K-Means

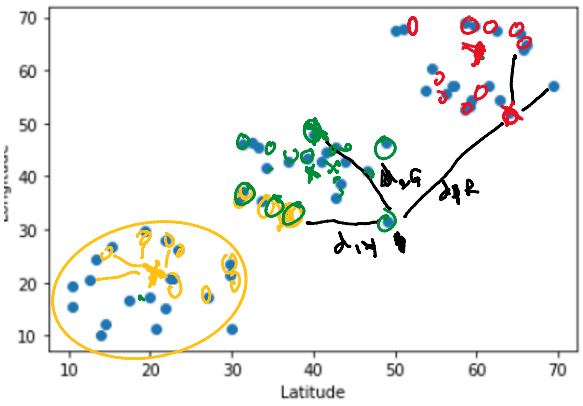
02 March 2022 20:51

Unsupervised Learning (No Target Variable)

→ Clustering Group observations such that With in same group observations are similar (Homogenous)
Across groups different (Heterogenous)

→ K-Means ✓ K - no. of clusters "K = 5"

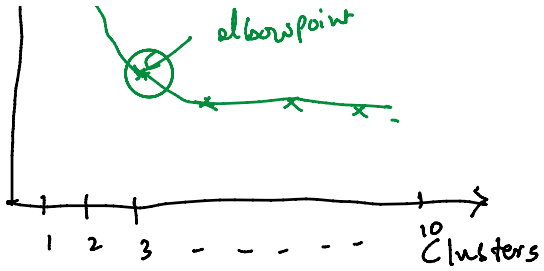
K



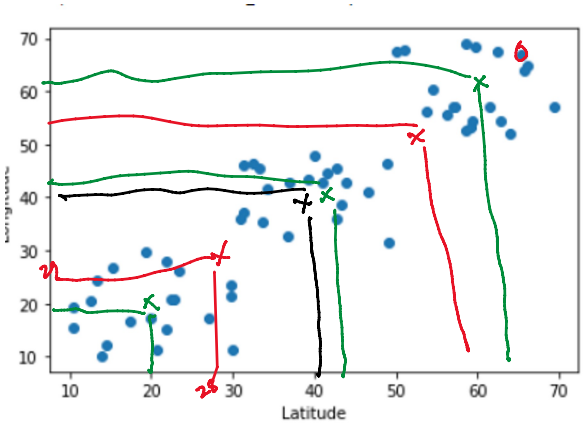
Euclidian distance

$(n-3) \times 3$

Elbow Method



$S_1 = 1000$



$x_1 = (10-40)^2 + (11-40)^2 + \dots$

$x_2 = (10-40)^2 + (11-40)^2 + \dots$

$x_1 \; (10-29)^2 + \dots$ | $x_1 \; (70-55)^2 + \dots$ $= 700$

$S_2 \quad x_2 \; (11-28)^2 - 1 \dots$ | $x_2 \; (70-60)^2 + \dots$

400 | 400

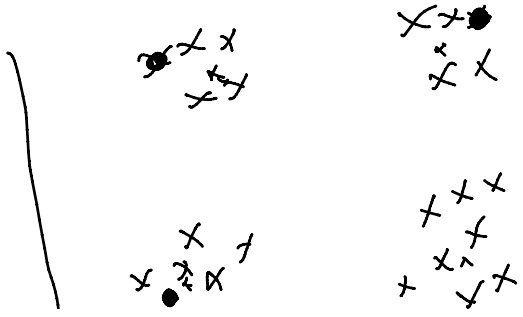
S_3 150 200 300

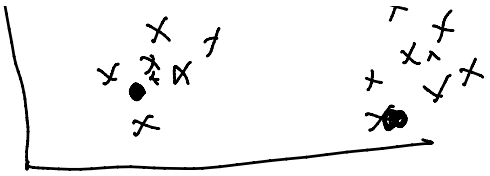
$S_3 = 650$

$S_4 = 620$

Initialization : (Kmeans ++)

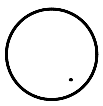
It chooses cluster Centers one by one





Silhouette Score. $(-1, +1)$

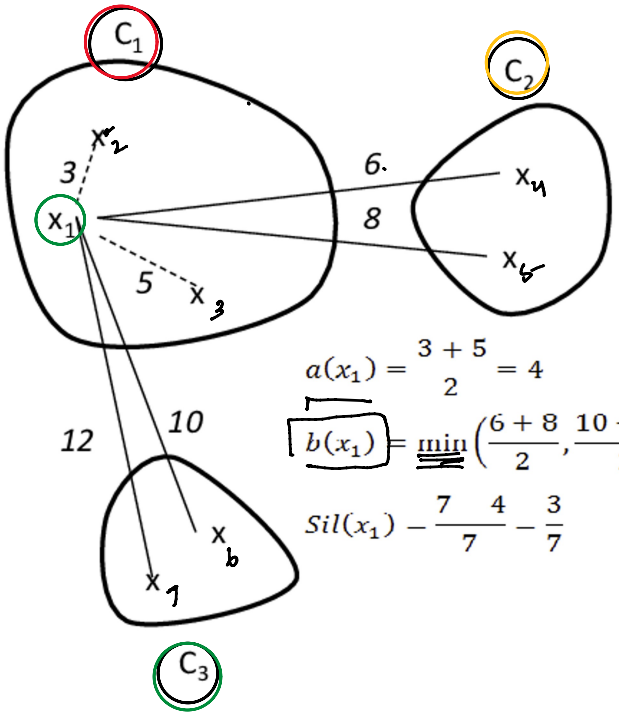
Clusters



a - within-cluster distance.

b - out of cluster distance.

Example of calculation of Silhouette score:



$$a(x_1) = \frac{3+5}{2} = 4$$

$$b(x_1) = \min\left(\frac{6+8}{2}, \frac{10+12}{2}\right) = 7$$

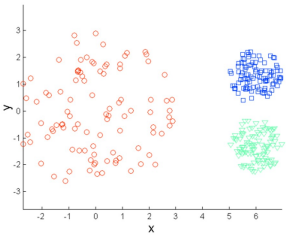
$$Sil(x_1) - \frac{7}{7} - \frac{4}{7} - \frac{3}{7}$$

$a(x_1) \rightarrow \{C_1\}$

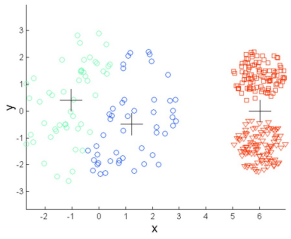
$a = \frac{3+5}{2} = 4.$

$\min\left(\frac{6+8}{2}, \frac{10+12}{2}\right) = 7$

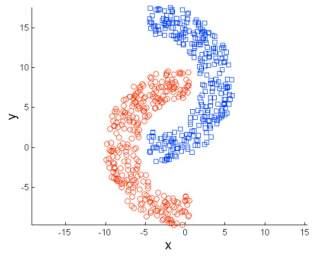
$Sil \frac{(b-a)}{\max(a,b)} = \frac{7-4}{7} = \left(\frac{3}{7}\right)$



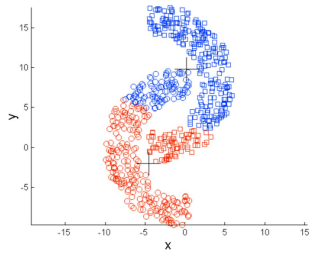
Original Points



K-means (3 Clusters)



Original Points



K-means (2 Clusters)