24/11/21

eg) Datapoints = {2, 3, 4, 10, 11, 12, 20, 25, 30}

Given two cluster, divide these points into 2 cluster using k-means algo?

Soln

Step 1°- Select randomly two points from Datapoints.

$M_1 = 4$ $\quad M_2 = 12$

$K_1 = \{2, 3, 4, \}$ $\quad K_2 = \{10, 11, 12, 20, 25, 30\}$

$X = \{2, 3, 4, 10, 11, 12, 20, 25, 30\}$

Step 2°- We calculate centroid for each cluster

$C_1 = \frac{2+3+4}{3} = 3$, $C_2 = \frac{10+11+12+20+25+30}{6}$

$C_2 = \frac{108}{6} = 18$

Step3:- again form the cluster using centroid (3, 18)

K_1, K_2

$K_1 = \{2, 3, 4, 10\}$

$K_2 = \{11, 12, 20, 25, 30\}$

$X = \{2, 3, 4, 10, 11, 12, 20, 25, 30\}$

Step4:- again we calculate centroid

$C_1 = \frac{2+3+4+10}{4} = 4.75 \approx 5$

$C_2 = \frac{11+12+20+25+30}{5} = 19.6 \approx 20$

new centroid (5, 20)

$K_1 = \{2, 3, 4, 10, 11, 12\}$

$K_2 = \{20, 25, 30\}$

$X_2 = \{2, 3, 4, 10, 11, 12, 20, 25, 30\}$

again Centroid

$$C_1 = \frac{2+3+4+10+11+12}{6} = 7$$

$$C_2 = \frac{20+25+30}{3} = 25$$

new centroid = (7, 25) ✓

$K_1 = \{2, 3, 4, 10, 11, 12\}$

$K_2 = \{20, 25, 30\}$

$X = \{2, 3, 4, 10, 11, 12, 20, 25, 30\}$

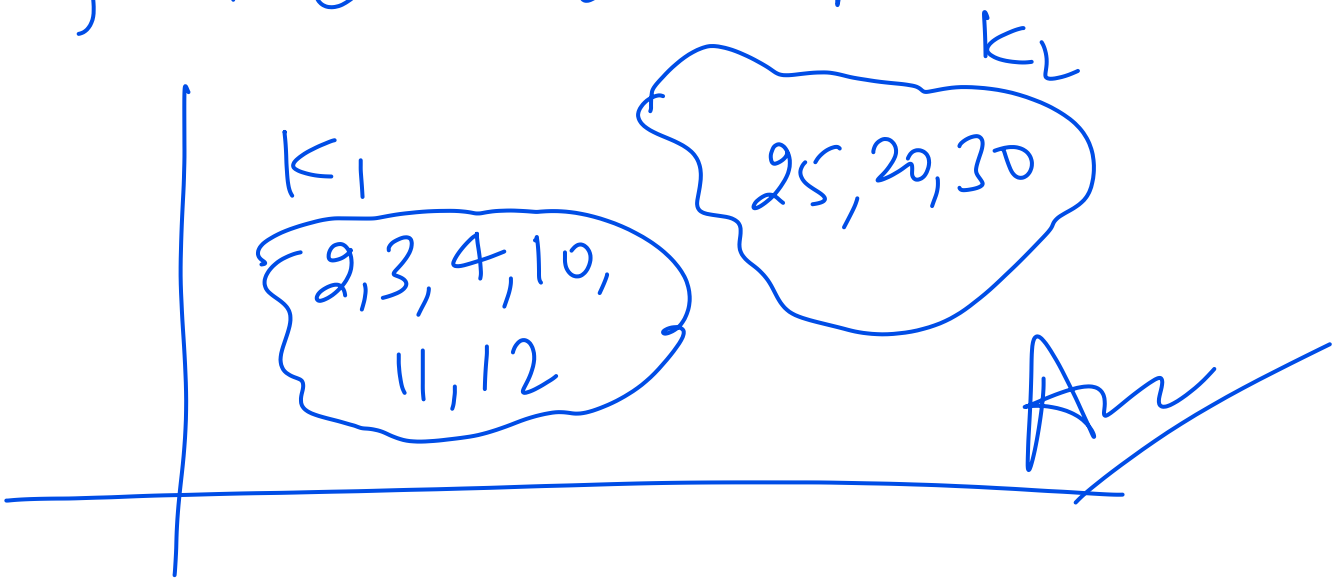
again calculate centroid

$$C_1 = \frac{2+3+4+10+11+12}{6} = 7$$

$$C_2 = \frac{20+25+30}{3} = 25$$

Now new centroid is (7, 25) ✓ which is same as previous centroid

The final cluster data points are



Ans

ex2

	Height(H)	Weight(W)
R1	185	72
R2	170	56
R3	168	60
R4	179	68
R5	182	72
R6	188	77
R7	180	71
R8	180	70
R9	183	84
R10	180	88
R11	180	67
R12	170	76

Divide the given sample data into two cluster using K-means algo and Euclidean distance?

Soln Euclidean distance $= \sqrt{(x_H - c_1)^2 + (x_W - c_2)^2}$

x_1, x_W : observed value (Dataset)

c_1, c_2 : centroid value

Step 1 :- Select two random points

	H	W	Centroid
R_1 :	185	72	(185, 72)
R_2 :	170	56	(170, 56)

$K_1 = \{R_1, \quad\}$

$K_2 = \{R_2, \quad\}$

	H	W
$R_3 \rightarrow$	168	60

E.D of R_3 from K_1 $= \sqrt{(168-185)^2 + (60-72)^2}$

$= \sqrt{289 + 144}$

$= 20.80$

E.D of R_3 from K_2

$$k_2 = \sqrt{(168-170)^2 + (60-56)^2}$$

$$= \sqrt{4 + 16}$$

$$= 4.48$$

$K_1 = \{R_1\}$

$K_2 = \{R_2, R_3\}$

again calculate centroid

$K_1 = \{R_1\}$ $\quad C_1 = (185, 72)$

$K_2 = \{R_2, R_3\}$ $\quad C_2 = \left(\frac{170+168}{2}, \frac{60+56}{2}\right)$

$= (169, 58)$

H	W
179	68

R4's-

ED of R4

k_1 $= \sqrt{(179-185)^2 + (68-72)^2}$

$= 6.32$

k_2 $= \sqrt{(179-169)^2 + (68-58)^2}$

$= 14.14$

$K_1 = \{R_1, R_4\} \rightarrow C_1 = \left(\frac{185+179}{2}, \frac{72+68}{2}\right)$

$K_2 = \{R_2, R_3\}$ $\quad C_1 = (182, 70)$

$C_2 = (169, 58)$

R_5 :-

$K_1 = \{R_1, R_4, R_5, R_6, R_7, R_8, R_9, R_{10}\}$

$K_2 = \{R_2, R_3, \quad\}$

	R_5	R_6	R_7	R_8	R_9
K_1	(182, 71)	(185, 74)	(182.5, 72.5)	(181.25, 71.25)	(182.125, 77.625)
K_2	(169, 58)	(169, 58)	(169, 58)	(169, 58)	(169, 58)

	R_{10}	R_{11}	R_{12}
K_1			
K_2			

The final cluster data points are

$K_1 = \{R_1, R_4, R_5, R_6, R_7, R_8, R_9, R_{10}, R_{11}, R_{12}\}$

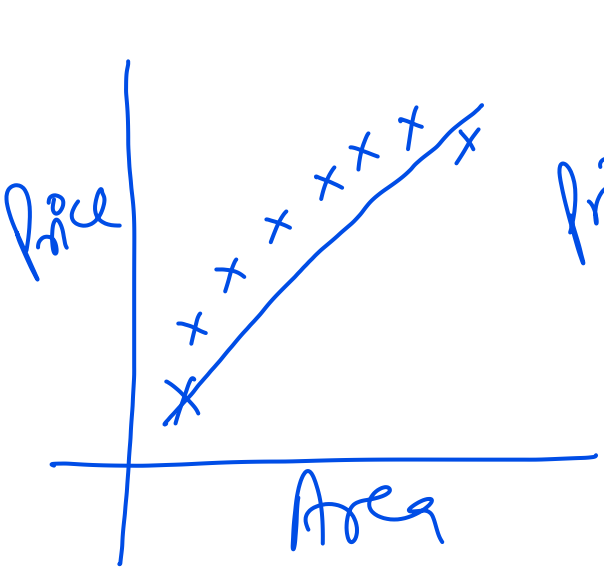
$K_2 = \{R_2, R_3\}$

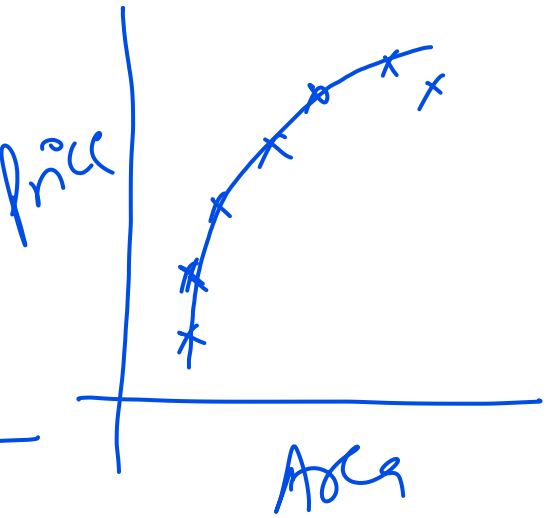
eg

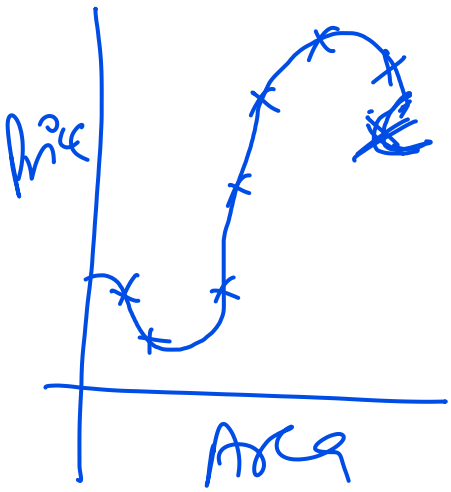
x	y
1	1
1.5	2
3	4
5	7
3.5	5
4.5	5
3.5	4.5

two cluster using k-means and Euclidean distance?

Regularization







$h = \theta_0 + \theta_1 x$ (underfitting)

$\theta_0 + \theta_1 x + \theta_2 x_2^2$ (Best fit)

$\theta_0 + \theta_1 x_1 + \theta_2 x_2^2 + \theta_3 x_3^3 + \theta_4 x_4^4$ (overfitted)

To overcome the overfitted model

1. Reduce the no. of feature manually.
2. use Regularization.
   i) keep all feature but reduce the magnitude of parameter $\theta_0, \theta_1, \theta_2, \ldots$

(ii) if features are very high. We consider all significant features.