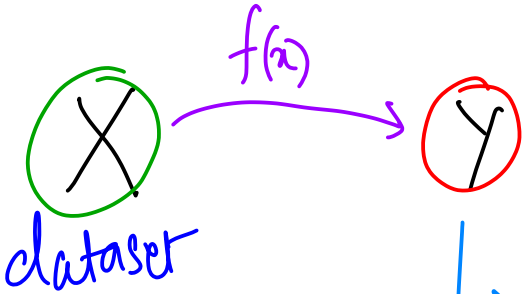
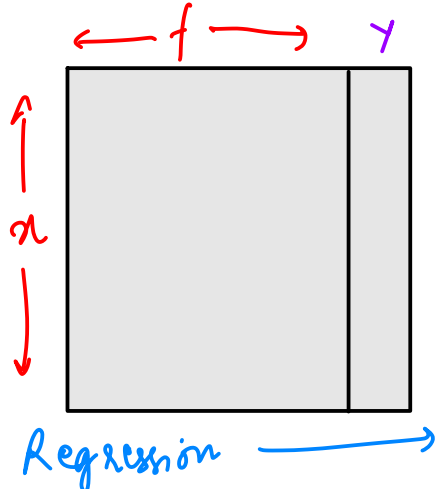
K MEANS

1. Unsupervised learning
2. Higher dimensional analyse/visualisation
3. GMM
4. Anomaly detection



↳ Discrete value [Classification]

↳ Real Continuous value [Regression]

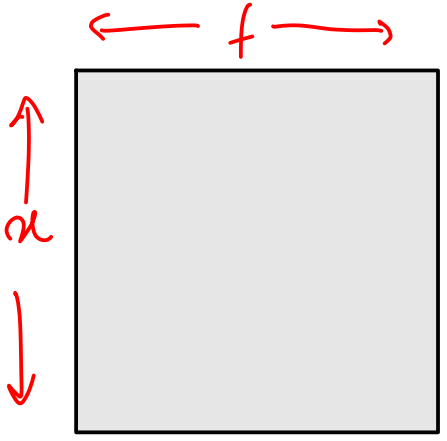


Supervised

→ Classification

$\mathcal{D} = [(x_i, y_i), x_i \in \mathbb{R}^d, y_i \in \{0, 1\}]$

$\mathcal{D} = [(x_i, y_i), x_i \in \mathbb{R}^d, y_i \in \mathbb{R}^1]$

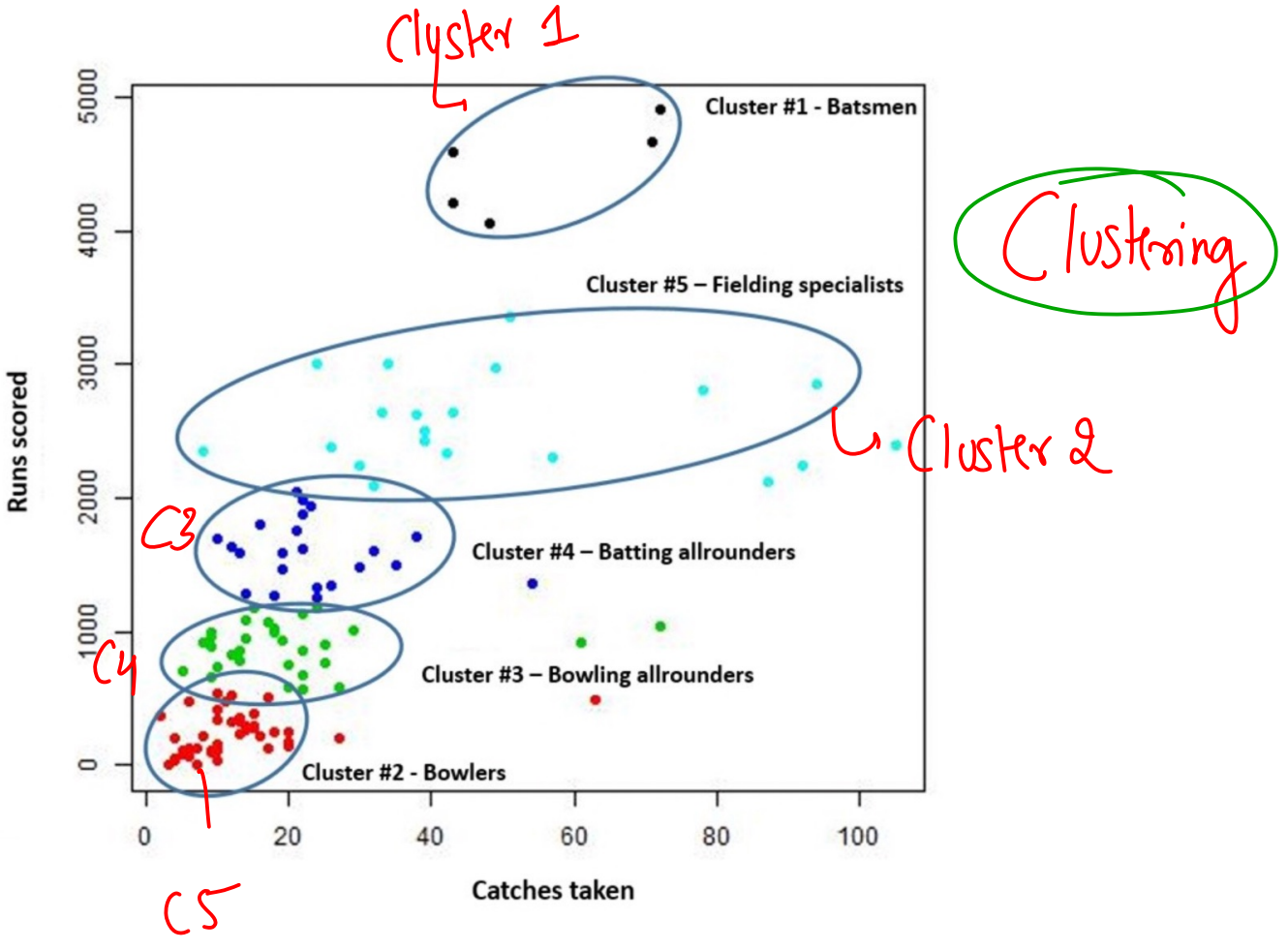


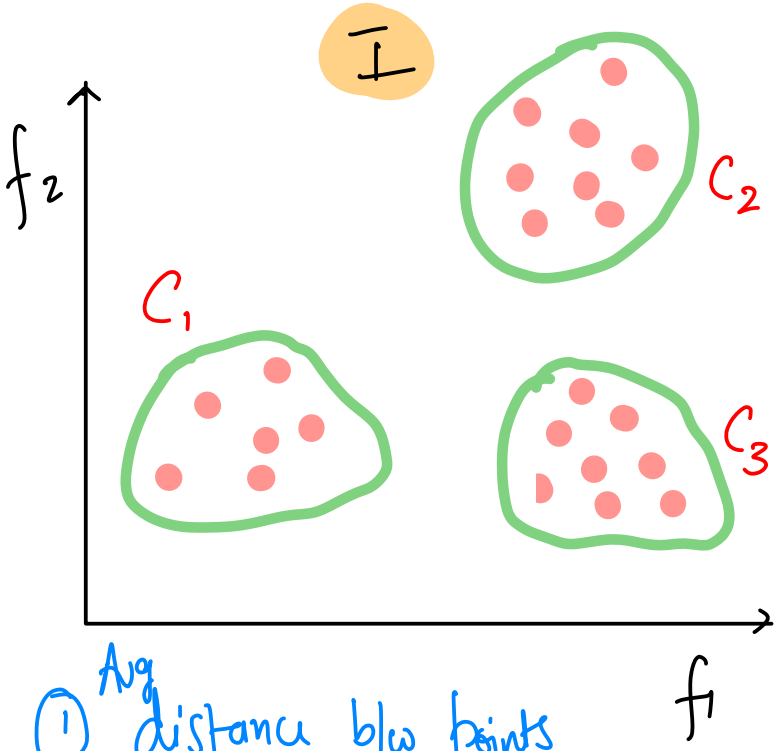
Unsupervised

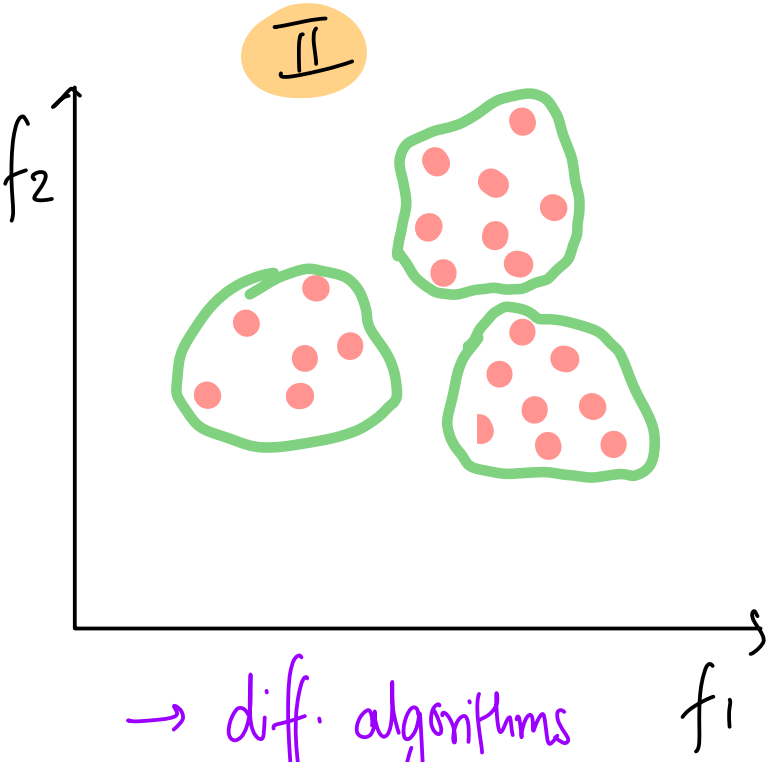
Amazon wants to segment their customers so that they can provide relevant and similar items to their customers Which algorithm do you think can be used here?

8 users have participated

	A	Regresssion	0%
	B	Classification	13%
✓	C	Clustering	87%
	D	None of the above	0%


Cluster 1
Cluster #1 - Batsmen
Clustering
Cluster #5 – Fielding specialists
Cluster 2
C3
Cluster #4 – Batting allrounders
C4
Cluster #3 – Bowling allrounders
Cluster #2 - Bowlers
C5
Runs scored
Catches taken
0
1000
2000
3000
4000
5000
0
20
40
60
80
100





① Avg distance b/w points within grps should be less.

② Avg distance b/w points in one group to another grp should be more.

→ diff. algorithms

→ Good or Bad.

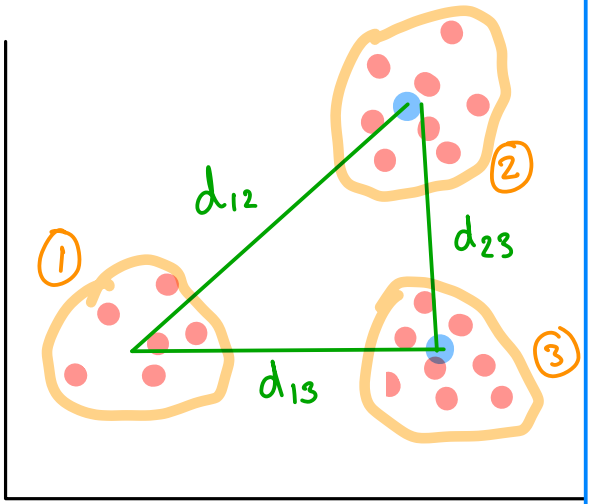
"Cluster output should make biz sense."

$C_1, C_2, f_1 f_2$

1. Inter cluster distance — b/w different clusters
2. Intra cluster distance — within Cluster

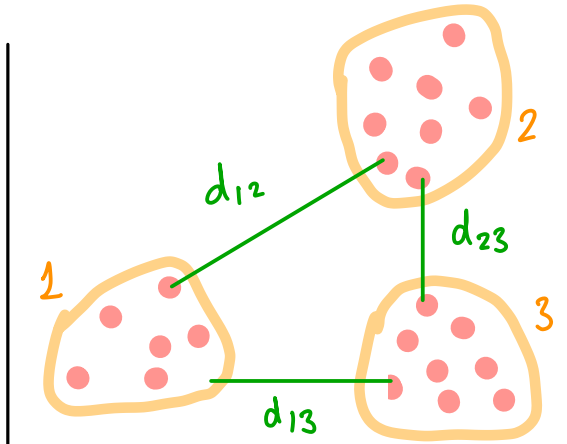
Inter cluster distance

I



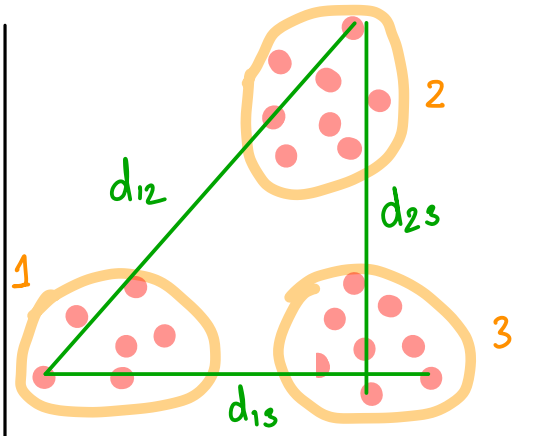
Distances b/w average values of cluster point.

II



Distance b/w closest point from the cluster.

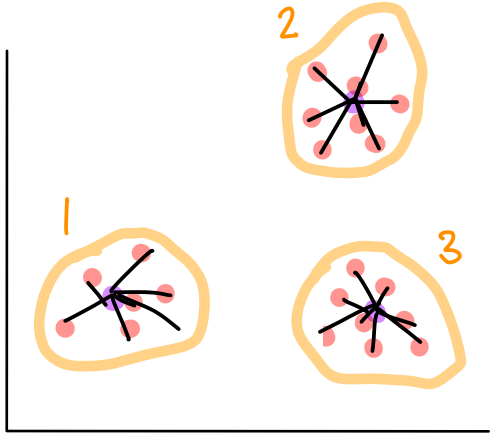
III



Distance b/w farthest point from cluster.

Intra Cluster Distances

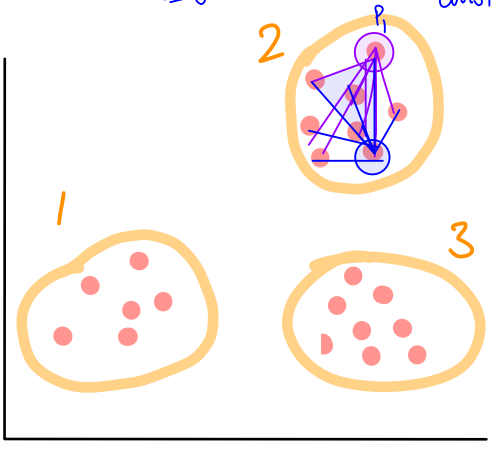
$\frac{1}{n}\sum d_i$



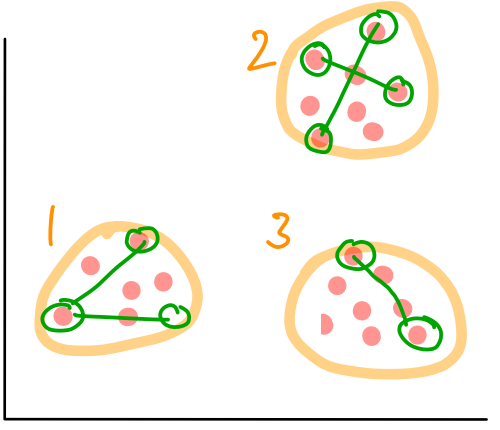
Avg distance w.r.t some central point

Avg

$P_1 \rightarrow$ Every other point in cluster



Avg distance b/w all pairs of points



distance b/w farthest points within cluster.

$$\mathcal{D} = \{ (x_i)_{i=1}^{n} \cdot x_i \in \mathbb{R}^d \}$$

inter cluster distance (↑ high)

intra cluster distance (↓ low)

Ways of finding distances ??

* Euclidean ⇒ low-dims
* Manhattan ⇒ low-medium dims
* Cosine ⇒ high dims.

text

If only one Inter or Intra distance is given, Can we judge how good or bad the clusters are?

It will not work.

We need both.

What metric can be used?

* Biz Case.

Raymond. → S, M, L, XL, XXL

would you create 100 Clusters?

Metric ⇒ DUNN-Index ⇒ metric used to evaluate clustering algorithm.

↳ OBJECTIVE: identify cluster with small variance b/w cluster members and are well separated.

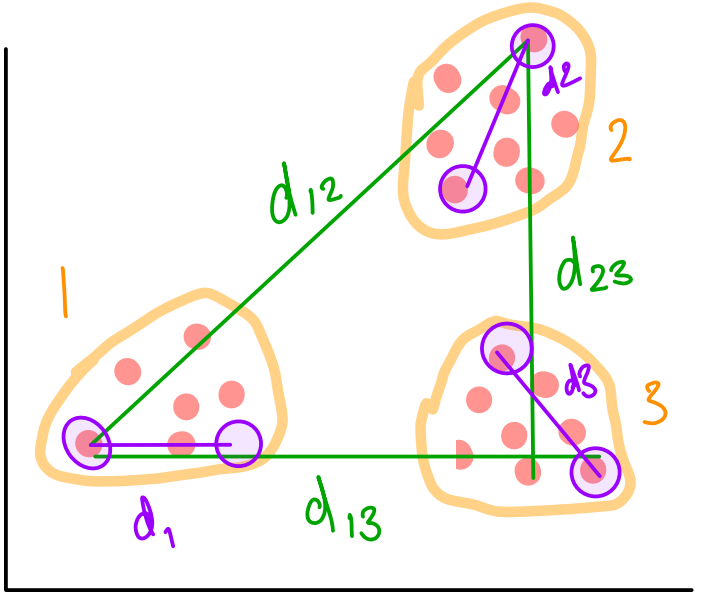
$$D = \frac{\min_{i,j} \text{distance}(i,j)}{\max_{k} \text{distance}(k)}$$

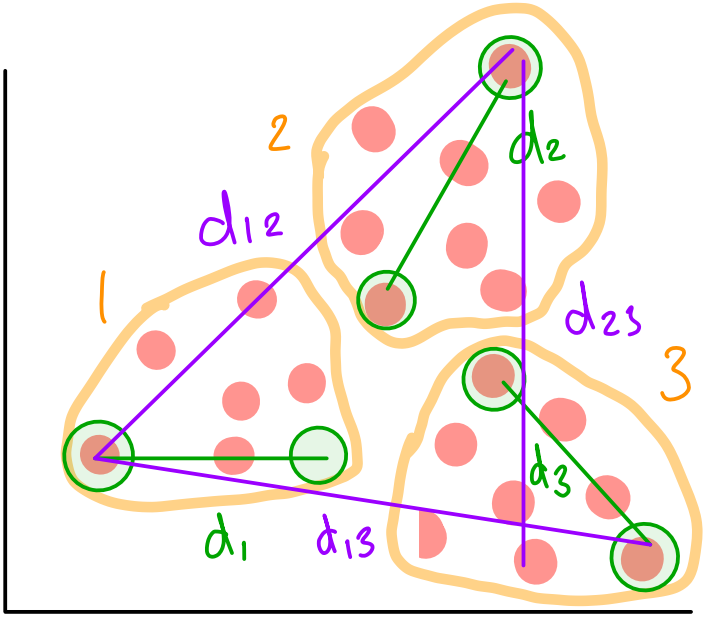
$D(i,j)$ => Inter cluster distance

↳ Distance b/w farthest point of the cluster C_1 & C_2.

$D(k)$ => Intra cluster distance

↳ Distance b/w farthest points within cluster k^{th}





$d(i,j) \mid \min(d_{12}, d_{13}, d_{23})$

$d'(k) \mid \max(d_1, d_2, d_3)$

$d(k) \mid \max(d_1, d_2, d_3)$

$d(i,j) \mid \min(d_{12}, d_{13}, d_{23})$

k clusters

Dunn Index = D = $\frac{\uparrow \min_{i,j} d(i,j)}{\downarrow \max_{k} d(k)}$

$\uparrow \min_{i,j} d(i,j)$ ⇒ Inter cluster

$\downarrow \max_{k} d(k)$ ⇒ Intra cluster.

$D_1 > D_2$

D_1 good

$\frac{\text{Inter } 10}{\text{Intra } 5}$ (1) > $\frac{\text{Inter } 10}{\text{Intra } 10}$ (2)

In the Dunn Index formula, what does "distance(i, j)" represent?

8 users have participated

A	The average distance between all points in cluster i and all points in cluster j.	13%
B	The distance between the centroids of cluster i and cluster j.	0%
C	The distance between the closest points in cluster i and cluster j.	12%
✓ D	The distance between the farthest points in cluster i and cluster j.	75%

We have two clustering algorithms, and if we have to choose one of them, the algorithms chosen should have:

8 users have participated

✓ A	higher dunn-index	100%
B	lower dunn-index	0%
C	Dunn-index doesn't matter	0%