K-Means Clustering

Centroid Based Clustering (K-Means)

- Select k number of classes and initialize their respective center points
- The center points are vectors of the same length as each data point vector
- Each data point is classified to be in the group whose center is closest to it
- Based on these classified points, recompute the group center
- Repeat these steps until the expected result
- Consider the following 2D points and find 3 cluster centroids for those

X	**1**	**-3**	**0**	**-7**	**0**	**6**	**2**	**-1**	**0**	**10**
Y	**2**	**6**	**3**	**-4**	**7**	**9**	**3**	**-1**	**-100**	**10**

Centroid Based Clustering (K-Means)

X	1	-3	0	-7	0	6	2	-1	0	10
Y	2	6	3	-4	7	9	3	-1	-100	10
Cluster										

Distance (C1)										
Distance (C2)										
Distance (C3)										

C1	(1,2)
C2	(-3,6)
C3	(0,3)

For simplicity, we will use squared distances

Centroid Based Clustering (K-Means)

X	1	-3	0	-7	0	6	2	-1	0	10
Y	2	6	3	-4	7	9	3	-1	-100	10
Cluster	1	2	3	3	2	3	1	1	1	1

Distance (C1)	0	32	2	100	26	74	2	13	10405	145
Distance (C2)	32	0	18	116	10	90	34	53	11245	185
Distance (C3)	2	18	0	98	16	72	4	17	10609	149

C1	(1,2)
C2	(-3,6)
C3	(0,3)

Update the cluster centroid

C1	(2.4,-17.2)
C2	(-1.5,6.5)
C3	(-0.3,2.7)

Centroid Based Clustering (K-Means)

X	1	-3	0	-7	0	6		-1	0	10
Y	2	6	3	-4	7				-100	10
Cluster										

Distance (C1)										
Distance (C2)										
Distance (C3)										

Do Yourself

C1	
C2	
C3	,2.7)

Update the cluster centroid

C1	
C2	
C3	

How to Find a Proper k value?

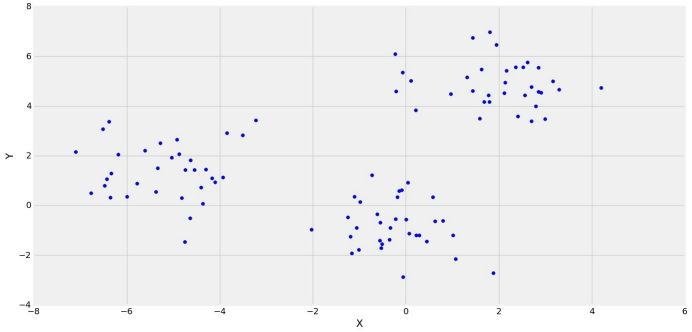
A number of analysis are used:

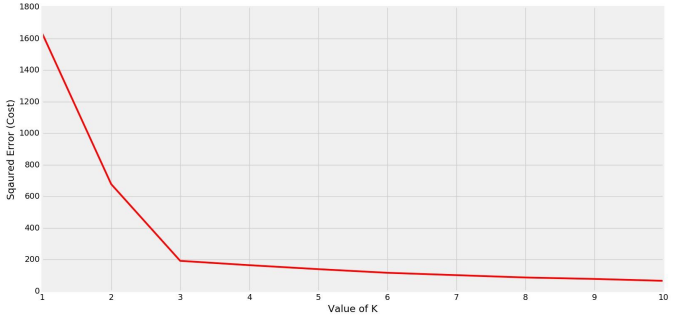
- Elbow Method
- Average Silhouette Method
- Gap Statistic Method

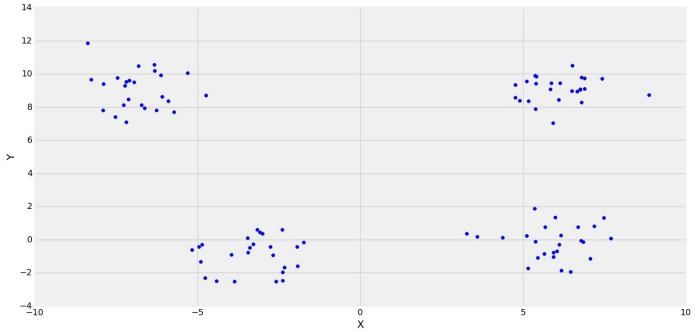
Elbow Method

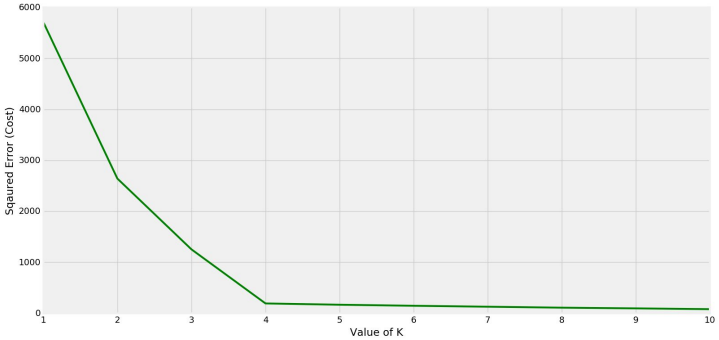
- Compute clustering algorithm for different values of k
- For each k, calculate the sum of intra-cluster squared error (sse)
- Plot the curve of **sse vs k**
- The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number (k) of clusters

Elbow Method









Average Silhouette Method

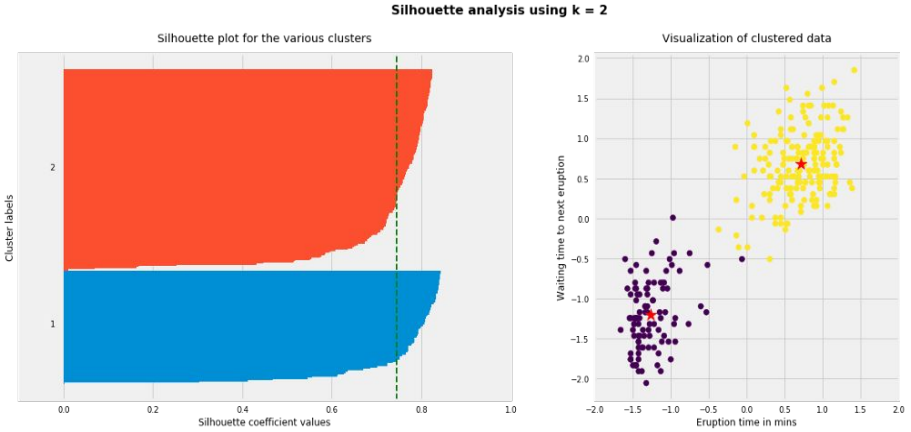
- Compute clustering algorithm for different values of k
- For each k, calculate the average silhouette of observations (avg.sil)
- Plot the curve of **avg.sil vs k**
- The location of the maximum is considered as the appropriate number (k) of clusters
- Silhouette Coefficient is calculated as:

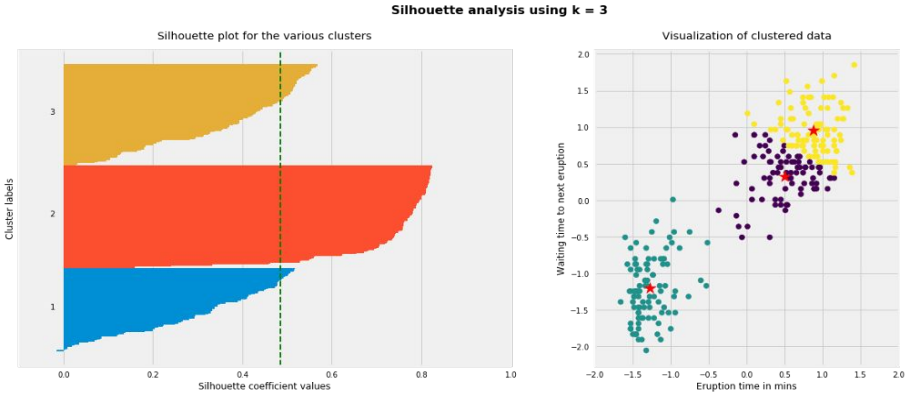
$$silCof = \frac{\beta - \alpha}{\max(\beta, \alpha)}$$

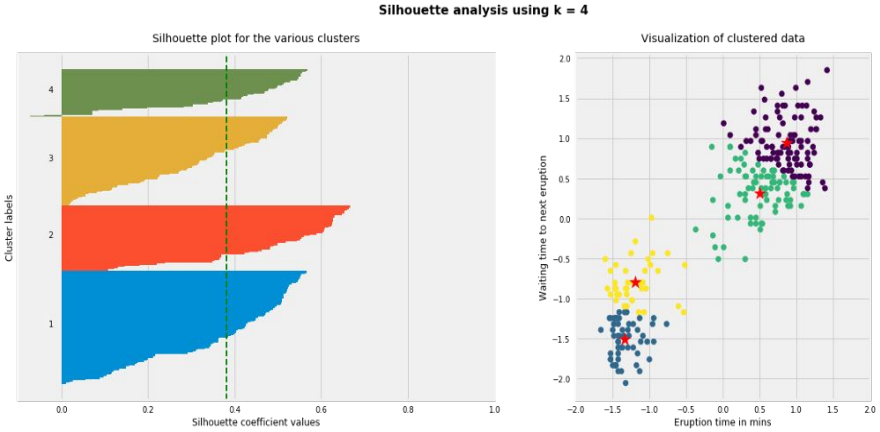
$$\alpha = average\ intra - cluster\ distance$$

$$\beta = minimum\ average\ inter - cluster\ distance$$

Average Silhouette Method



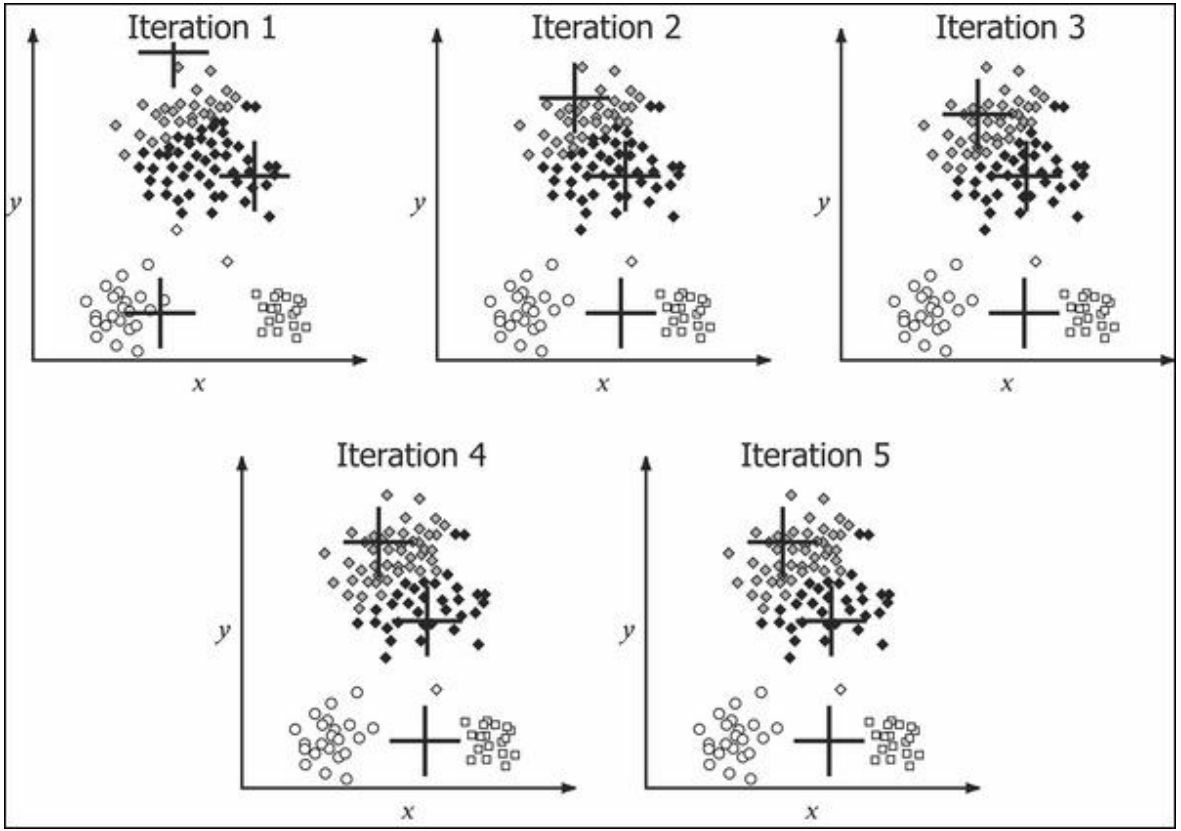




Drawbacks of K-Means

- Works poor with complex geometric shapes of clusters
- Can’t handle outlier
- Can’t guarantee to find the global optimum clusters
- Can’t handle non-numerical data
- …

Drawbacks of K-Means



Assignment

Find the optimum number of cluster for the given dataset using Elbow Method.

Dataset Preparation: If your student id is
ABCD-E-FG-HIJ

X	-1	7	2	0	9	-3	5	8	-6	4
Y	A	B	C	D	E	F	G	H	I	J

Thank You

References

- https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1
- https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a
- https://www.geeksforgeeks.org/ml-determine-the-optimal-value-of-k-in-k-means-clustering/#:~:text=There%20is%20a%20popular%20method,fewer%20elements%20in%20the%20cluster.
- http://user.ceng.metu.edu.tr/~akifakkus/courses/ceng574/k-means/