
Kuldeep Y	154.4	Rishab P	109.2	Kedar J	134.2
Mohamed S	149.0	Lokesh R	93.8	Ravindra J	130.4

Dinesh Karthik not Rohit sharma seems to be similar to David Warner with the shortest distance of 26.8

3) Testing AB De Villiers (Player of choice)

Assigning the same feature space for AB De Villiers and computing distances from other players for KNN, we have,

Kohli	137.8	Jasprit B	177.5	MS Dhoni	122.1
$\boxed{\text{Rohit S} \quad 39.0}$		Bhuvaneshwar T	123.6	Dinesh K	184.1
Mayank A	NA (did not play ODI)	Yuzvendra Chahal	185.1	Hardik P	176.3
Kuldeep Y	182.3	Rishabh P	257.2	Kedar J	163.9
Mohammad S	167.6	Lokesh R	243.9	Ravindra J	77.0

Rohit Sharma is most similar to AB De Villiers with the smallest distance of 39.0.

↳ The results of KNN are not meaningful here as parameters such as age, height, no of matches etc are all on different scales.

We can define a weighted distance function to improve similarity.

↳ Using weights for standardization/normalization

In the euclidean distance formula, we define w, such that in

$$d_{x,b} = \left(\sum_{i=1}^{n} (w_i (x_i - b_i)^2)\right)^{1/2}$$

w is the reciprocal of each measurement's variance. Hereby all measurements will be on same scale (say normalized from 0 to 1)

↳ Using weights for importance

(more imp ones)

We can assign higher weights to certain properties to ensure ~~that~~ they have more contribution in the weight func.

By using these techniques the result of the above KNN we applied can be improved.

3)

↳ $x_1 = [1,2,3]^T$ $x_2 = [0,3,4]^T$, $x_3 = [2,4,4]^T$

Find a w such that $w^T x_1 < 0$ $w^T x_2 > 0$ $w^T x_3 > 0$

∴ 3 eq obtained are, considering w^T $[w_1, w_2, w_3]^T$

$$w_1 + 2w_2 + 3w_3 < 0 \quad (1)$$

$$3w_2 + 4w_3 > 0 \quad (2)$$

$$2w_1 + 4w_2 + 4w_3 > 0 \quad (3)$$

Solving, (3) − (2), $2w_1 + w_2 > 0$ $2w_1 > -w_2$ or $w_1 > \frac{-w_2}{2}$

We can also write $4w_1 > -2w_2 \rightarrow (4)$

adding (1) + (4), $w_1 + 3w_3 < 4w_1$

$3w_3 < 3w_1$ or $w_3 < w_1$ or $\boxed{4w_3 < 4w_1}$

By (2), $\boxed{4w_3 > -3w_2}$

Combining we have, $-3w_2 < 4w_3 < 4w_1$

dividing by 4, $\boxed{-\frac{3}{4} w_2 < w_3 < w_1}$ ☓

∴ Let $w_1 = 0.2$, $w_2 = 2.05$ $w_3 = -1.5$

~~So $w^T = [\ ,\ ,\ ]$~~ ~~$w^T = [\ ,\ ,\ ]$~~

~~(Satisfies the inequality $\frac{-3}{4} < 2 <$)~~

So $w^T = [0.2, 2.05, -1.5]$ $w^T = [0.2, 2.05, -1.5]$

satisfies the inequality obtained too, $-\frac{3}{4} \times 2.05 < -1.5 < 0.2$

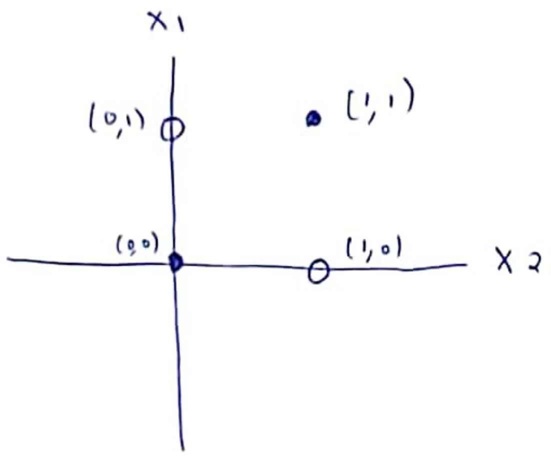
i.e. $-1.5375 < -1.5 < 0.2$

A set of w such $w^T = [0.2, 2.05, -1.5]$ satisfies the above eqs

) b) Four points x_1, x_2, x_3, x_4 in 4D

no w should keep x_1, x_2 on one side, x_3, x_4 on the other.

Let us consider the linear inseparable pattern of logical XOR function. The 4 points cannot be separated by a single line (w) such that 2 points (0,0) and (1,1) are on one side and (1,0), (0,1) on the other.



x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	0

$$y = x_1 \oplus x_2$$

Similarly extending this to 4D, Let us say the 4 points are

$$\left(\begin{array}{ll} x_1 & 0,0,0,0 \\ x_4 & 0,1,0,0 \\ x_3 & 1,0,0,0 \\ x_2 & 1,1,0,0 \end{array}\right)$$

We could not / proved no line separates this in 2D, hence no plane(w) can keep x_1 and x_2 on one side and x_3 and x_4 on the other with these set of points in 4D

(0, 0, 0, 0) (0, 1, 0, 0) (1, 0, 0, 0) (1, 1, 0, 0)