Introduction

Unsupervised ML

- Kmeans Clustering
- Hierarchical Clustering
- Gaussian Mixture Models
- Outlier/Novelty detection Algos
- PCA / T-SNE / U-map

Topics

- Intro to Unsupervised ML
- Case Study: Customer Segmentation
- Clustering
- Dunn Index
- K-means Intro
- Mathematical formulation of K-means
- Lloyd's Algo
- Implementation of LLoyd's Algo
- Determining K
- Home-work

Intro to Unsupervised ML

- ML
  - Supervised
    - Classification
    - Regression
  - Unsupervised
  - Reinforcement

Classification $\Rightarrow \{\underbrace{x_i}_{\text{Features}}, \underbrace{y_i}_{\text{Labels}} \; ; \; x_i \in \mathbb{R}^d$

$y_i \in \{0, 1, 2 \ldots n\}\}$

Binary $\Rightarrow$ $y_i \in \{0, 1\}$

Regression $\Rightarrow \{\underbrace{x_i}_{\text{Features}}, \underbrace{y_i}_{\text{Labels}} \; ; \; x_i \in \mathbb{R}^d$

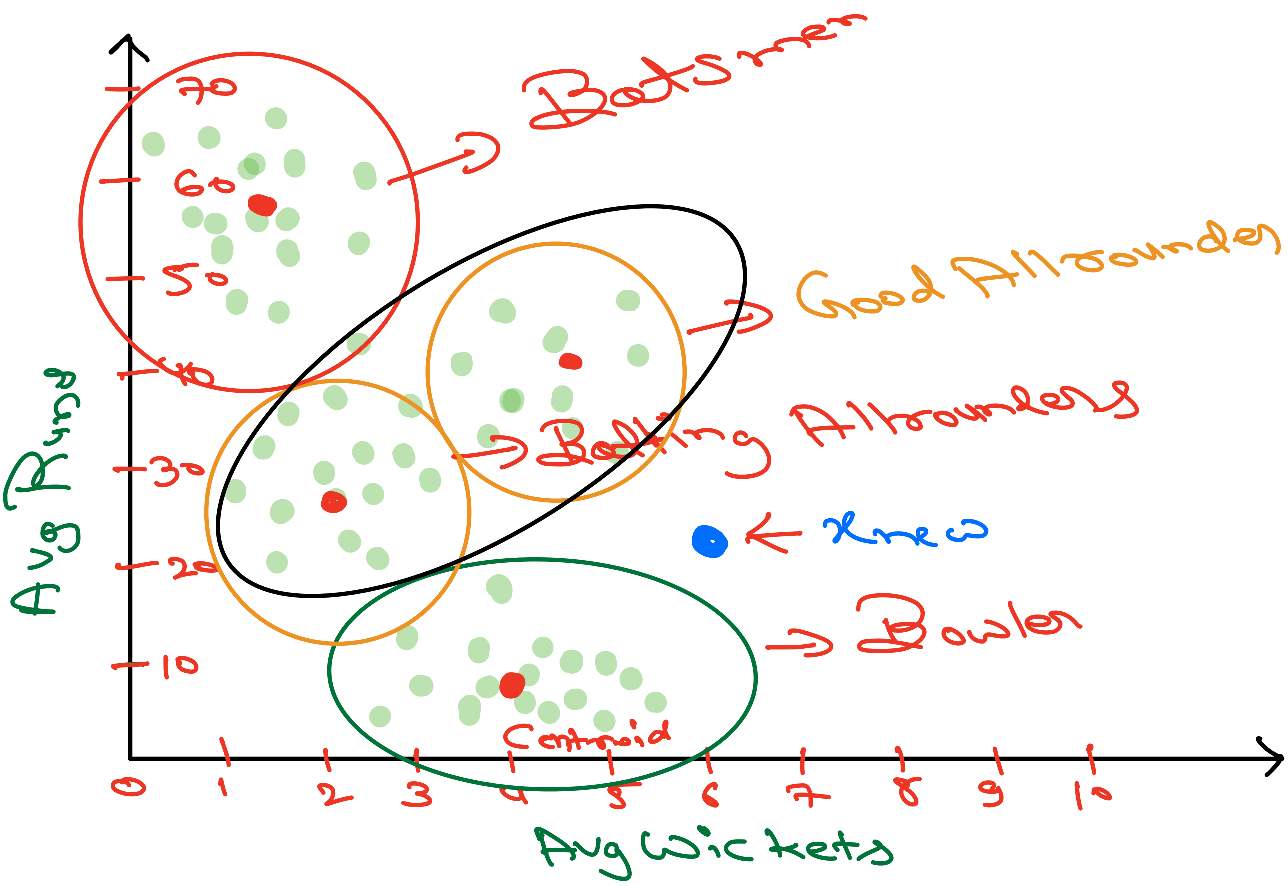
$y_i \in \mathbb{R}$

* Unsupervised $\Rightarrow$

$D \Rightarrow \{x_i \; ; \; x_i \in \mathbb{R}^d\}$

⇒ Create Meaningful Groups of Customers who have Similar Behaviour ⇒ Create Clusters

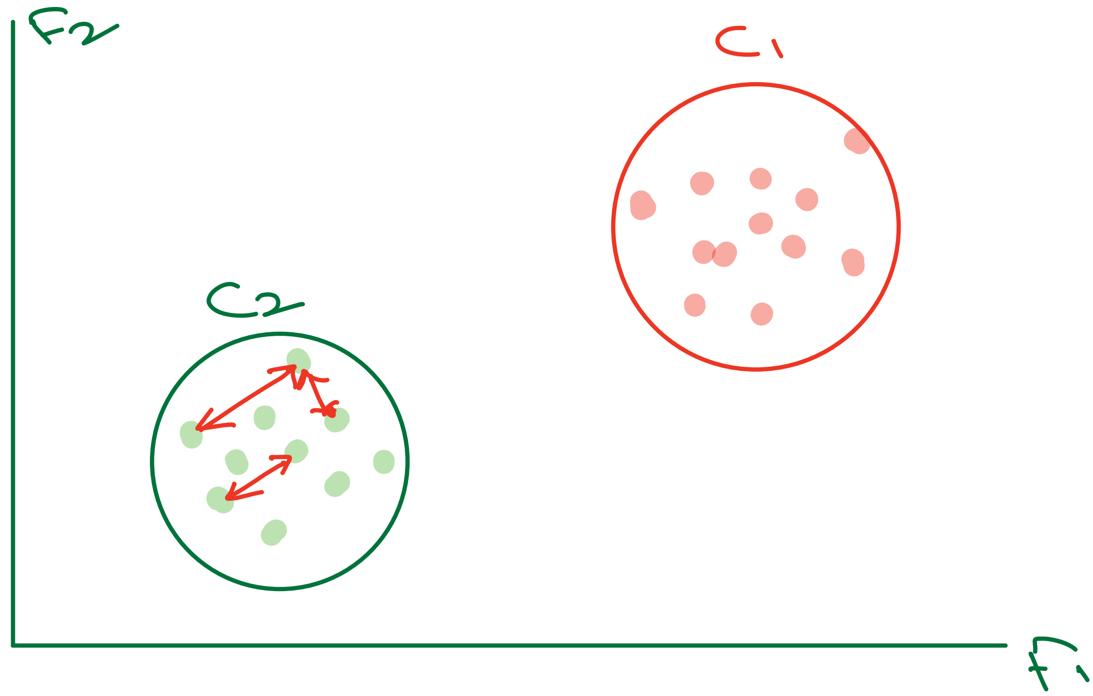
$\boxed{\text{Clustering}}$



* n-cluster ⇒ Hyperparameter
* Every point will belong to one of the clusters
* Data points in a Cluster are Close Each Other

Evaluating Clusters

① No Ground Truth, hence we can't Classification



* Intra Cluster Distance

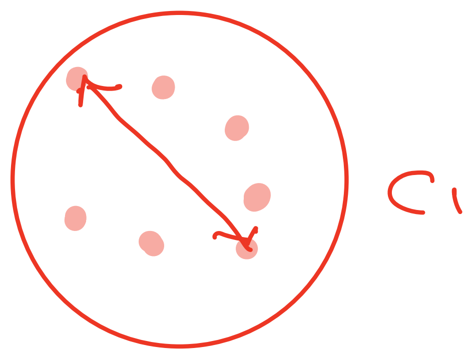
① Avg distance among All pairs

or

② Distance of Farthest pair

or

③ Closest points Distance



Q: Intra-Cluster : Minimum

* Inter Cluster Distance

Maximum

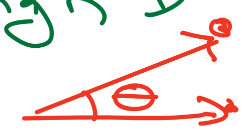
* Avg of All pairs
* Distance among farthest pair
* Distance among Closest pair

Summary

Intra Cluster Distance Minimum

Inter Cluster Distance Maximum

* Methods to Calculate Distance
  1) Euclidean Distance → Low D
  2) Manhattan Distance → Moderate D
  3) Cosine Similarity → High D



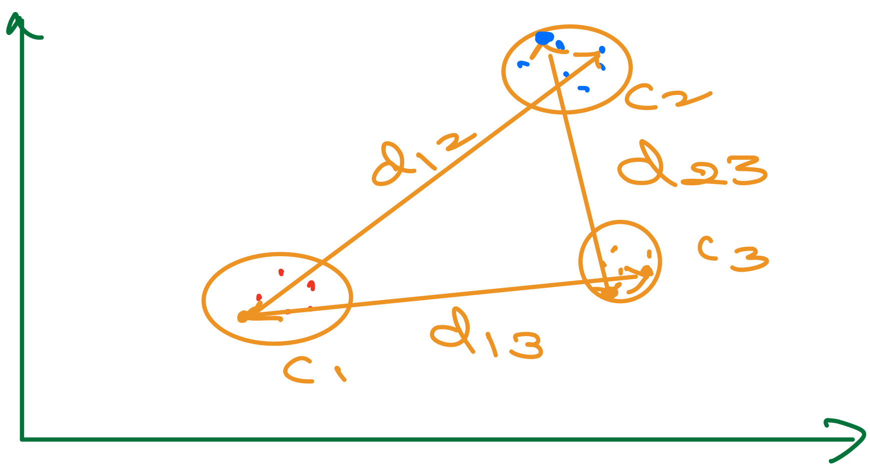
Dunn Index

$$\boxed{D_i \Rightarrow \frac{\min_{i,j} \text{distance}(i,j)}{\max_k \text{distance}'(k)}}$$

* distance(i, j) : Inter-Cluster Distanc

  Farthest point in cluster i and Cluster j

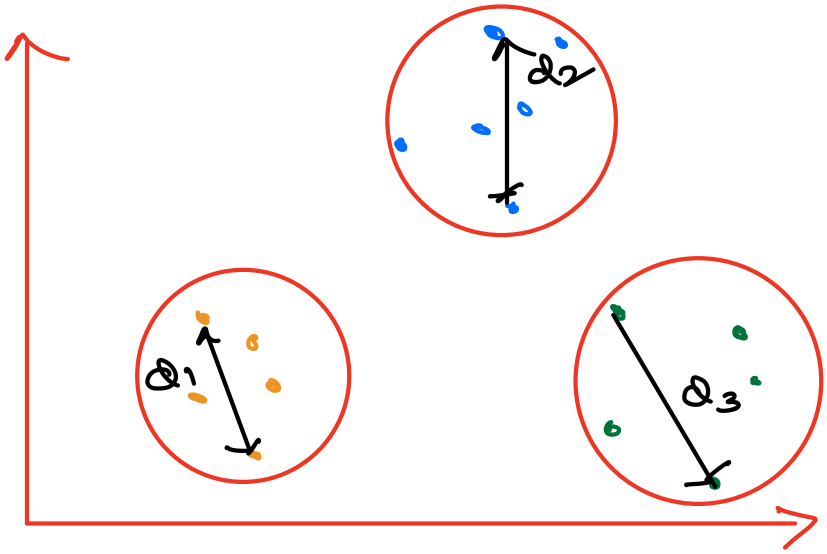
* distance'k : Intra Cluster Distance farthest pair



d(i, j) ⇒ d_{13}, d_{12}, d_{23}

min

d_{23} ← numerator



(d_1, d_2, d_3)

$\downarrow$

maximum

$\downarrow$

d_3

Dunn-index $\Rightarrow$ $\dfrac{d_{23} \leftarrow \text{High}}{d_3 \leftarrow \text{Low}}$

$\Rightarrow$ Higher Value of Dunn-index is Better

$\Rightarrow$ Range $\Rightarrow$ $[0, \infty)$

$\boxed{\text{K-means}}$

↓

n-clusters (No of Clusters)

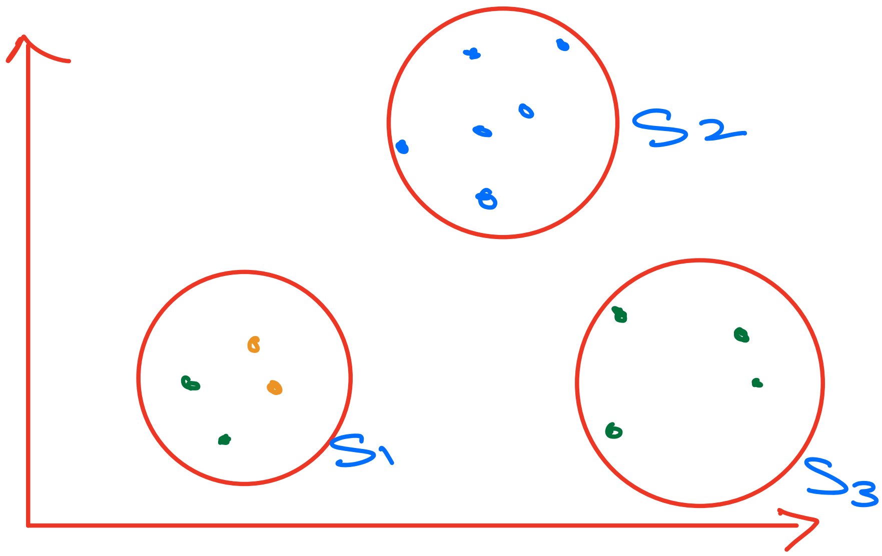
↓

Similar to K-nn

$x_i \in \mathbb{R}^d$

n inputs

d features



$S_n \Rightarrow D$

$S_1 \cup S_2 \cup S_3 \Rightarrow D \Rightarrow S_n$

$S_1 \cap S_2 \cap S_3 \Rightarrow \phi$

$S_1 \cap S_2 \Rightarrow \phi \Rightarrow S_2 \cap S_3$

* K-means is a Centroid Based algorithm where

$$C_1 \Rightarrow \frac{x_1 + x_2 + x_3 -- x_n \in S_1}{len(S_1)}$$

$$C_1 \Rightarrow \frac{x_1 + x_2 + x_3 + x}{len(S_1)} \quad x_i,$$

$$\frac{y_1 + y_2 + y_3 + y_4}{len(S_1)} \quad y_i$$

$$C_1 \Rightarrow x_i, y_i$$

$$\boxed{C_i \Rightarrow \frac{1}{|S_i|} \sum_{x_j \in S_i} x_j}$$

jth vector belonging to ith Set

Size / Cardinality of ith set

Objective Function

$C_1, C_2, C_3 \ldots C_n$

① max (Inter Cluster Distance)

and

② min (Intra Cluster Distance)

Difficult to solve with Traditional Optimization

* Approximate Algo:

$\boxed{\text{LLoyd's Algorithm}}$

→ Simple and Easy to Calculate

→ Doesn't Guarantee the Best Solution