Clustering Algorithms: k-Means and k-Medoids

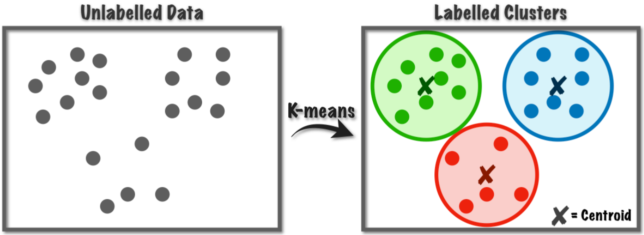
Gandholi Sarat

Sri Sathya Sai Institute of Higher Learning

March 31, 2025

k-Means

- k-Means is a widely used clustering algorithm due to its efficiency.
- Time complexity: $O(Nmq)$, where q is the number of iterations, and m is number of clusters.
- Suitable for large datasets.



Advantages of k-Means

- Fast and computationally efficient.
- Simple to implement and interpret.
- Can be extended for different clustering problems.

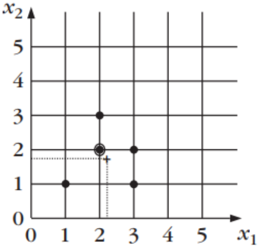
Drawbacks of k-Means

- Sensitive to outliers and noise.
- Struggles with non-spherical clusters.
- Generally applicable to data sets with continuous valued feature vectors

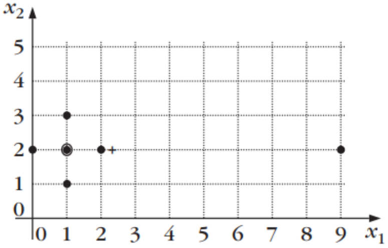
Introduction to k-Medoids

- In the k-medoids methods,each cluster is represented by a vector selected among the elements of X, and we will refer to it as the **medoid**
- More robust to **outliers**.
- Works for both continuous and discrete datasets.

K-mean Vs K-Medoids



(a)



(b)

Partitioning Around Medoids (PAM)

- Used to determine the set of the m medoids that best represent the data set
- Iteratively swaps medoids with non-medoids to minimize cost.
- Time Complexity: $O(m(N-m)^2)$.

PAM Algorithm: Optimization Approach

- PAM minimizes $J(\mathcal{M}, U)$, where $\mathcal{M}$ is the set of medoids.
- Constraints: Medoids are actual elements from dataset X.
- Two sets of medoids $\mathcal{M}$ and $\mathcal{M}'$ are **neighbors** if they share $m-1$ elements.
- A neighbor $\mathcal{M}_{ij}$ results from replacing x_i with x_j.

PAM Algorithm: Iterative Improvement

- Start with a random set $\mathcal{M}$ of m medoids.
- For each neighbor $\mathcal{M}_{ij}$, compute:

$$\Delta J_{ij} = J(\mathcal{M}_{ij}, U_{ij}) - J(\mathcal{M}, U)$$

- If $\Delta J_{qr} = \min(\Delta J_{ij}) < 0$, replace $\mathcal{M}$ with $\mathcal{M}_{qr}$.
- Repeat until no further improvement.

Computation of ΔJ_{ij}

- ΔJ_{ij} is computed by summing individual point contributions:

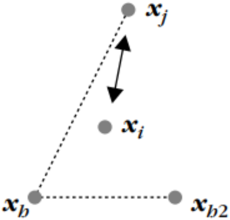
$$\Delta J_{ij} = \sum_{h \in X \setminus \mathcal{M}} C_{hij}$$

- Four cases determine C_{hij}, depending on:
    - Whether x_h belongs to the cluster of x_i.
    - Whether x_j is closer than the second nearest medoid.

Case 1: Retains Second Closest Medoid

- x_h belongs to the cluster of x_i.
- After replacing x_i with x_j, x_h is now represented by the second closest medoid x_{h2}.
- The cost change is:

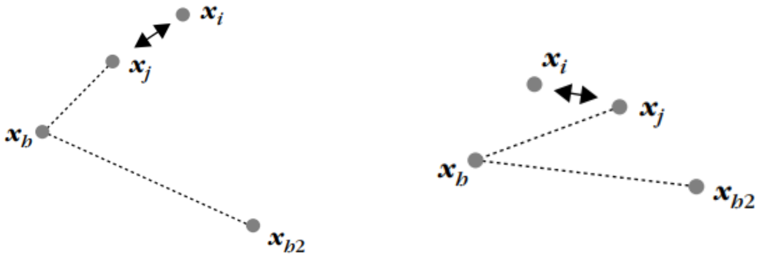
$$C_{hij} = d(x_h, x_{h2}) - d(x_h, x_i) \geq 0$$



Case 2: Switches to New Medoid

- x_h was initially assigned to x_i.
- After replacing x_i with x_j, x_h now moves to x_j.
- The cost change is:

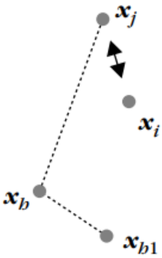
$$C_{hij} = d(x_h, x_j) - d(x_h, x_i)$$



Case 3: Remains in the Same Cluster

- x_h is not assigned to x_i, and the replacement does not affect its assignment.
- Thus, there is no change in cost:

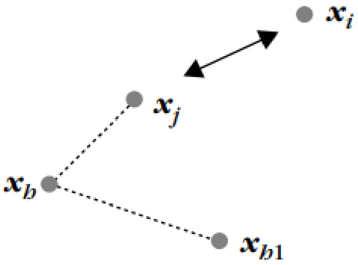
$$C_{hij} = 0$$



Case 4: Moves from a Different Medoid

- x_h was initially assigned to a different medoid x_{h1}.
- After the replacement, x_h is now assigned to x_j.
- The cost change is:

$$C_{hij} = d(x_h, x_j) - d(x_h, x_{h1}) \geq 0$$



CLARA and CLARANS

- **CLARA**: Draw randomly a sample X' of size N' from the entire data set, X and to determine the set of the medoids M' that best represents X' using the PAM algorithm
- **CLARANS**:PAM is applied on the entire data set X, but with a slight modification. At each iteration, not all neighbors of the current set of medoids are considered. Instead, only a randomly selected fraction

$$q < m(N - m)$$

of them is utilized.

CLARA and CLARANS

- CLARANS is more accurate but computationally expensive.
- In some cases CLARA runs significantly faster than CLARANS. It must be pointed out that CLARANS retains its quadratic computational nature and is thus not appropriate for very large data sets.

Conclusion

- **k-Means**: Simple and efficient but sensitive to initialization.
- **k-Medoids**: More robust but computationally heavier.
- **CLARA and CLARANS**: Trade-off between speed and accuracy.
- **Suggested numbers in CLARA**: Experimental studies suggest that five X' and N' $= 40 + 2$m lead to satisfactory results.
- **Suggested numbers in CLARANS**: Experimental studies suggest that q can be chosen as the maximum between 0.12m (N - m) and 250.