SNS COLLEGE OF ENGINEERING

(Autonomous)

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

Artificial Intelligence & Machine Learning

Unit 3 – Unsupervised Learning

K-Means Clustering

Prepared by,
P.Ramya
Assistant Professor/ECE
SNS College of Engineering

K-Means Clustering

- Clustering is an unsupervised machine learning technique. It is the process of division of the dataset into groups in which the members in the same group possess similarities in features.

Contd…

- It is the simplest and commonly used iterative type unsupervised learning algorithm. In this, we randomly initialize the K number of centroids in the data (the number of k is found using the Elbow method which will be discussed later in this article) and iterates these centroids until no change happens to the position of the centroid.

Contd…

Let’s go through the steps involved in K means clustering for a better understanding.

1) Select the number of clusters for the dataset (K)

2) Select K number of centroids

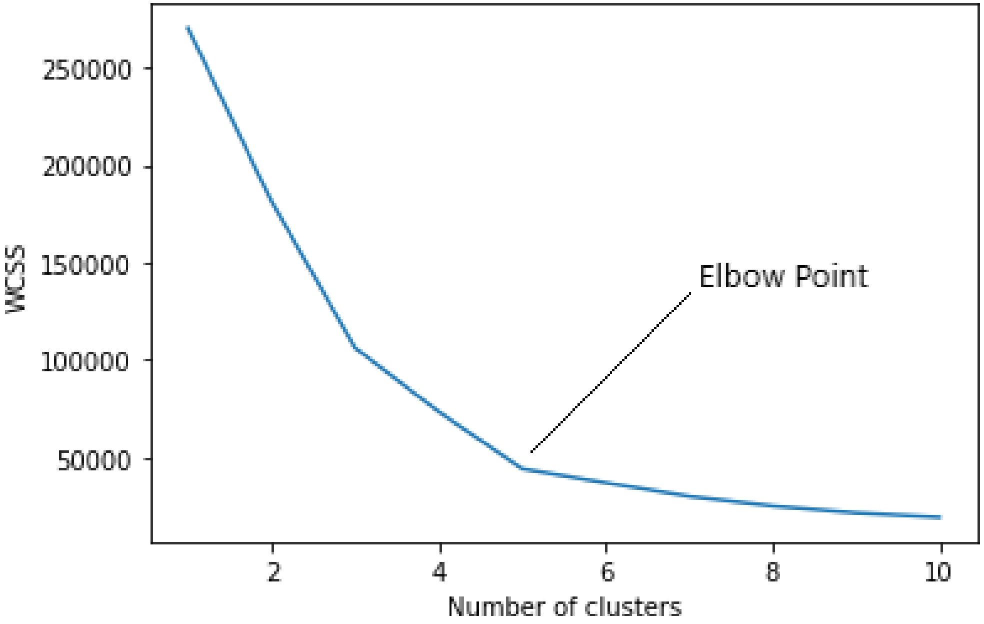
3) By calculating the Euclidean distance or Manhattan distance assign the points to the nearest centroid, thus creating K groups

4) Now find the original centroid in each group

5) Again reassign the whole data point based on this new centroid, then repeat step 4 until the position of the centroid doesn’t change.

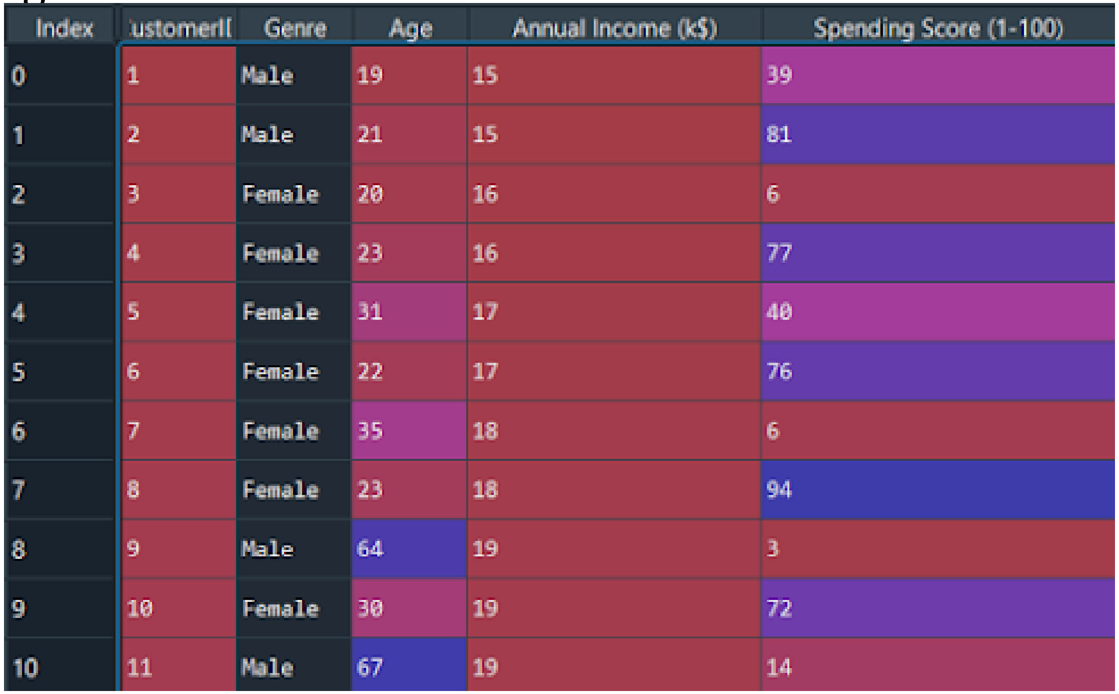
Elbow Method

In the Elbow method, we are actually varying the number of clusters (K) from 1 – 10. For each value of K, we are calculating WCSS (Within-Cluster Sum of Square). WCSS is the sum of squared distance between each point and the centroid in a cluster.



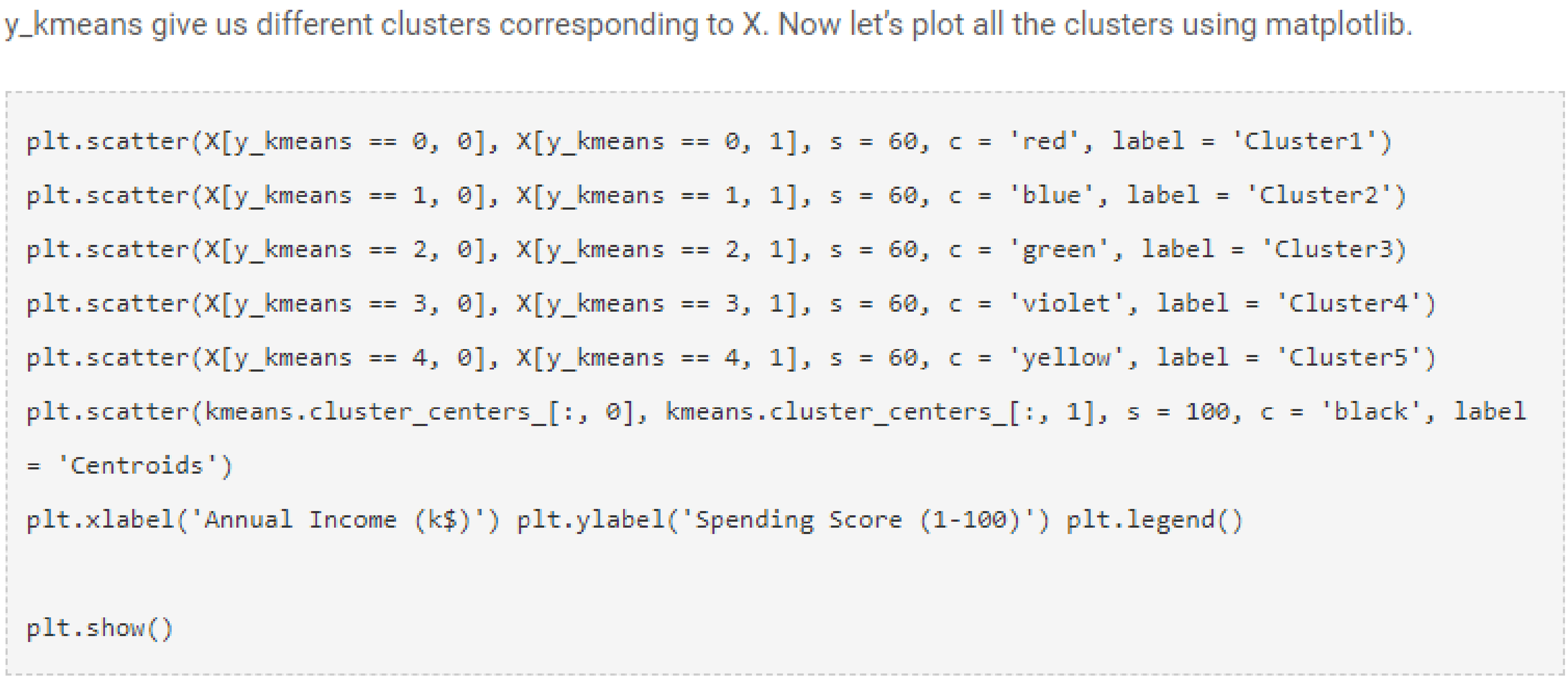
Implementation

Dataset we are using here is the Mall Customers data (Download here). It’s unlabeled data that contains the details of customers in a mall (features like genre, age, annual income(k$), and spending score). Our aim is to cluster the customers based on the relevant features annual income and spending score.



Index	ustomerI[	Genre	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40
5	6	Female	22	17	76
6	7	Female	35	18	6
7	8	Female	23	18	94
8	9	Male	64	19	3
9	10	Female	30	19	72
10	11	Male	67	19	14

Contd...

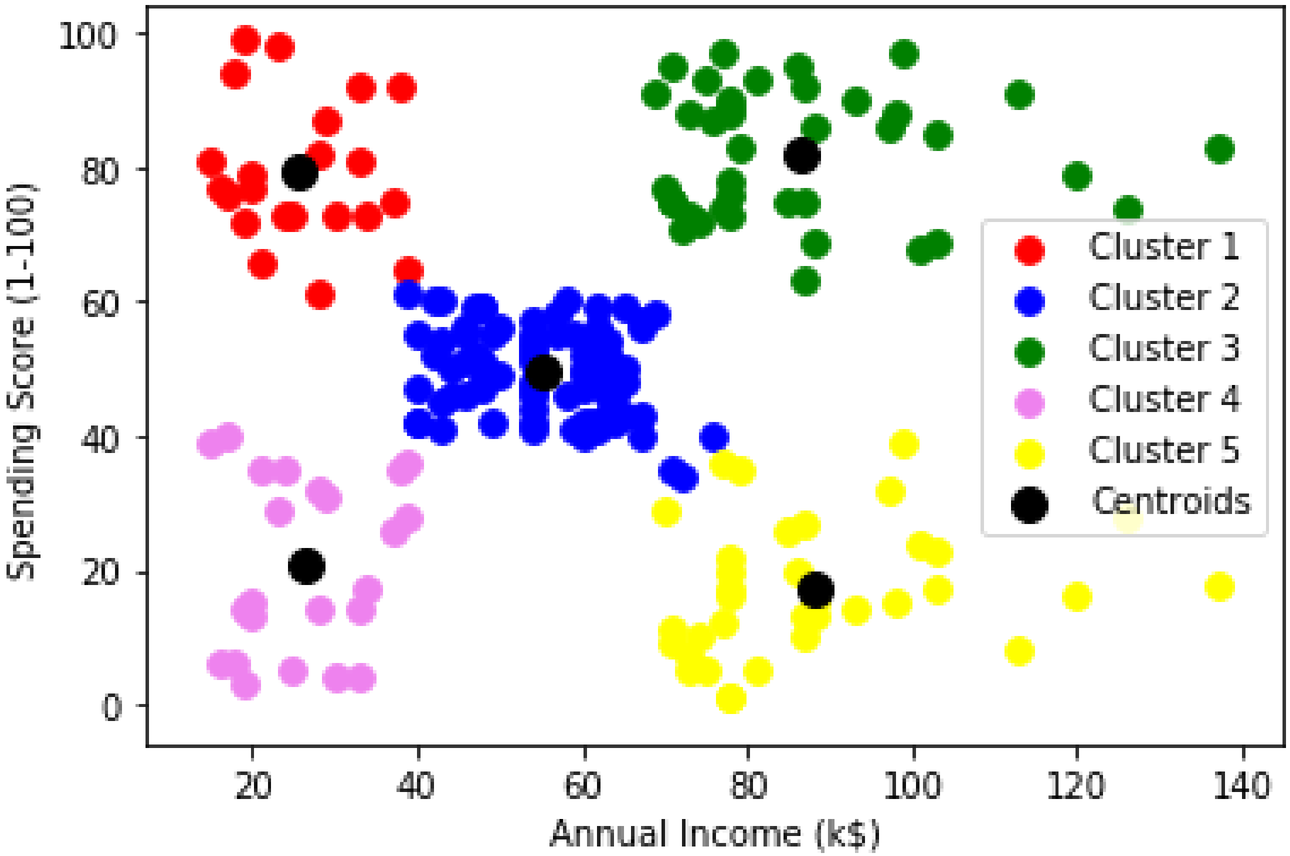


y_kmeans give us different clusters corresponding to X. Now let's plot all the clusters using matplotlib.

```
plt.scatter(X[y_kmeans == 0, 0], X[y_kmeans == 0, 1], s = 60, c = 'red', label = 'Cluster1')
plt.scatter(X[y_kmeans == 1, 0], X[y_kmeans == 1, 1], s = 60, c = 'blue', label = 'Cluster2')
plt.scatter(X[y_kmeans == 2, 0], X[y_kmeans == 2, 1], s = 60, c = 'green', label = 'Cluster3)
plt.scatter(X[y_kmeans == 3, 0], X[y_kmeans == 3, 1], s = 60, c = 'violet', label = 'Cluster4')
plt.scatter(X[y_kmeans == 4, 0], X[y_kmeans == 4, 1], s = 60, c = 'yellow', label = 'Cluster5')
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1], s = 100, c = 'black', label
= 'Centroids')
plt.xlabel('Annual Income (k$)') plt.ylabel('Spending Score (1-100)') plt.legend()


plt.show()
```

Cluster formation



thank
you