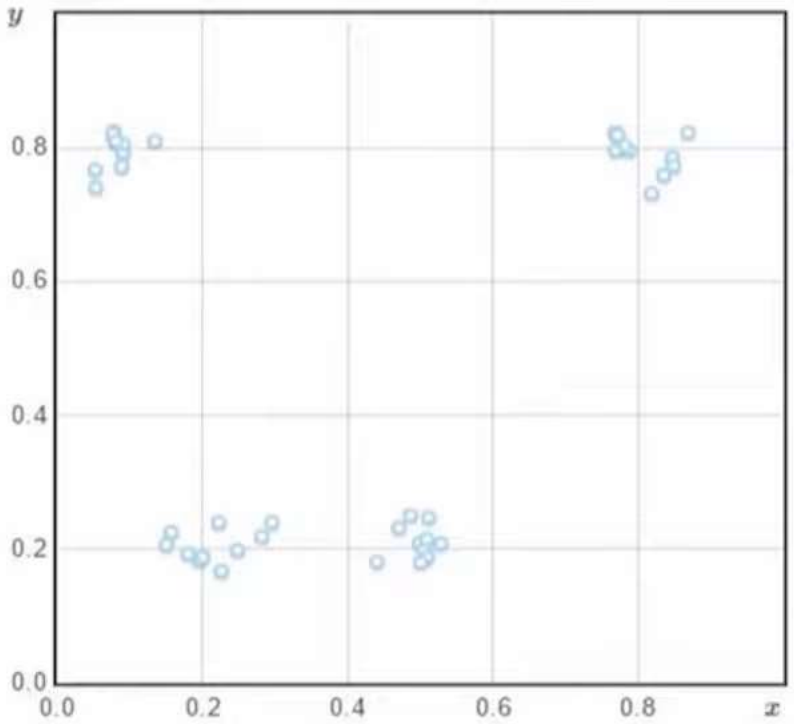
What is K-means clustering?

- A technique to partition N observations into K clusters (K≤N) in which each observation belongs to cluster with nearest mean

- One of the simplest unsupervised algorithms

- Works well for all distance metrics where mean is defined (ex. Euclidean distance)



Description of K-means clustering

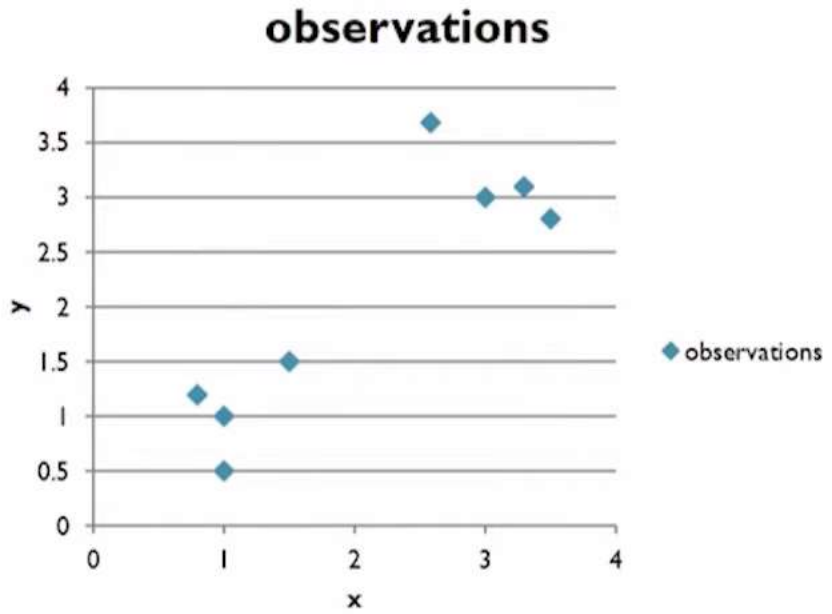
Given N observations $(x_1, x_2, ..., x_N)$, K-means clustering will partition n observations into K ($K \leq N$) sets $S=\{s_1, ..., s_k\}$ so as to minimize the within cluster sum of squares (WCSS)

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

Where $\mu(i)$ is the mean of points in s(i)

Algorithm with an example

Obervation	x	y
1	1	1
2	1.5	1.5
3	1	0.5
4	0.8	1.2
5	3.3	3.1
6	2.58	3.68
7	3.5	2.8
8	3	3



Example continued…

Step 1: Randomly choose two points as the cluster centers

	Individual	Mean x	Mean y
Group 1	1	1	1
Group 2	8	3	3

Step 2: Compute the distances and group the closest ones

Observation	distance 1	distance 2	Group
1	0	2.8284271	1
2	0.7071068	2.1213203	1
3	0.5	3.2015621	1
4	0.2828427	2.8425341	1
5	3.1144823	0.3162278	2
6	3.111077	0.7992496	2
7	3.0805844	0.5385165	2
8	2.8284271	0	2

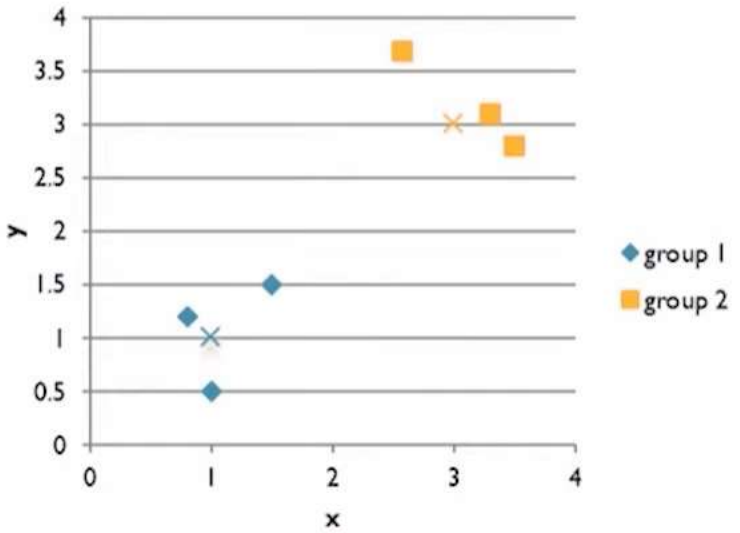


Fig: Plot of groups after step 1 & 'x' - mean

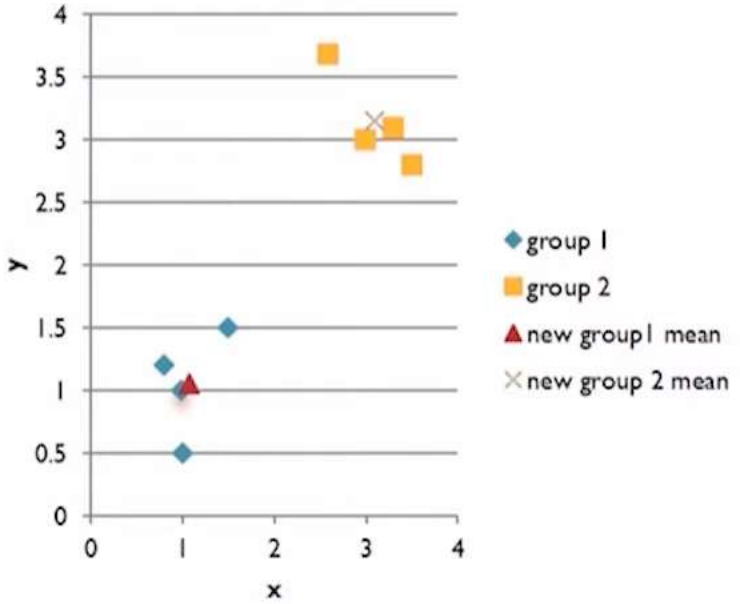
Example continued…

Step 3: Compute the new mean and repeat step 2

	Individual	Mean x	Mean y
Group1	1,2,3,4	1.075	1.05
Group 2	5,6,7,8	3.095	3.145

Step 4: If change in mean is negligible or no reassignment then stop the process

Observation	distance 1	distance 2	Group
1	0.0901388	2.9983412	1
2	0.6189709	2.2912988	1
3	0.5550901	3.374174	1
4	0.3132491	3.0083301	1
5	3.0254132	0.2098809	2
6	3.0301691	0.7425968	2
7	2.9905058	0.5320244	2
8	2.7400958	0.1733494	2



Determining number of clusters(K)

- Elbow method – looks at percentage of variance explained as a function of number of clusters
- The point where marginal decrease plateaus is an indicator of the optimal number of clusters
- We will see a demonstration of this in the example

Disadvantages of K-means

- This algorithm could converge to a local minima, therefore role of initial position is very important
- If the clusters are not spherical, then K-means can fail to identify the correct number of clusters