Clustering

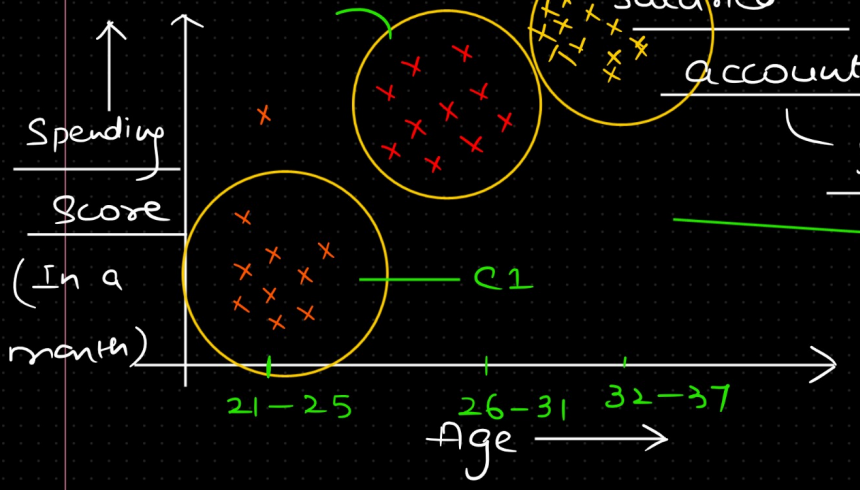
↳ unsupervised machine learning

↳ labelled data not available

HDFC Bank

Salaried account

Bengaluru



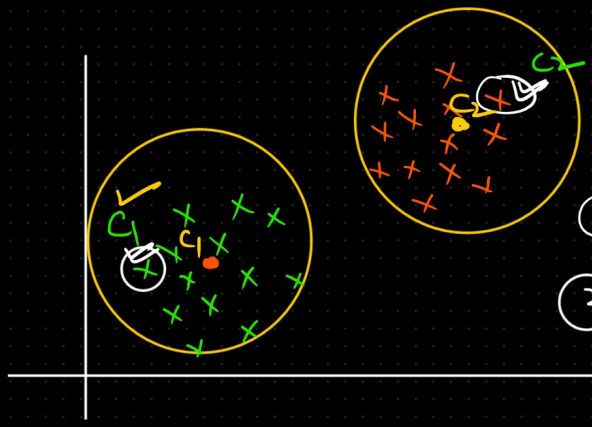
→ Customer segmentation

Clusters

K-Means Clustering

↳ Elbow Method

↳ Optimal value of k (num of clusters)



① Euclidean

② Manhattan

Euclidean Distance $K = 2$

C1 (Distance of all data points from cluster C1 & C2)

C2

$P_1(x_1, y_1)$ $P_2(x_2, y_2)$

$$\sqrt{(y_2-y_1)^2+(x_2-x_1)^2}$$

Manhattan Distance

P_1 (x_1, y_1) P_2 (x_2, y_2)

$$|y_2-y_1| + |x_2-x_1|$$

P1(2,10) P2(2,5) P3(8,4) P4(5,8) P5(7,5)

C1 C2 C3

P6(6,4) P7(1,2) P8(4,9)

(new value)

$C_2 \rightarrow$

k = 3 → P1(2,10) — C1

P4(5,8) — C2

P7(1,2) — C3

num of Clusters

P1(2,10)

P3, P4, P5, P6, P8

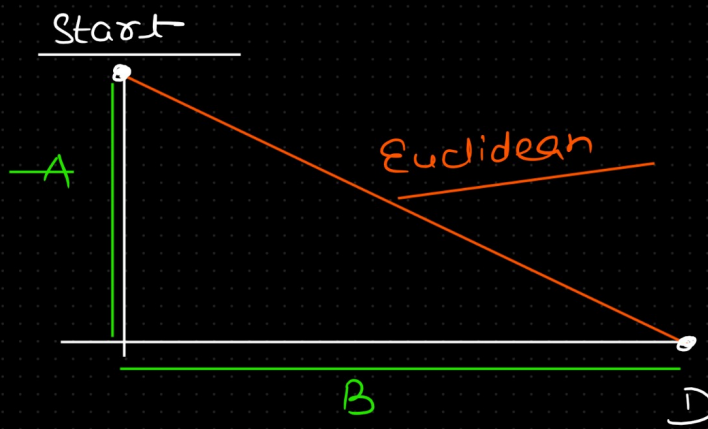
P2, P7

$x = \frac{8+5+7+6+4}{5}$

$y = \frac{4+8+5+4+9}{5}$

$x = \frac{2+1}{2}$

$y = \frac{5+2}{2}$



A + B = manhattan Distance

Distance Euclidean

$(C_1, P_1) = 0$	$(C_2, P_1) = 5$	$(C_3, P_1) = 10$
$(C_1, P_2) = 10$	$(C_2, P_2) = 7$	$(C_3, P_2) = 5$
(C_1, P_3)		
(C_1, P_4)		
⋮	⋮	⋮
(C_1, P_8)	(C_2, P_8)	(C_3, P_8)

Stopping criteria

↳ in the next iteration, new values of (c_1, c_2, c_3) if almost similar which we had in the previous iteration.

form a stable cluster

default

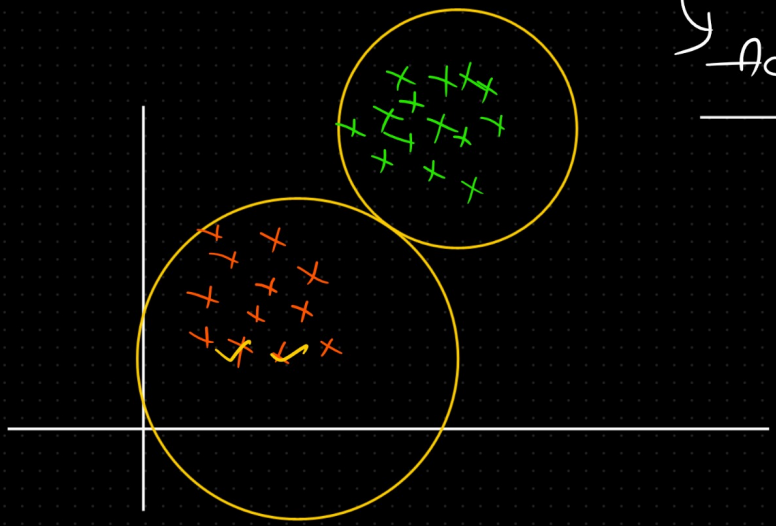
k-means++

↳ Researches

↳ Advance k-means

↳ centroids

↳ at far distances

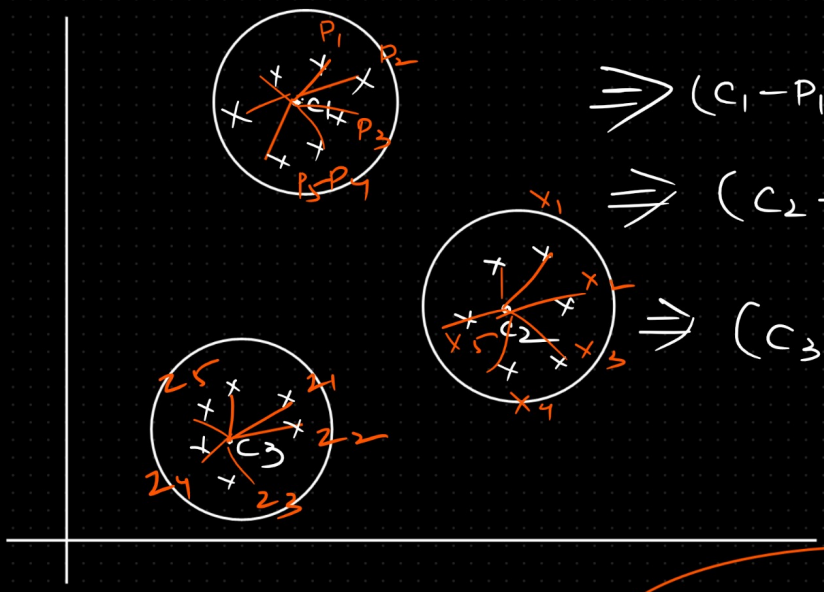


$K = 1$ — WCSS = max — Optimal value of K

↳ WCSS → Within cluster Summation of squared

data points

$K = n$ — WCSS = 0



$$\Rightarrow (c_1 - P_1)^2 + (c_1 - P_2)^2 + \dots (c_1 - P_3)^2$$

$$+$$

$$\Rightarrow (c_2 - Y_1)^2 + (c_2 - X_2)^2 + \dots + (c_2 - X_3)^2$$

$$+$$

$$\Rightarrow (c_3 - z_1)^2 + (c_3 - z_2)^2 + \dots +$$

$$(c_3 - z_5)^2$$

Elbow Method