k-Medoids Clustering

Sudipto Ghosh

M.Sc. CS Semester I
Department of Computer Science
University of Delhi

March 12, 2023

Motivation

- In k-means clustering, we calculate the arithmetic cluster means and calculate distance from every other point to the cluster means. The cluster mean does not necessarily correspond to a data point.
- Can we pick some actual data point as representative elements of clusters, and calculate distances from them?

k-Medoids Clustering

- Partitioning is performed based on the principle of minimizing the sum of residuals between each data point and its corresponding representative.
- Absolute Error Criterion is used

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, o_i)$$

where E is the sum of absolute error for all objects p in the dataset and o_i is the representative point of cluster C_i.

Comparison b/w k-Means and k-Medoids

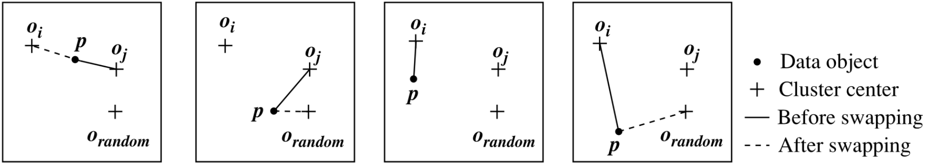
- k-Medoids method is more robust than k-Means in the presence of noise and outliers because a medoid is less influenced by outliers or other extreme values than a mean.
- Complexity of each iteration in the k-Medoid method is $O(k \cdot (n - k)^2)$. For large databases where n and k are very high, such computations become costlier than k-Means.
- Both methods require the user to specify k – the number of clusters.

PAM

- **P**artitioning **A**round **M**edoids.
- Approaches the clustering problem in an iterative, greedy way.
- Like the k-Means algorithm, the initial representatives are chosen arbitrarily.
- Next, we consider whether replacing a representative by a non-representative point would improve the clustering quality.
- All possible replacements are carried out. The process continues until the quality of the resultant clustering cannot be improved by any replacement. Complexity is given by $O(k \cdot (n - k)^2)$.
- PAM works well for small databases but not for larger databases. To deal with larger datasets, sampling-based methods can be used.

PAM

1. Arbitrarily choose k-objects in D as initial representatives
2. Until Convergence
    1. Assign each remaining object to the cluster with the nearest representative.
    2. Randomly select a non-representative object o_{random}.
    3. Compute the total cost S, of swapping representative point o_j with o_{random}.
    4. If $S < 0$ then swap o_j with o_{random} to form the new set of k-representatives.



(a) Reassigned to o_i **(b)** Reassigned to o_{random} **(c)** No change **(d)** Reassigned to o_{random}

CLARA

- **C**lustering **LAR**ge **A**pplications.
- Instead of taking the whole dataset into consideration, CLARA uses a random sample of the dataset.
- The PAM algorithm is then applied to compute the best medoids from the sample.
- CLARA builds clustering from multiple random samples and returns the best clustering as the output.

CLARA

- Complexity of computing the medoids on a random sample is given by $O(ks^2 + k(n-k))$, where s is the sample size, k is the no. of clusters and n is $|D|$.
- Effectiveness of CLARA depends on the sample size.
- If an object is one of the best k-medoids but is not selected during sampling, CLARA will never find the best clustering.

CLARANS

- **C**lustering **L**arge **A**pplications based upon **RAN**domized **S**earch.
- Presents a tradeoff betweem the cost and effectiveness of using random samples to obtain clustering.

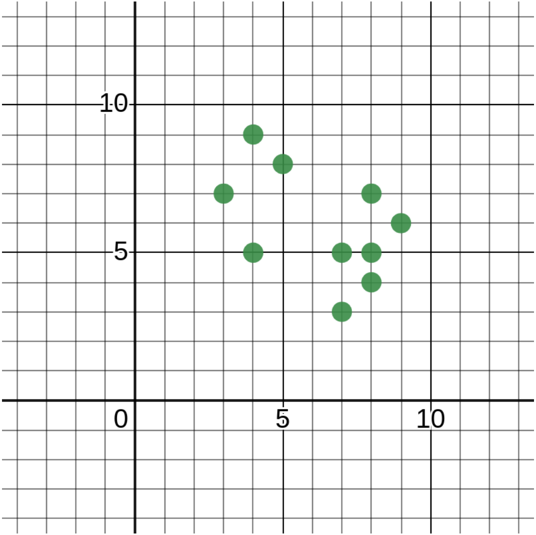
CLARANS

1. Arbitrarily select k objects in the dataset as the initial medoids.
2. Randomly select a current medoid x and an object y that is not one of the current medoids.
3. Replace x by y if it improves the absolute error criterion. Conduct such Randomized Search l times.
4. The set of current medoids after the l steps is considered a local optimum.
5. Repeat this randomized process m-times and return the best local optimal as the final result.

PAM - Dry Run

x	y
8	7
3	7
4	9
9	6
8	5
5	8
7	3
8	4
7	5
4	5

PAM - Dry Run



PAM - Dry Run

x	y
8	7
3	7
4	9
9	6
8	**5**
5	8
7	3
8	4
7	5
4	**5**

PAM - Dry Run

x	y	$d(o, c_1)$	$d(o, c_2)$
8	7	6	2
3	7	3	7
4	9	4	8
9	6	6	2
8	**5**	-	-
5	8	4	6
7	3	5	3
8	4	5	1
7	5	3	1
4	**5**	-	-

$$[(4, 5), (3, 7), (4, 9), (5, 8)]$$
$$[(8, 7), (9, 6), (8, 5), (7, 3), (8, 4), (7, 5)]$$

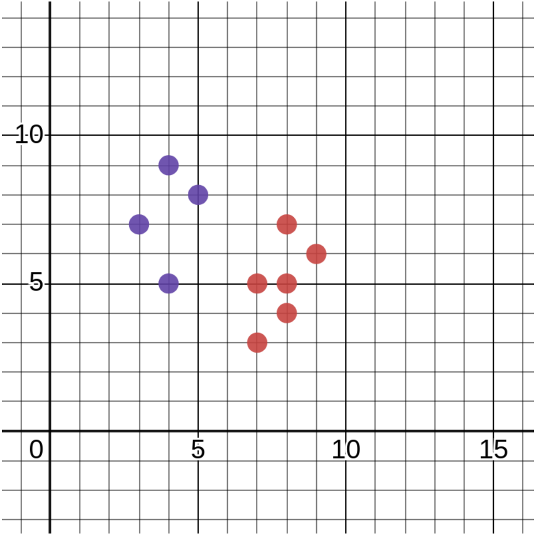
$$\text{Cost} = (3 + 4 + 4) + (3 + 1 + 1 + 2 + 2) = 20$$

PAM - Dry Run

x	y	$d(o, c'_1)$	$d(o, c'_2)$
8	7	6	3
3	7	3	8
4	9	4	9
9	6	6	3
8	5	4	1
5	8	4	7
7	3	5	2
8	**4**	-	-
7	5	3	2
4	**5**	-	-

$$\text{New Cost} = (3 + 4 + 4) + (2 + 2 + 1 + 3 + 3) = 22$$
$$\text{New Cost} > \text{Cost} \implies \text{Undo Swap}$$
$$\therefore (8,\ 5) \text{ and } (4,\ 5) \text{ are the final medoids}$$

PAM - Dry Run



References

[1] Jiawei Han, Micheline Kamber, and Jian Pei. "10 - Cluster Analysis: Basic Concepts and Methods". In: *Data Mining (Third Edition)*. Ed. by Jiawei Han, Micheline Kamber, and Jian Pei. Third Edition. The Morgan Kaufmann Series in Data Management Systems. Boston: Morgan Kaufmann, 2012, pp. 443–495. ISBN: 978-0-12-381479-1. DOI: https://doi.org/10.1016/B978-0-12-381479-1.00010-1. URL: https://www.sciencedirect.com/science/article/pii/B9780123814791000101.