Experiment 3: Program for K-mean clustering.

K means clustering is an algorithm, where the main goal is to group similar data points into a cluster. In K means clustering, K represents the total number of groups or clusters. K means clustering runs on Euclidean distance calculation. Now, let us understand K means clustering with the help of an example.

Say, we have a dataset consisting of height and weight information of 10 players. We need to group them into two clusters based on their height and weight.

Height	Weight
180	80
172	73
178	69
189	82
164	70
186	71
180	69
170	76
166	71
180	72

Step 1: Initialize a cluster centroid

Initial Clusters	Height	Weight
K1	185	70
K2	170	80

Step 2: Calculate the Euclidean distance from each observation to the initial clusters

$$\text{Euclidean Distance} = \sqrt{(x_{height} - H_{centroid})^2 + (x_{weight} - W_{centroid})^2}$$

Observation	Height	Weight	Distance from Cluster 1	Distance from Cluster 2	Assign Clusters
1	180	80	11.18	10	2
2	172	73	13.3	7.28	2
3	178	69	7.07	13.6	1
4	189	82	12.64	19.10	1
5	164	70	21	11.66	2
6	186	71	1.41	18.35	1
7	180	69	5.09	14.86	1
8	170	76	16.15	4	2
9	166	71	19.02	9.84	2
10	180	72	5.38	12.80	1

Step 3: Find the new cluster centroid

Observation	**Height**	**Weight**	**Assign Clusters**
1	180	80	2
2	172	73	2
3	178	69	1
4	189	82	1
5	164	70	2
6	186	71	1
7	180	69	1
8	170	76	2
9	166	71	2
10	180	72	1
New Cluster 1	(178+189+186+180+180)/5	(69+82+71+69+72)/5	
New Cluster 2	(180+172+164+170+166)/5	(80+73+70+76+76+71)/5	

New Cluster 1 = (182.6, 72.6)
New Cluster 1 = (170.4, 89.2)

Step 4: Again, calculate the Euclidean distance

Calculate the Euclidean distance from each observation to both Cluster 1 and Cluster 2

Repeat Steps 2, 3, and 4, until cluster centers don’t change any more

Now, let us look at the hands-on given below to have a deeper understanding of K-means algorithm.

Q1: Implement K-means Clustering Algorithm Using Sklearn in Python- Iris Dataset

Solution:

```
# Required Libraries
from sklearn.datasets import load_iris
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt


# Load the Iris dataset
iris = load_iris()
data = iris.data
print(data)
target = iris.target
print(iris.target)

# Perform K-means clustering and take cluster centers
kmeans_model = KMeans(n_clusters=3, random_state=1).fit(data)
labels = kmeans_model.labels_
clusters_sklearn = kmeans_model.cluster_centers_

print(labels)
print(kmeans_model.cluster_centers_)


# Function to plot final clusters
def plot_clusters_sklearn(data, labels, clusters):
    # Plot data points, choosing first two features for visualization
    plt.scatter(data[:, 0], data[:, 1], c=labels, cmap='viridis', label='Data points')
    # Plot cluster centers, also focusing on the same two features
    plt.scatter(clusters[:, 0], clusters[:, 1], s=200, color='red', marker='X', label='Centers')
    plt.title('Visualizing Clusters with Matplotlib using Iris Dataset')
    plt.xlabel('Sepal Length (cm)')
    plt.ylabel('Sepal Width (cm)')
    plt.legend()
    plt.grid(True)
    plt.show()

# Visualize the clusters
plot_clusters_sklearn(data, labels, clusters_sklearn)
```