CNAK

The CNAK module provides a K-means based algorithm which does not require pre-declaration of K. It can detect appropriate cluster number within the data and useful for handling big data.

This method can solve a few pertinent issues of clustering a dataset:

- Detection of a single cluster in the absence of any other cluster in a dataset.
- The presence of hierarchy.
- Clustering of a high dimensional dataset.
- Robustness over dataset having cluster imbalance.
- Robustness to noise.

However, This method is not applicable to different shaped dataset.

Please cite the paper below if you make use of the software.

Fundamental observation

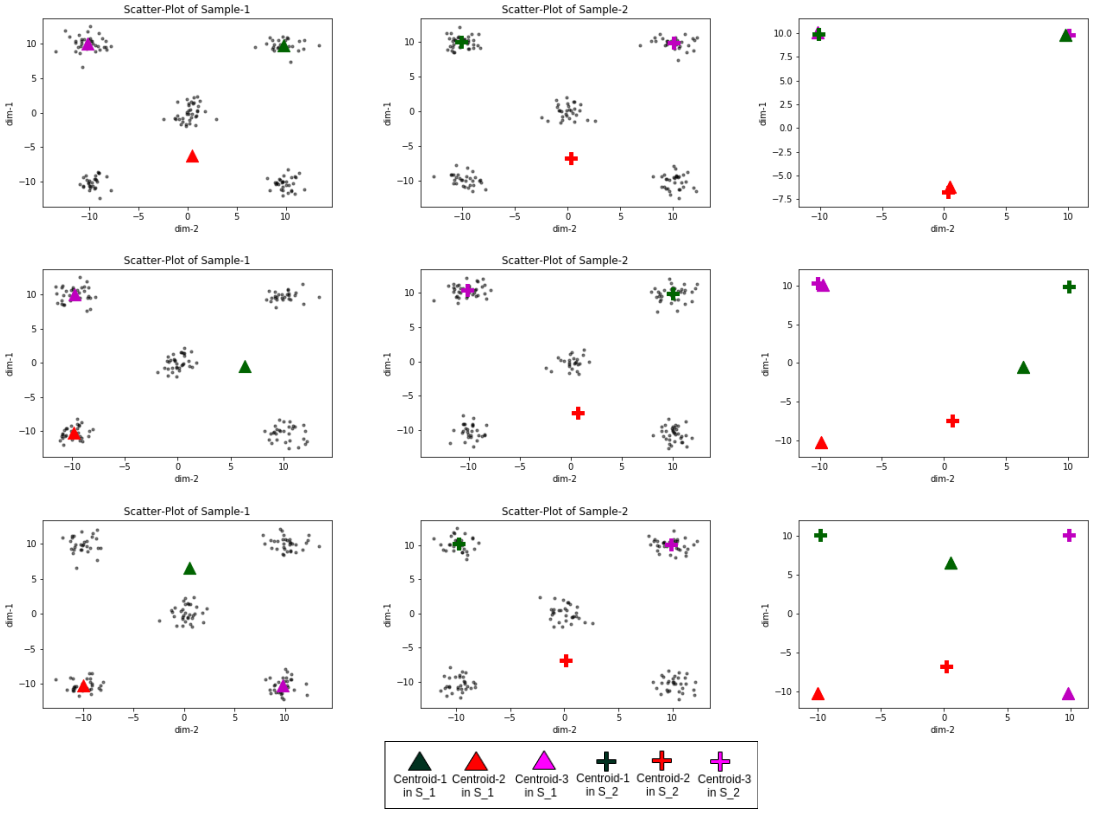

Figure 1: Each row shows the positional behavior of cluster centroids chosen by *K-means++* in a pair of samples H (S_1, S_2). First and second column of each row depicts randomly selected sample distributions with cluster centroids generated by *K-means++* with $K = 3$. Third column depicts a plot where set of centroids generated from S_1 and S_2 are plotted.

Intuition : The choice of K is governed by the positional behavior of any two sets of the centroids.

1. $K < \hat{K}$, $\hat{K}$ is optimal cluster number
2. $K = \hat{K}$, $\hat{K}$ is optimal cluster number
3. $K > \hat{K}$, $\hat{K}$ is optimal cluster number

Experiment : Experiment is conducted on a dataset where 5 clusters are well separated. i.e, $\hat{K} = 5$. We have conducted with K=3 ($K < \hat{K}$), 5($K = \hat{K}$), and 7($K > \hat{K}$).

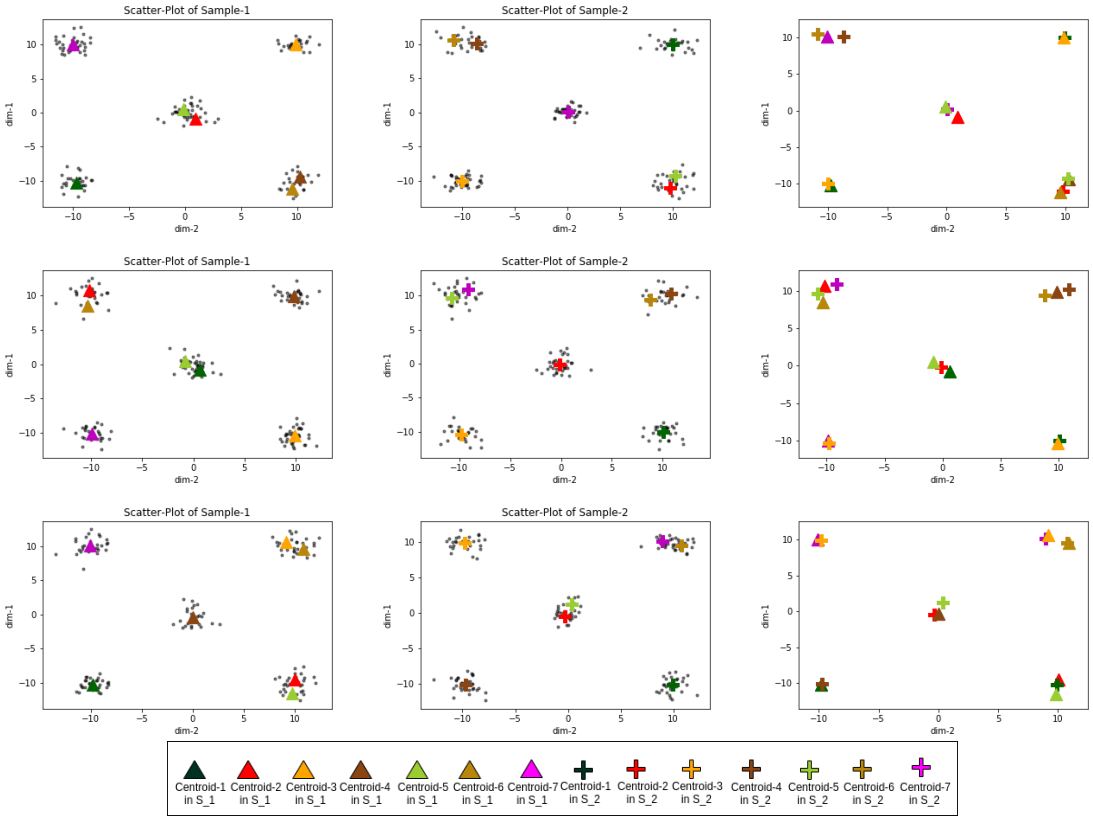

Figure 2: Each row shows the positional behavior of cluster centroids chosen by *K-means++* in a pair of samples H (S_1, S_2). First and second column of each row depicts randomly selected sample distributions with cluster centroids generated by *K-means++* with $K = 7$. Third column depicts a plot where set of centroids generated from S_1 and S_2 are plotted.

Observation :

- $K = \hat{K}$ **:** The centroids of two samples for a given pair are very close to each other for all three cases. However, the situation is different when $K \neq \hat{K}$.

- $K \neq \hat{K}$ **:** It has a very interesting observation.
  - They still may be close to each other even when all the cluster centroids do not represent true centroids.
  - They may be randomly scattered. The probability of occurring the case is increasing with a growing number of observed pairs of samples.

Intuitively, more than one centroids get assigned to a compact cluster when a data is partitioned with $K > \hat{K}$. Interestingly, a compact cluster that may split varies in different samples.

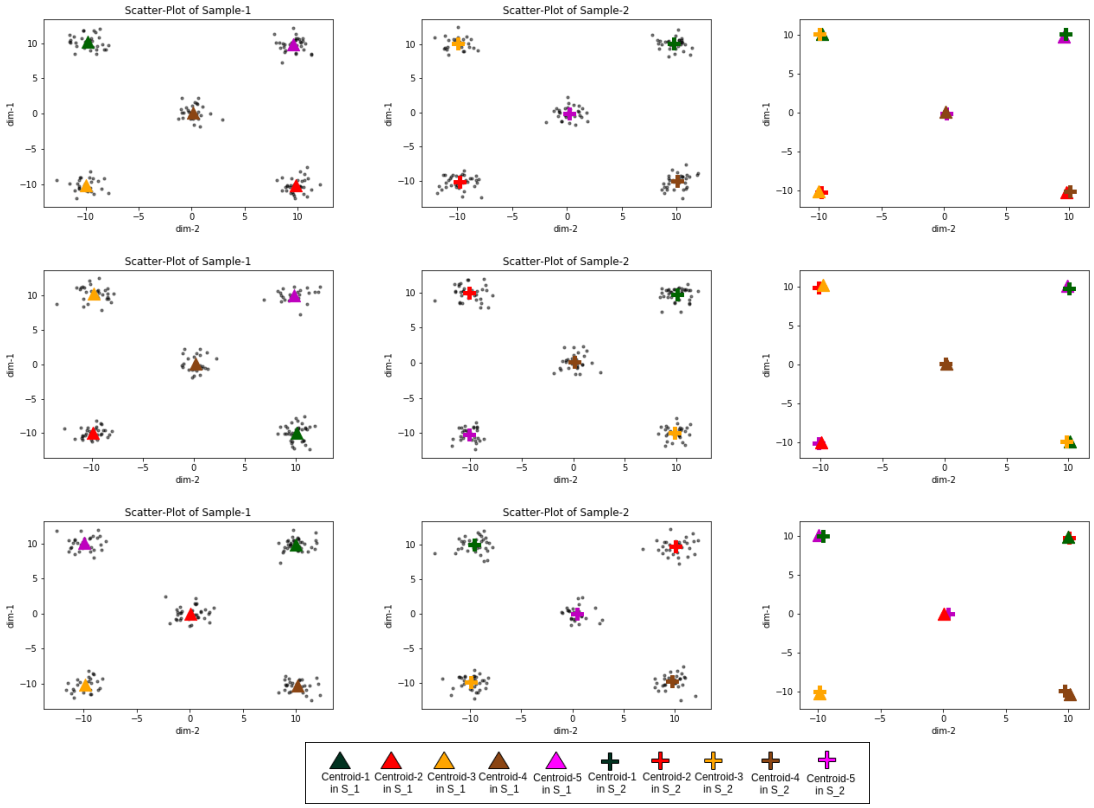

Figure 3: Each row shows the positional behavior of cluster centroids chosen by *K-means++* in a pair of samples H (S_1, S_2). First and second column of each row depicts randomly selected sample distributions with cluster centroids generated by *K-means++* with $K = 5$. Third column depicts a plot where set of centroids generated from S_1 and S_2 are plotted.

Conclusion on the experiment :

- It indicates that for a dataset, other than $K = \hat{K}$, there is a high chance of having a pair of centroids whose distance is larger compared to the other pairs under observation.
- The number of such cases increases with increasing pairs of observations.
- Therefore, the average distance between a pair of centroids is expected to be higher for $K \neq \hat{K}$.
- The distortion would be high when K is non-optimal. Hence, the positions of centroids in various samples provide random behavior for $K = 3$ and $K = 7$.

Dependency

- numpy
- sklearn
- munkres (by Brian M. Clapper)

Publications

```
@article{SAHA2020,
author = "Jayasree Saha and Jayanta Mukherjee",
title = "CNAK : Cluster Number Assisted K-means",
journal = "Pattern Recognition",
pages = "107625",
year = "2020",
issn = "0031-3203",
doi = "https://doi.org/10.1016/j.patcog.2020.107625",
url = "http://www.sciencedirect.com/science/article/pii/S0031320320304283",
}
```