Centroid-based Clustering

Centroid Based Clustering (K-Medoids)

1. Initialize: select k random points out of the n data points as the medoids.
2. Associate each data point to the closest medoid by using any common distance metric methods.

$$c = \sum_{Ci} \sum_{Pi \in Ci} |Pi - Ci|$$

3. While the cost decreases:
For each medoid m, for each data point d which is not a medoid:
a. Swap m and d, associate each data point to the closest medoid, recompute the cost.
b. If the total cost is more than that in the previous step, undo the swap.

- 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-Medoids)

C1 = **(1,2),** (-1, -1), (0, -100), (10, 10)(2, 3)
C2 = **(-3,6),** (0, 7), (6, 9)
C3 = **(0,3),** (-7, -4)

$$c = \sum_{Ci} \sum_{Pi \in Ci} |Pi - Ci|$$

c = (c1's cost) + (c2's cost) + (c3's cost)

C1's cost = sum(2+3, 1+99, 9+8, 1+1) = 5+100+17+2 = 124
C2's cost = sum(3+1, 9+3) = 4+12 = 16
C3's cost = sum(7+7) = 14

c = 124+16+14 = 154

X	1	-3	0	-7	0	6	2	-1	0	10
Y	2	6	3	-4	7	9	3	-1	-100	10
	C1	C2	C3	C3	C2	C2	C1	C1	C1	C1

Centroid Based Clustering (K-Medoids)

C1 = **(0, 7)**, (-7, -4), (6,9), (2, 3)
C2 = **(-3,6)**, (1, 2), (-1, -1)
C3 = **(0,3)**, (0, -100), (10, 10)

$$c = \sum_{Ci} \sum_{Pi \in Ci} |Pi - Ci|$$

c = (c1's cost) + (c2's cost) + (c3's cost)

C1's cost = sum(2+3, 1+99, 9+8, 1+1) = 5+100+17+2 = 124
C2's cost = sum(3+1, 9+3) = 4+12 = 16
C3's cost = sum(7+7) = 14

c = 124+16+14 = 154

X	1	-3	0	-7	0	6	2	-1	0	10
Y	2	6	3	-4	7	9	3	-1	-100	10
	C1	C2	C3	*C1*	*C2*	*C1*	*C1*	*C2*	*C3*	*C3*

Centroid Based Clustering (K-Medoids)

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

$$(e\_dist(x_1\ y_1)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Centroid Based Clustering (**K-Medoids**)

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) <1,2>	**0**	5.66	1.41	10	5	8.6	1.41	**3.6**	**102.005**	**12.04**
Distance (C2) <-3, 6>	5.66	**0**	4.24	10.77	**3.16**	**3.46**	5.83	7.28	106.04	13.60
Distance (C3) <0, 3>	1.41	4.24	**0**	**9.9**	4	8.45	**2**	4.12	103	12.21

$$c = \sum_{Ci} \sum_{Pi \in Ci} |Pi - Ci|$$

C1	(1,2), (-1, -1), (0, -100), (10, 10)	Cost = 0+ 3.6 + 102.005 + 12.04 = 117.645	(-1, -1)
C2	(-3,6), (0, 7), (6, 9)	Cost = 0+ 3.16 + 3.46 = 6.62	(-3, 6)
C3	(0,3), (-7, -4), (2, 3)	Cost = 0+ 9.9 + 2 = 11.9	(0, 3)

How to Find a Proper k value?

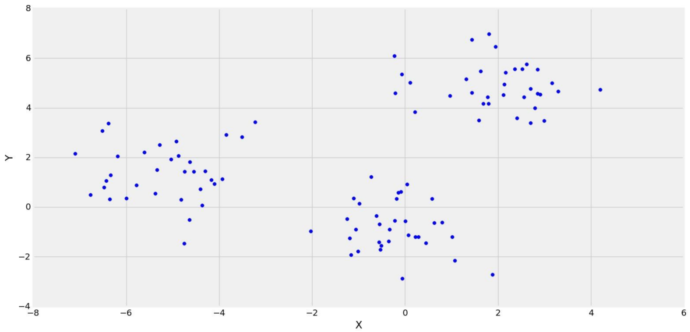
A number of analysis are used:

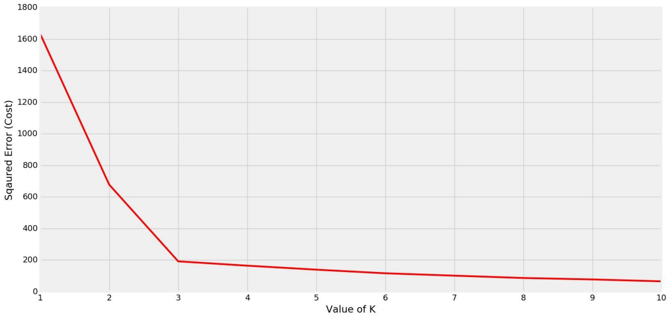
- Elbow Method
- Average Silhouette Method
- Gap Statistic Method

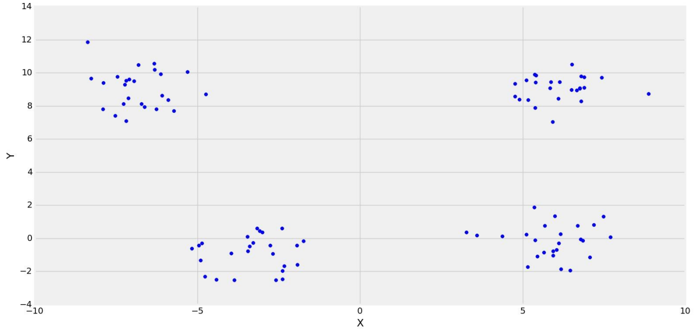
Elbow Method

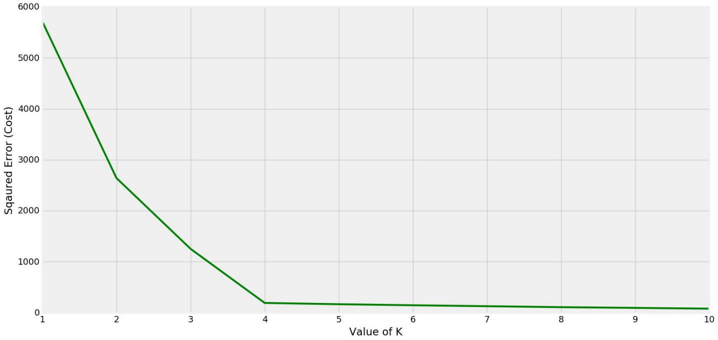
- Compute clustering algorithm for different values of k
- For each k, calculate the sum of intra-cluster squared error (sse)
- Plot the curve of **sse vs k**
- The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number (k) of clusters

Elbow Method









Average Silhouette Method

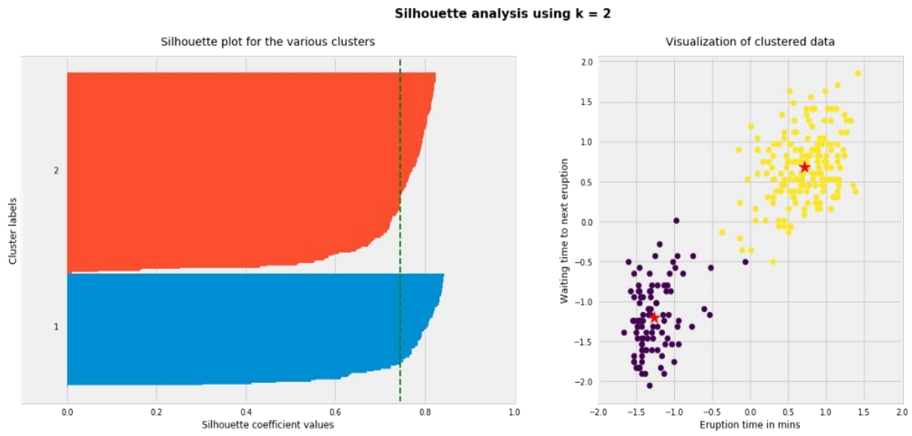
- Compute clustering algorithm for different values of k
- For each k, calculate the average silhouette of observations (avg.sil)
- Plot the curve of **avg.sil vs k**
- The location of the maximum is considered as the appropriate number (k) of clusters
- Silhouette Coefficient is calculated as:

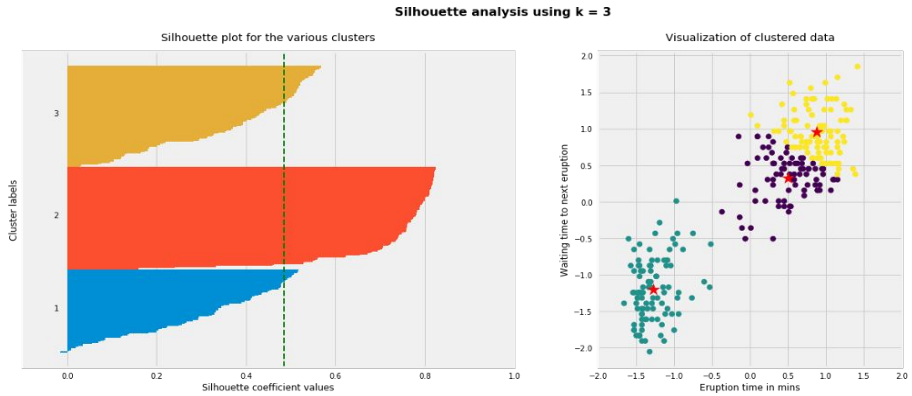
$$silCof = \frac{\beta - \alpha}{\max(\beta, \alpha)}$$

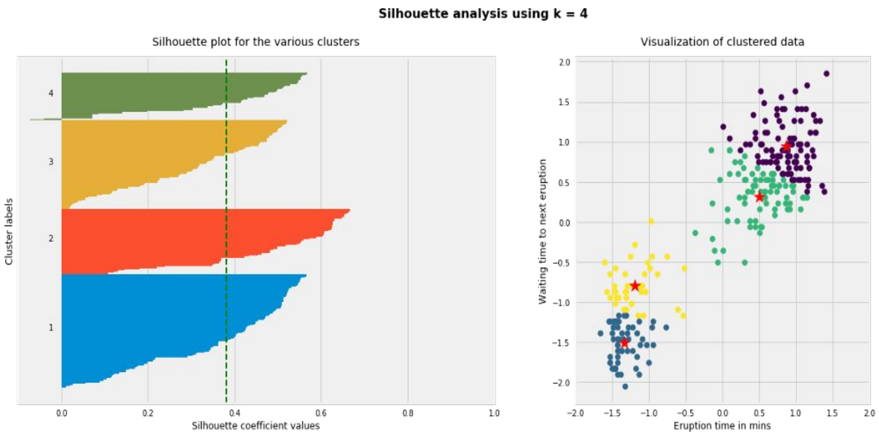
$$\alpha = average\, intra - cluster\, distance$$

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

Average Silhouette Method



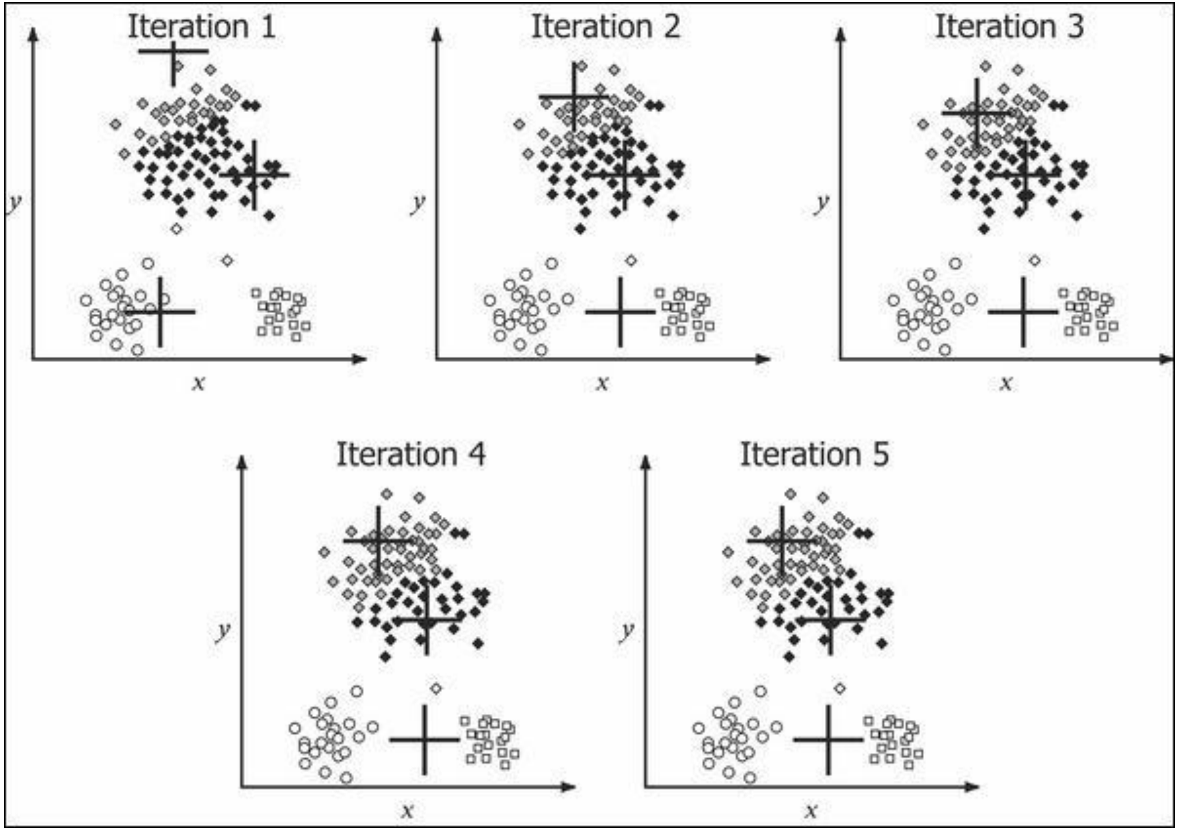




Drawbacks of K-Means

- 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