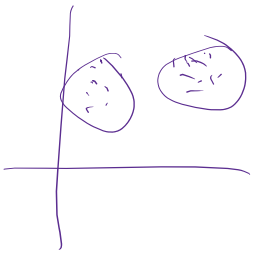
* Introduction to K-Means
* K means core idea and key Hyperparameters (K, init, max-iter)
* choosing number of clusters and practical tips (Elbow Method, Scaling)
* Implementing K-means on a Dataset
* Model Evaluation, Limitations and Use Cases



Introduction to Clustering and K-Means

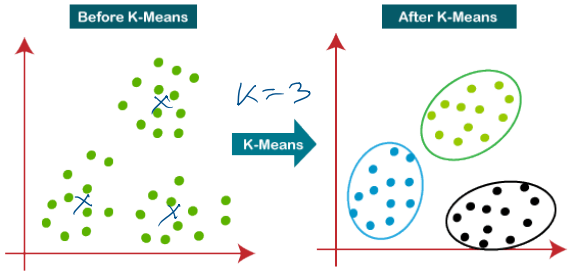
* what is clustering?

Group data points such that -

i) Intra-cluster similarity is high

ii) Inter-cluster similarity is low

It is a unsupervised learning problem.

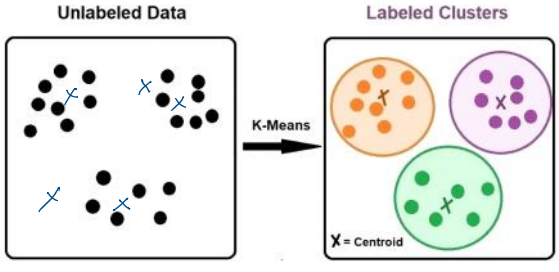


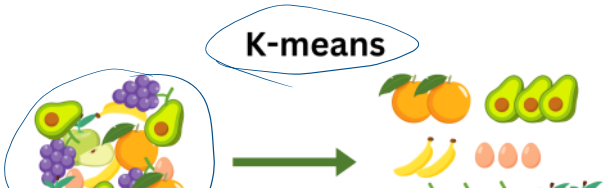
* what is K-Means?

- K-Means is a partition based clustering algorithm.

It divides N data points into k disjoint clusters

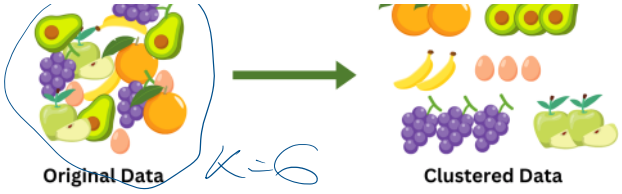
= * assigning points to the nearest centroid (mean)

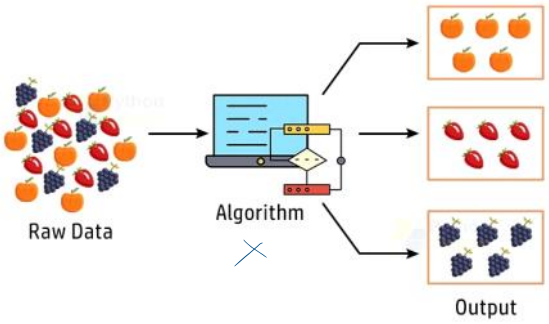




nearest centroid (mean)

= * updating centroids as the mean of assigned points





K-Means Core Idea and Key Hyperparameters

Core steps of K-means:

1. choose k (number of clusters)
2. Initialize k centroids (randomly or a smarter method)
3. Assign each points to the closest centroid
4. update centroids (mean of points in each cluster)
5. Repeat steps 3 and 4 until assignments stop changing (convergence)

k=2

[1, 2, 3, 4] [10, 12, 14]

Key hyperparameters:

i) k: number of clusters

ii) init: centroid initialization
- ↳ random
- ↳ k-means++

iii) max-iter: maximum iteration

* Elbow method

~ scaling

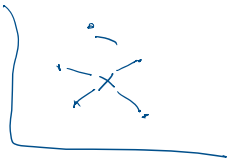
* Elbow method
* Scaling
* K-means++

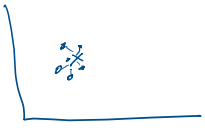
Choosing Number of Clusters and Practical Tips

* Elbow method:

Inertia

$$WCSS = \sum_{j=1}^{k} \sum_{x_i \in C_j} \| x_i - \mu_j \|^2$$



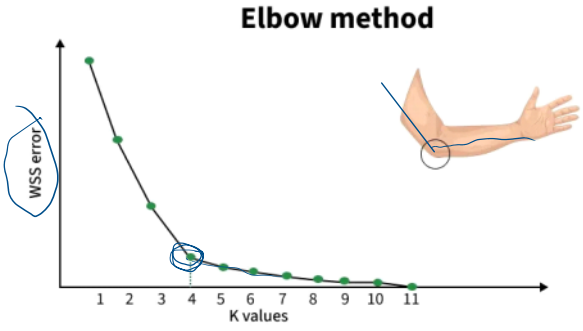


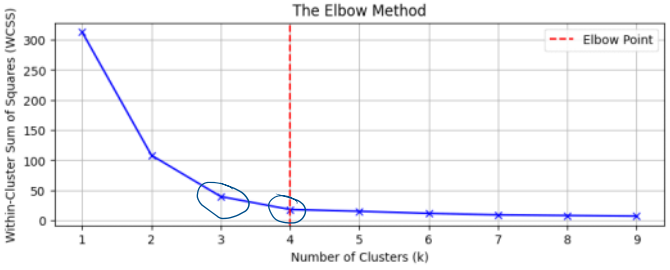
k = number of clusters

C_j = j-th cluster

x_i = data point of cluster

μ_j = j-th cluster centroid





points:

(2,3), (3,4), (4,5)

$\mu = \left(\frac{2+3+4}{3}, \frac{3+4+5}{3} \right)$

$= (3,4)$

point	Distance to (3,4)	squared
(2,3)	$\sqrt{2}$	2
(3,4)	0	0
(4,5)	$\sqrt{2}$	2

Inertia = 2 + 0 + 2 = 4

Multiple clusters,

K = 2

Inertia = Intertia of cluster 1 + Intertia of cluster 2

K = 2

Intertia = Intertia of cluster 1 + Intertia of cluster 2

For Elbow method,

K = 1, 2, 3, --- None

for each k, we calculate inertia

Then we plot it

K vs inertia

* Scaling:

Example:

Age range: 0 to 100

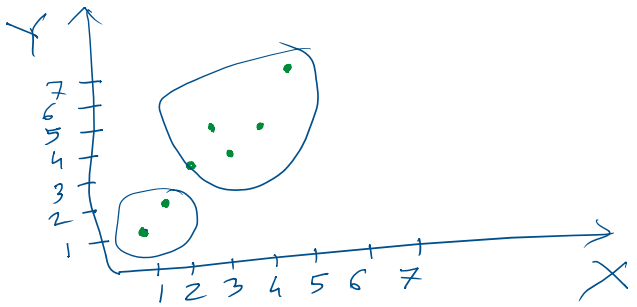
Income range: 0 to 10,0000,0

K-means++

Worked Example: K-Means with k = 2

A small 2D dataset of 7 points

point (ID)	x (variable 1)	y (variable 2)
→ 1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
	5.0	7.0
→ 4	3.5	5.0
5	4.5	5.0
6	3.5	4.5
7		



Step 1: Initialization

$K = 2$,

$m_1^{(0)} = (1.0, 1.0)$, $m_2^{(0)} = (5.0, 7.0)$

Step 2: Compute disances and assign clusters (Iteration 0)

we use Eucledian distance: $d((x,y),(a,b)) = \sqrt{(x-a)^2 + (y-b)^2}$

point	Distance to m_1	Distance to m_2	Assign
1	$\sqrt{(1-1)^2+(1-1)^2} = 0.0$	$\sqrt{(1-5)^2+(1-7)^2} = 7.21$	c1
2	$\sqrt{(1.5-1)^2+(2-1)^2} = 1.12$	$\sqrt{(1.5-5)^2+(2-7)^2} = 6.10$	c1
3	3.61	3.61	c1
4	7.21	0.00 ↓	c2
5	4.72	2.50 ↓	c2
6	5.31	2.06 ↓	c2
7	4.30	2.92 ↓	c2

So, after Iteration 0, assignments are:

Cluster C1: {1, 2, 3} →

Cluster C2: {4, 5, 6, 7} →

Step-3: Update centroids

For C1 ({1, 2, 3})

$$m_1^{(1)} = \left(\frac{1.0+1.5+3.0}{3}, \frac{1.0+2.0+4.0}{3}\right) = (1.83, 2.33)$$

For C2 ({4, 5, 6, 7})

$$m_2^{(1)} = \left(\frac{5.0+3.5+4.5+3.5}{4}, \frac{7.0+5.0+5.0+4.5}{4}\right)$$

$$= (4.12, 5.38)$$

Step-3: Recompute distances and Reassign (Iteration 1)

point	d to $m_1^{(1)}$	d to $m_2^{(1)}$	Assign
1	$\sqrt{(1-1.83)^2+(1-2.33)^2}=1.57$	$\sqrt{(1-4.12)^2+(1-5.38)^2}=5.38$	c1
2	0.47 ↓	4.28	c1
3	2.04	1.78 ↓	c2 → c1
4	5.64	1.84 ↓	c2
5	3.15	0.73 ↓	c2
6	3.78	0.54 ↓	c2
7	2.74	1.08 ↓	c2

So, after Iteration 1, assignments:

- cluster c1 = {1, 2}
- cluster c2 = {3, 4, 5, 6, 7}

Step-5: Update centroids again

For c1 ({1, 2}) $m_1^{(2)} = \left(\frac{1.0+1.5}{2}, \frac{1.0+2.0}{2}\right) = (1.25, 1.50)$

For c2 ({3, 4, 5, 6, 7})

$$m_2^{(2)} = \left(\frac{3.0+5.0+3.5+4.5+3.5}{5}, \frac{4.0+7.0+5.0+5.0+4.5}{5}\right)$$

$$= (3.90, 5.10)$$

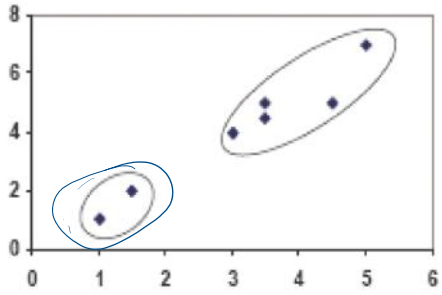
point	d from $m_1^{(2)}$	d from $m_2^{(2)}$	Assign

Final Assignments,

Cluster 1 : {1, 2}

Cluster 2 : {3, 4, 5, 6, 7}

PLOT



For, k = 3

PLOT