K- Medoid Clustering

A variation of K-Means clustering.

In K-Means, the mean of the values is taken. This mean may not be a real point.

However, in K- Mediaoid, a real point from the dataset is taken instead of the mean value.

A mediaoid is the most centrally placed point in the cluster.

This makes the algorithm less sensitive to outliers than compared to K-Means.

K Means minimizes the sum of square distances while K Mediaoid minimizes the sum of dissimilarities between points & their cluster centers.

Converges in fixed no of i but; slower than K-Means

K-Means	K- Medoid
$O(\underset{\text{Iterations}}{i}\ \underset{\text{clusters}}{k}\ \underset{\text{datapoints}}{n})$	$O(i\,k\,(n-k)^2)$

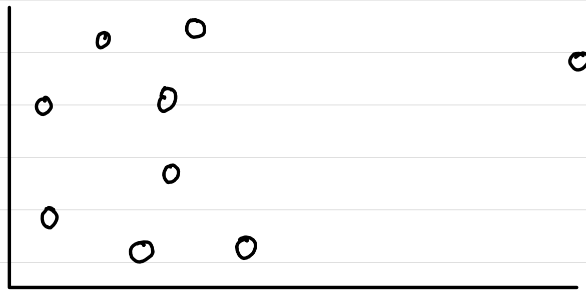
3- Main variations of algorithms

K-Medoid → PAM, CLARA, CLARANS

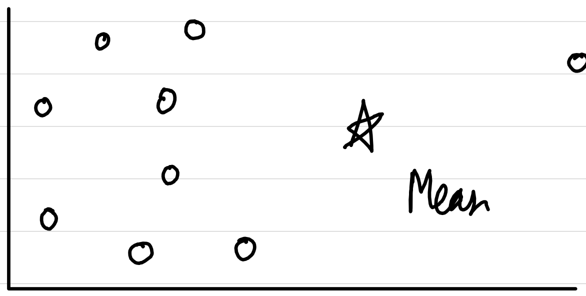
Advantages → Less sensetive to outliers & noise
Non spherical clusters

Disadvantages → High Time complexity
Sensativity to initial medoids (like k-Means)

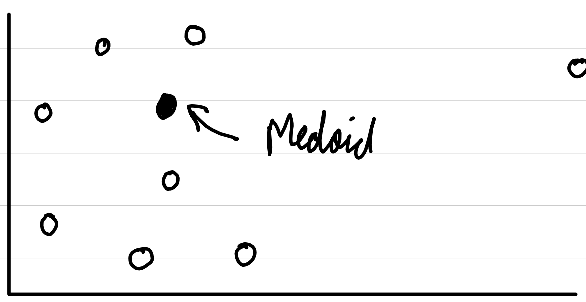
Consider data points



Mean is affected by outlier



Medoids are more robust to outliers



PAM algorithm
Partitioning Around Medoids

1. Choose k random points from dataset as initial medoids

2. Assign clusters based on distance

3. Calculate total cost (sum of all distances of each datapoint from its centroid)

4. Select a non medoid & swap with previous medioid & calculate cost

5. Repeat for all points, choose Medoid with lowest cost as new medoid.

6. Update cluster values.

7. Continue the repetitions till no changes in medioid take place.

Time complexity per iteration → $O(k(n-k)^2)$

$$k \times (n-k) \times (n-k)$$

for every Medoid

for all non mediord swaps

dist. of Non medoid datapoints from the medoid

from perspective of data size, $O(n^2)$ hence too large time for large datasets.

Update medioid is most expensive part as every point needs to be treated as potential medioid

CLARA

clustering Large Applications algorithm

Uses the divide & rule approach.

Breaks down the large dataset into smaller parts that can be clustered by PAM

1. Sampling $\rightarrow$ Dataset is sampled S times into S samples $S < n$, $S^2 << n^2$
These are drawn at random & can overlap

2. Run PAM on Sample 1 and find Mediods

3. Cluster dataset based on sample 1 Mediods

4. Calculate the dissimilarity of the dataset

5. Repeat 2-4 for sample 1 - S

6. Choose Mediods with lowest dissimilarity

This enables handling of large datasets

Just like ensemble learning, multiple samples are taken, & multiple models are built.

However, unlike ensemble learning, voting is not done. Only one best set of medoids is chosen and used for evaluation

CLARANS

Clustering Large Applications based on Randomized Search

Instead of treating every point as possible medoid, consider only the neighbours of the current medoid

Consider predetermined number of neighbouring datapoints as potential medoids.

More time efficient.

Bandit PAM (Tiwari & Zang)

Recent research has reduced time complexity from $O(n^2)$ to $O(n \log n)$

Uses technique inspired from multi armed bandits

Vornoi iteration

In each cluster, consider only points within that cluster as potential medoids.

Does not allow re-assigning points to other clusters.

Doesnt give good results.