Machine Learning (ML)

Chapter 7:

Unsupervised Learning

K-means, and K-medoids

Saeed Saeedvand, Ph.D.

Outline

In this Chapter:

- ✓ Concept of Unsupervised learning
- ✓ Applications of the Unsupervised learning
- ✓ K-means algorithm
- ✓ K-medoid algorithm
- ✓ Performance metrics
  - Within-cluster sum of squares (WCSS)
  - Silhouette score
  - Davies-Bouldin index
- ✓ Chose optimal number of clusters

Aim of this chapter:

- ✓ Understanding the concepts of the unsupervised learning algorithms, learn practical algorithms to solve classification problem and the evaluate them.

Unsupervised Learning

What is the Unsupervised Learning?

- ✓ In Unsupervised Learning algorithms data have no labels, and the **goal is to discover**:
  - ➢ Patterns
  - ➢ Structures
  - ➢ Relationships in the data
- ✓ The **main objective** is to extract useful information from the data without prior knowledge or assumptions of the underlying patterns of the data.

Unsupervised Learning's Techniques

- ✓ **Clustering** (in this chapter is the focus)**:**
  - ➢ We aim to group similar samples of data together into clusters, (K-means, …).
- ✓ **Dimensionality Reduction:**
  - ➢ We aim to reduce the number of features or independent variables in a dataset, **while** we keep important information as much as possible, (e.g. PCA, LDA, …).
- ✓ **Anomaly Detection:**
  - ➢ We aim to identify the data in a dataset that diverge significantly from expected behavior (**outliers**), (Variational Autoencoder, GANs, …).

Unsupervised Learning - Clustering

What is the Clustering

- ✓ As we mentioned we need to group similar samples together into clusters.
- ✓ Groups are is **based on** similarity or dissimilarity of **samples**.
- ✓ Same as unsupervised learning algorithms objective we want to find underlying structure or patterns in the data.
- ✓ In clustering, formally, we want to partition the data into clusters (groups), using a similarity or distance metric.

Unsupervised Learning - Clustering

Different types of clustering algorithms

- ✓ **Partitioning-based clustering** (in this chapter is the focus)**:**
  - ➢ Partition the data samples into k clusters, (**K-means**, and **K-medoids**).
- ✓ **Density-based clustering:**
  - ➢ Group the samples based on their density in the feature space, instead of a fixed number of clusters or a distance threshold, (DBSCAN, and OPTICS).
- ✓ **Hierarchical clustering:**
  - ➢ Make a hierarchy of nested clusters that each cluster is a subset of a larger cluster, (Agglomerative clustering, and Divisive clustering).

Unsupervised Learning - Clustering

Partitioning-based clustering:

K-means

- ✓ A popular clustering algorithm used in unsupervised machine learning.
- ✓ It group together data into a specified number of clusters.
- ✓ It uses similarity in the feature space.
- ✓ The **core idea** is iteratively assigning examples to the nearest centroid and update centeriod.

K-means

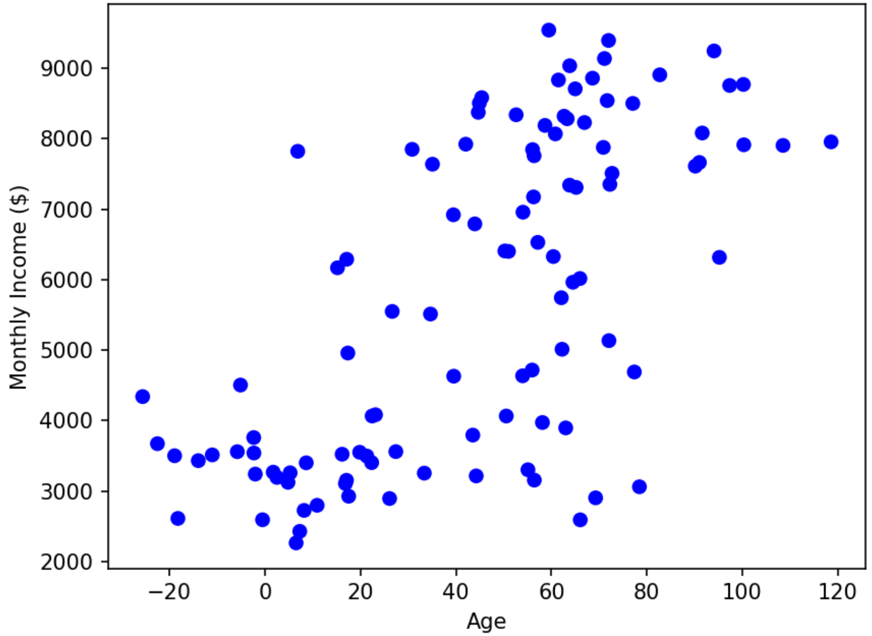
1. Initialize K centroids (randomly).
2. Assign each example to the nearest centroid.
3. Update the centroids.
4. Repeat steps 2-3 until convergence.

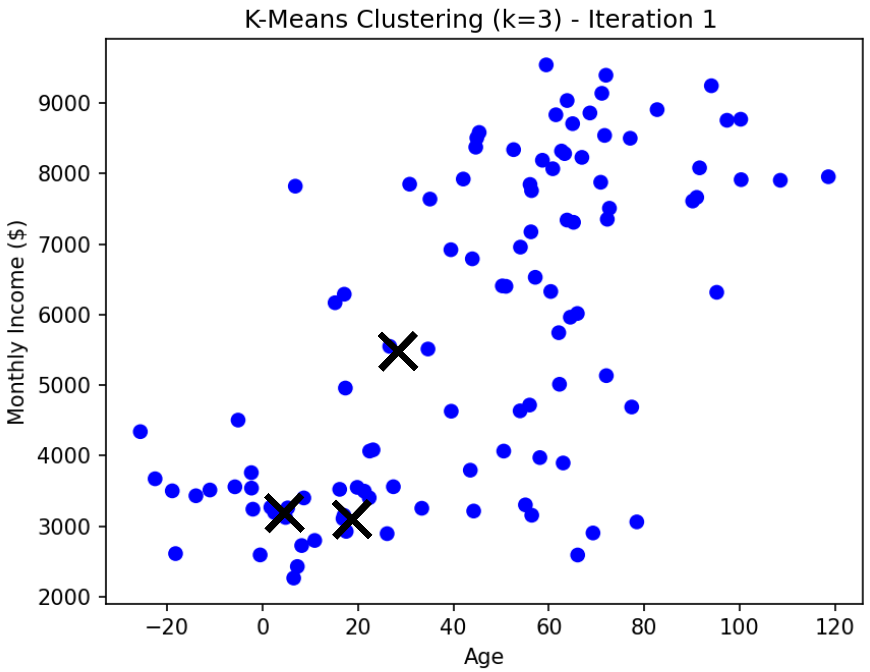
Unsupervised Learning - Partitioning-base Clustering

K-means

1. Initialize K centroids

    ✓ Choose K random **values** for the feature from the dataset as initial centroids.



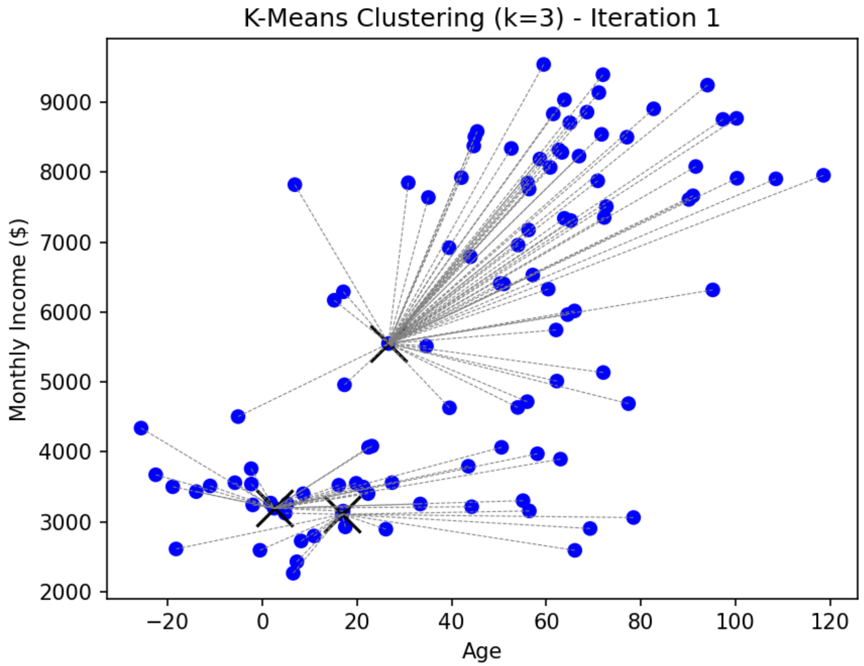


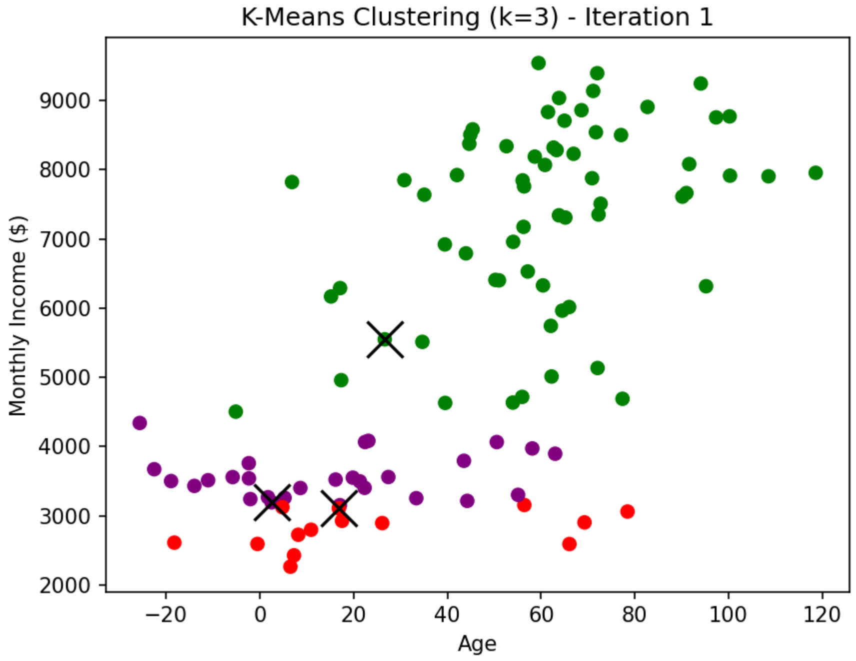
Unsupervised Learning - Partitioning-base Clustering

K-means

2. Assign each example to the nearest centroid

   ✓ Calculate the distance (**commonly Euclidean**) between each centroid and sample.



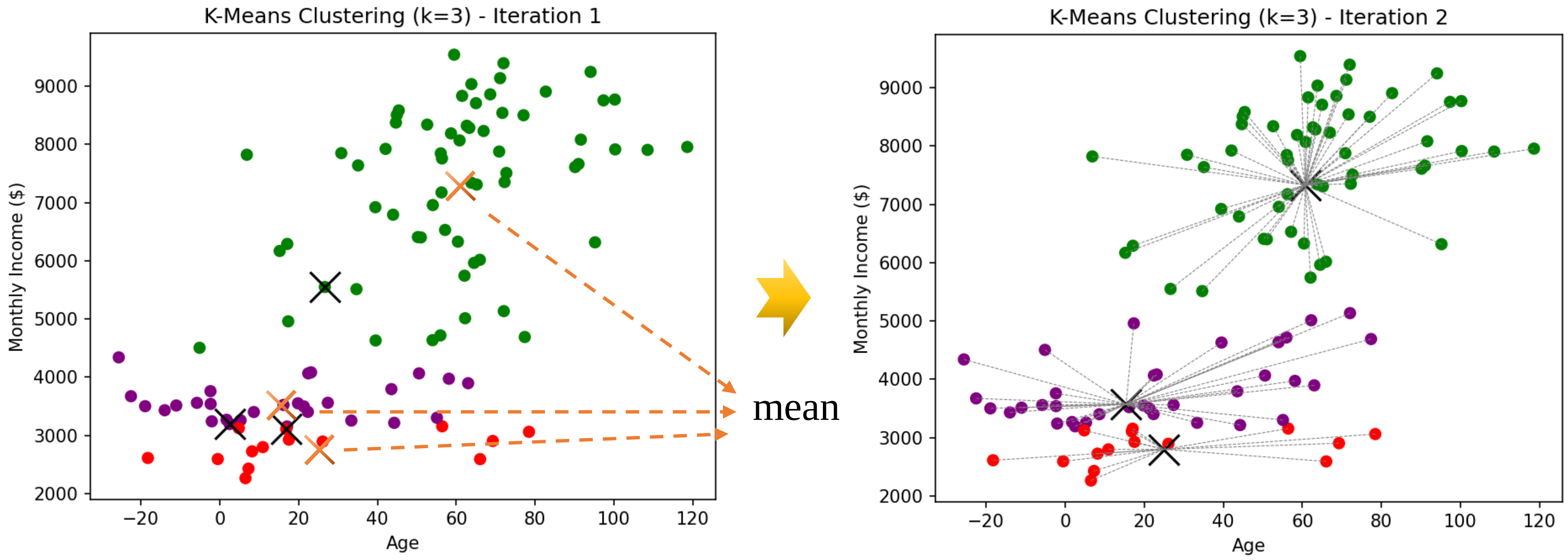


Unsupervised Learning - Partitioning-base Clustering

K-means

3. Update the centroids

   ✓ Calculate the mean of the assigned samples to each centroid

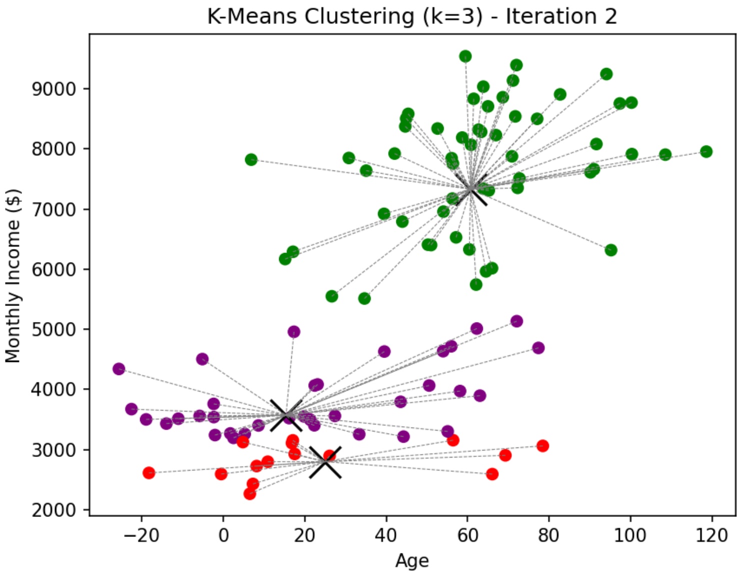


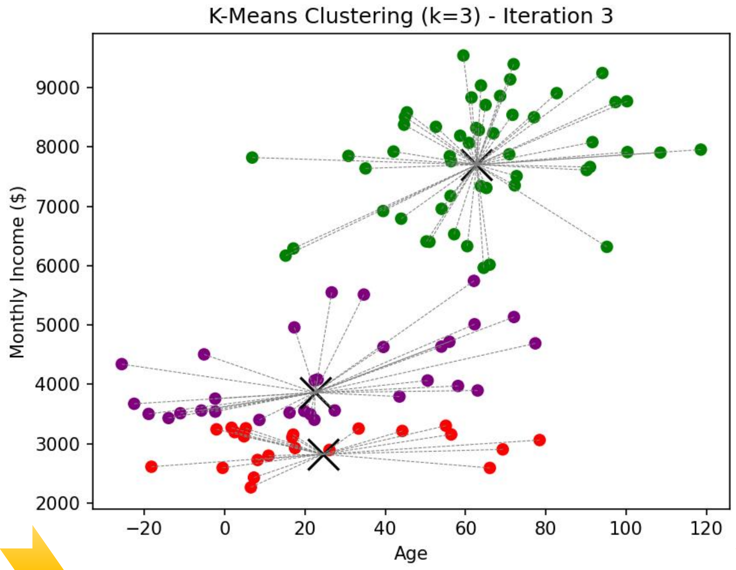
Unsupervised Learning - Partitioning-base Clustering

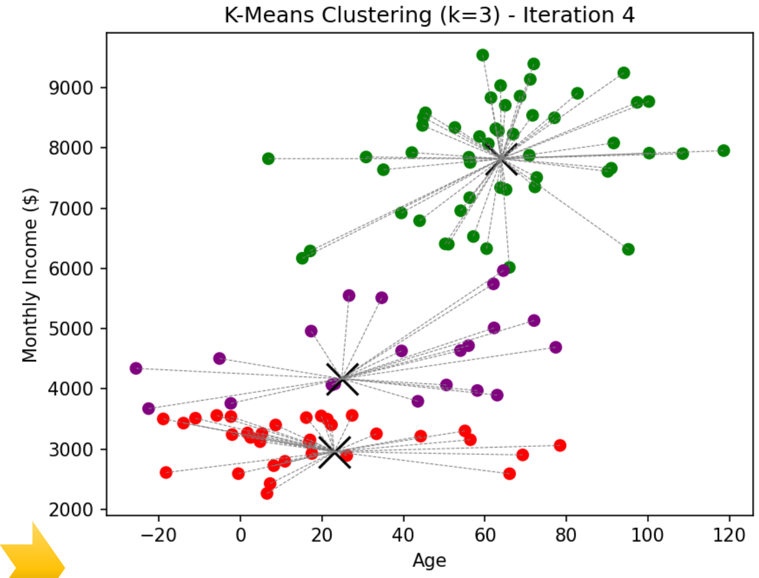
K-means

4. Repeat steps 2-3 until convergence

   - ✓ Until no centroids change or reach the maximum number of iterations.



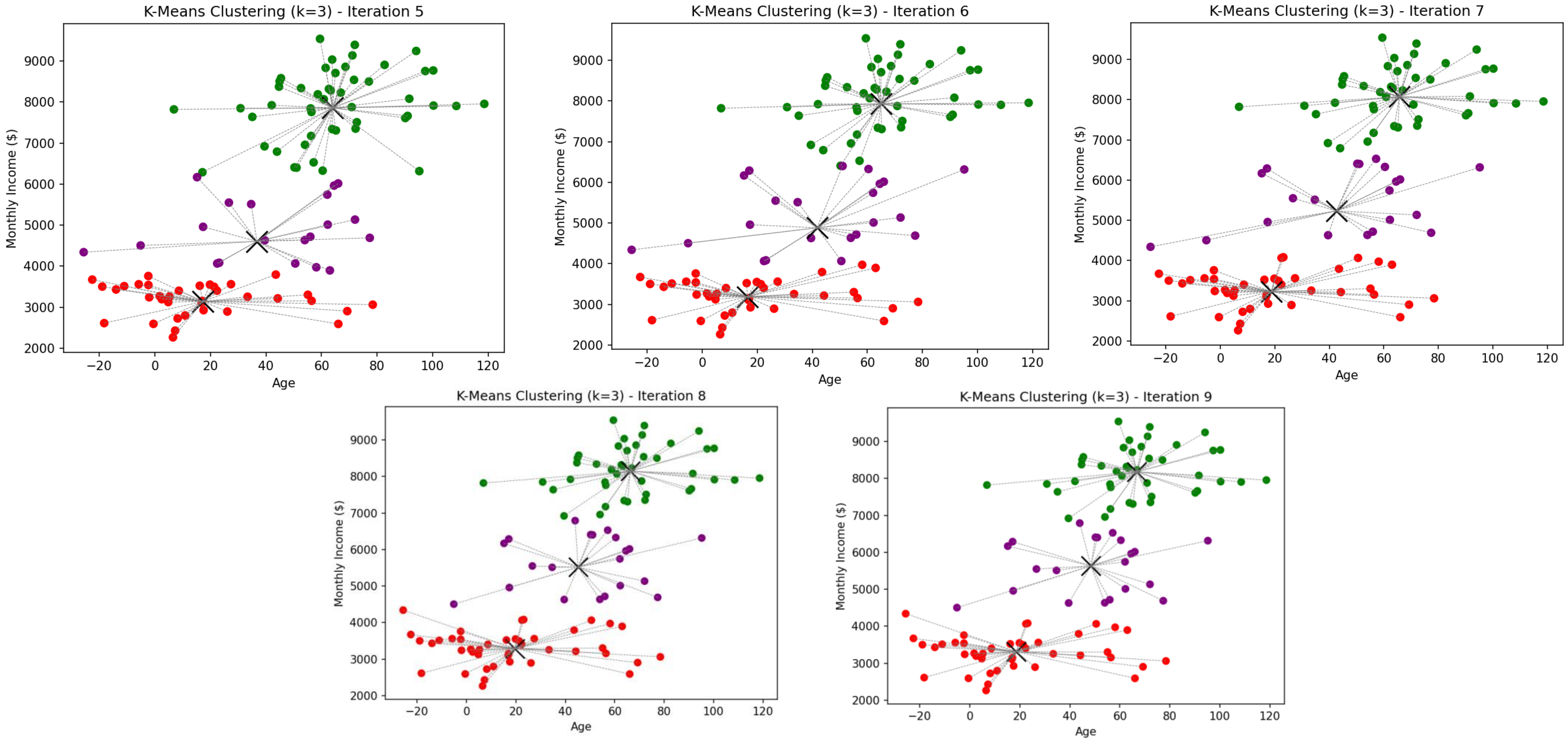




Unsupervised Learning - Partitioning-base Clustering

K-means

We continue Iterations until convergence:



K-medoids algorithm

- ✓ K-medoids is a variant of K-means clustering.
- ✓ Also aims to cluster data into a specified number of clusters.
- ✓ The key difference between K-means and K-medoids is that K-medoids uses **medoids** instead of centroids.

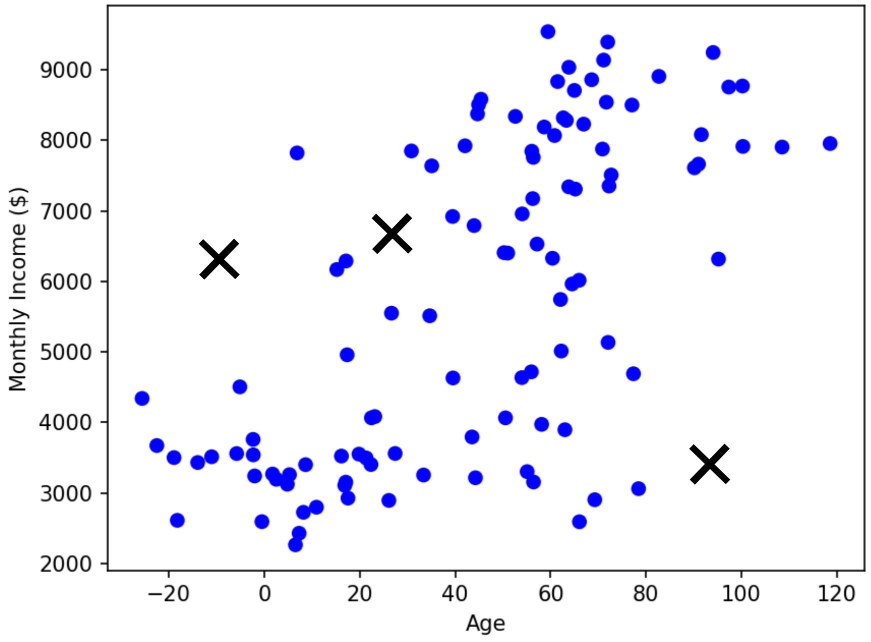
Medoids: After getting average the most close data to average point will be new medoid instead of centroid in K-means.

- ✓ All other steps are the same as K-means.

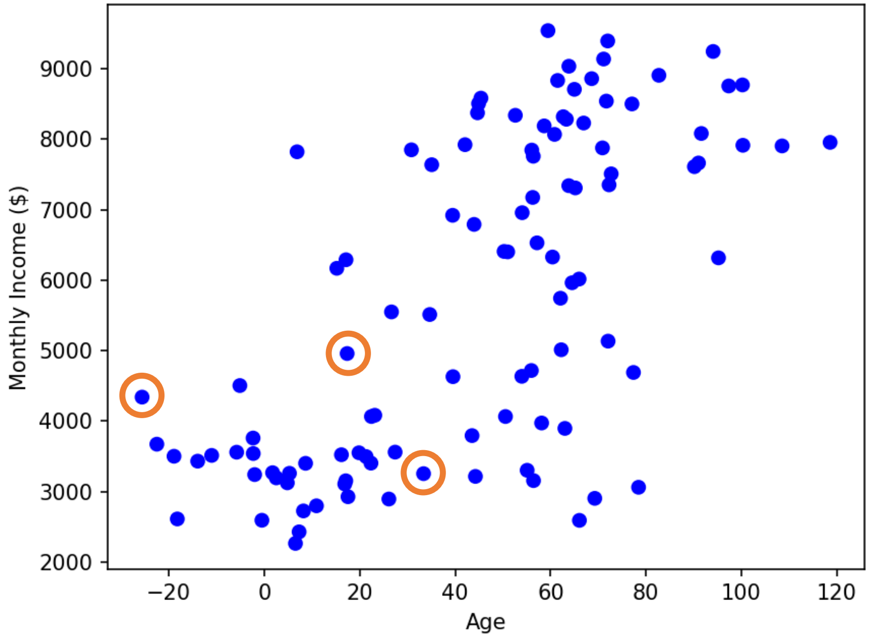
K-medoids

✓ Example for the initialization step example in both algorithms:

K-means



K-medoids



Unsupervised Learning - Partitioning-base Clustering

K-means

Advantages:

- ✓ Simple and computationally efficient.
- ✓ Widely used and well-known.
- ✓ Works well with continuous values for cluster centers.

Disadvantages:

- ✓ Sensitive to initial starting points and it may **stuck** in **local optima.**
- ✓ Cannot work with binary or categorical data.
- ✓ More challenging in handling outliers.

K-medoids

Advantages:

- ✓ Can handle categorical or binary data.
- ✓ Provides better results for clustering accuracy in general.
- ✓ More robust to outliers and noise.

Disadvantages:

- ✓ For large datasets or high-dimensional data can be computationally expensive.
- ✓ It can be more sensitive to the choice of distance metric.

Challenge Both require to specify clusters number in advance.

Unsupervised Learning - Partitioning-base Clustering

Can we use unsupervised learning to do Regression/Prediction?

- ✓ Unsupervised learning techniques are typically used for exploratory data analysis, dimensionality reduction, and clustering tasks.
- ✓ It is not suitable to use unsupervised learning techniques to predict a continuous output variable directly.
- ✓ Unsupervised learning techniques can be used as a preprocessing step for supervised learning tasks, such as regression or classification, (to predict a continuous or categorical output variable).

Unsupervised Learning - Evaluation

Performance Metrics for Unsupervised learning

✓ Different from the metrics used for classification algorithms.

✓ Because in clustering there are no predefined labels or ground truth to compare.

✓ The most **common metrics** used for Unsupervised Learning:

- Within-Cluster Sum of Squares (WCSS)
- Silhouette score
- Davies-Bouldin index
- …

Unsupervised Learning – Evaluation metric

Within-cluster sum of squares (WCSS)

- ✓ With WCSS we measure the sum of the squared distances **between** each cluster center its assigned data points.
- ✓ A lower WCSS indicates better clustering performance.

centroid of cluster

$$WCSS(K) = \sum_{i=1}^{K} \sum_{x \in c_i} ||x - \mu_i||^2$$

i^{th} cluster

Note: useful metric only when used in conjunction with other metrics!

Unsupervised Learning – Evaluation metric

Silhouette score

- ✓ It measures how well-separated the clusters are.
- ✓ It is based on the average distance between points/data in the same cluster and the average distance between points in different clusters.
- ✓ higher silhouette score indicates better clustering performance

Average distance between a data and all data of one next nearest cluster (clearly excluding its own cluster), called dissimilarity.

Average distance between a data to all data in the same cluster.

$$\text{Silhouette score} = \frac{(b - a)}{\max(a, b)}$$

Important Note: We run Silhouette score per each point like this in dataset and get average for all.

To normalize the score to the range [-1, 1]

Unsupervised Learning – Evaluation metric

Davies-Bouldin index

- ✓ It measures the average similarity between each cluster and all other clusters.
- ✓ A lower Davies-Bouldin index indicates better clustering performance.

Average distance between each point in cluster i and its centroid of cluster.

$$DB(K) = \frac{1}{K}\sum_{i=1}^{K} \max_{j \neq i} \frac{(R_i + R_j)}{d_{ij}}$$

Average distance between each point in cluster j and its centroid of cluster.

Distance between the centroids of clusters i and j as next cluster

Unsupervised Learning – Choose best K

How to choose optimal number of clusters?

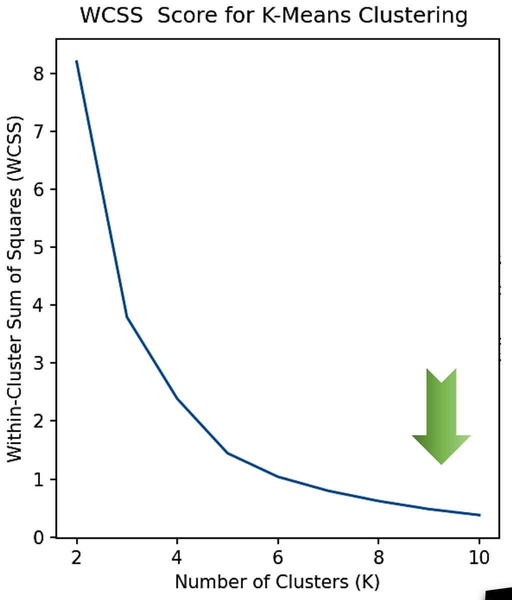
- ✓ We need to **run algorithm** per each cluster and evaluate them first.
- ✓ Then we use elbow point method as the most common technique.

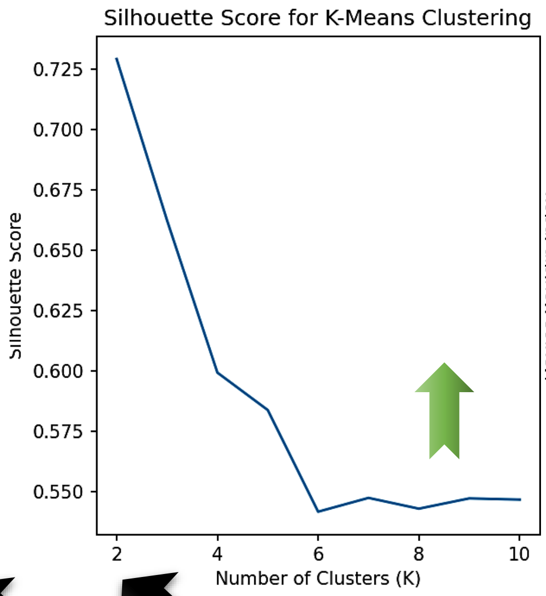
Elbow point method

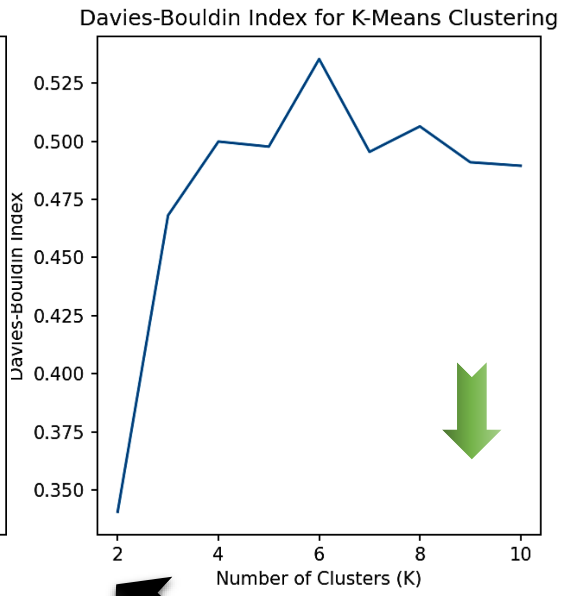
- ✓ **Elbow point** is indicating the k in which WCSS's decrease rate is significant (largest).
- ✓ Usually it forms a bend (**an elbow**) in the WCSS versus k plot.
- ✓ Indicates the point where adding more clusters (k) does not lead to a significant improvement in clustering performance (WCSS).

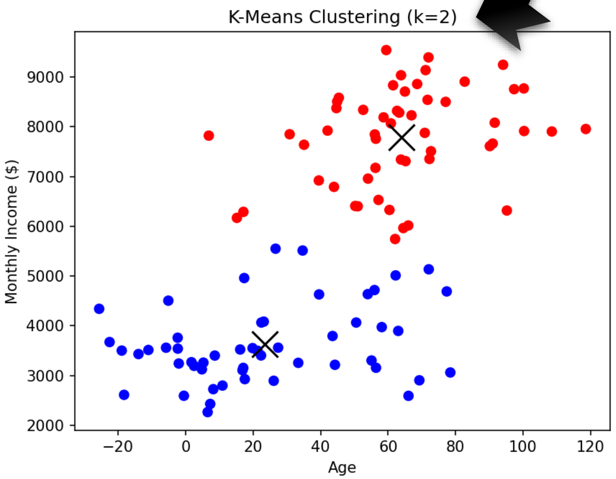
Unsupervised Learning – Choose best K

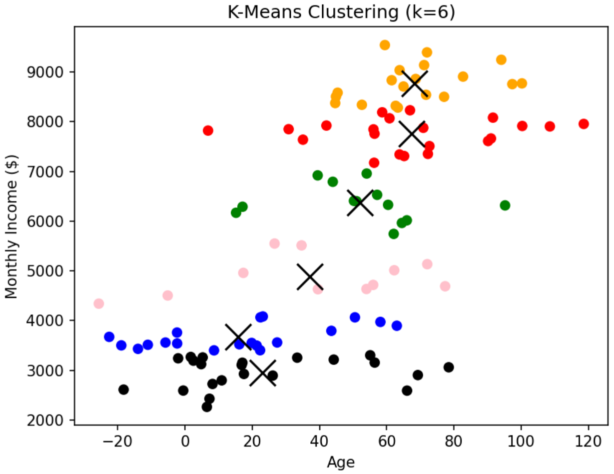
Metrics for different K values









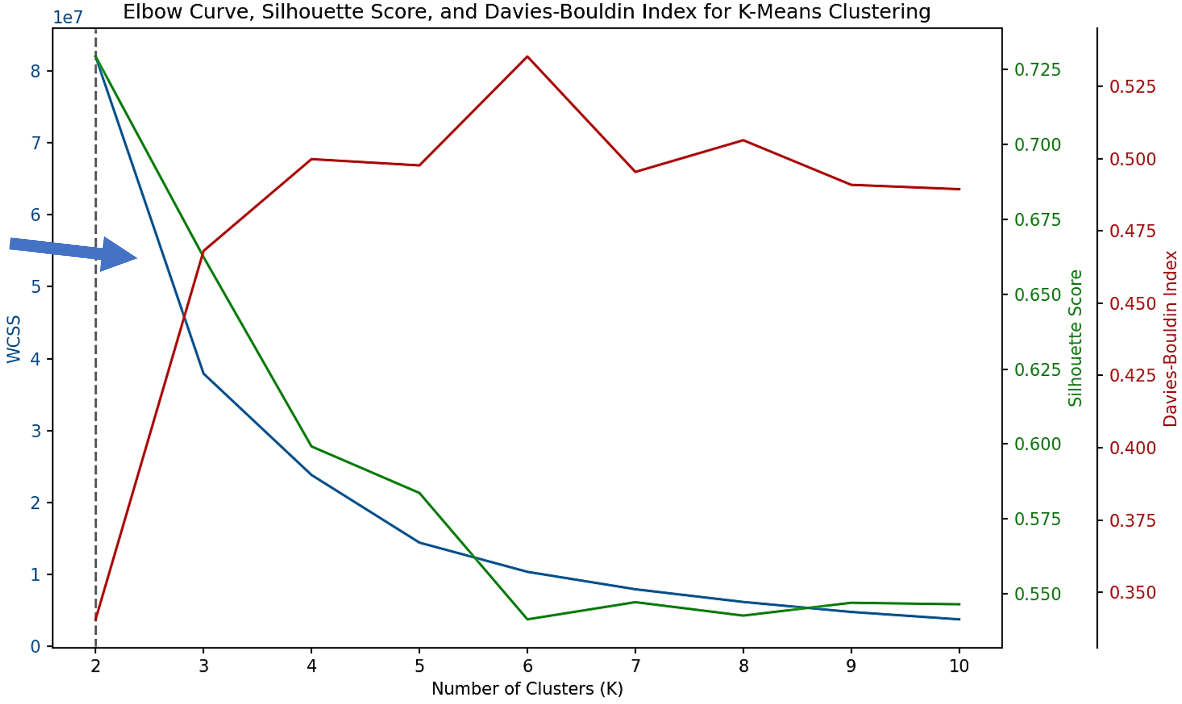


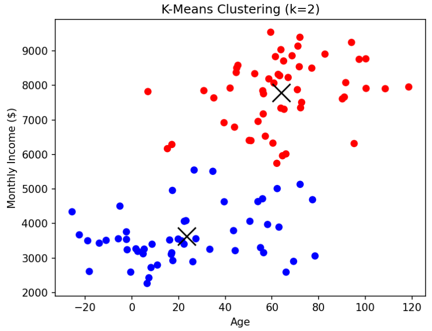
Unsupervised Learning – Choose best K

Elbow point method

- ✓ Bend or an elbow in the WCSS versus k plot with significant improvement.

Significant drop



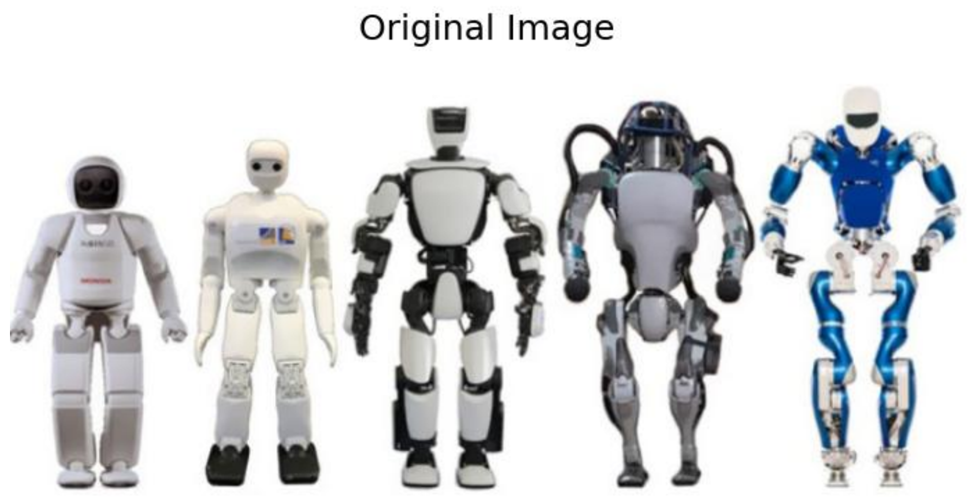


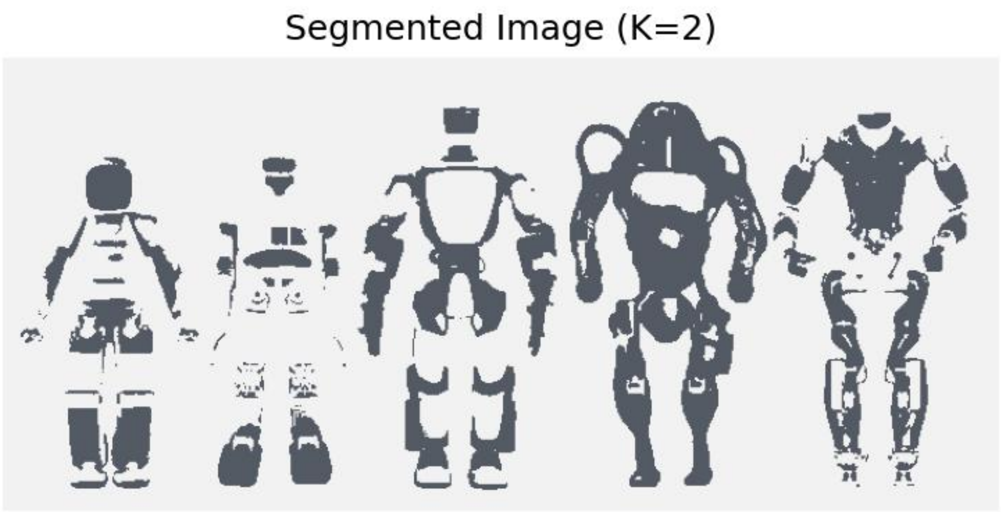
- ✓ **Silhouette score** and **Davies-Bouldin** index can be used as validation elbow method afterward to evaluate the quality of the k.

Unsupervised Learning

Example for Image Segmentation

- ✓ Segmentation Based on color only:





Note: replacement color for each segment here is the average color of all pixels within that cluster, and clusters are based on color similarity.

Unsupervised Learning

Example for Image Segmentation

✓ Segmentation Based on color and pixels position (weighted)

If we run based on color only:

Original Image

Segmented Image (K=3)

Segmented Image (K=3)

Practice

Write K-means without using library for given data than apply elbow point method.

Assignment

-Use K-means algorithm for image segmentation (based on color, and position), and apply elbow method.

Summery

- ✓ We understood the concept of unsupervised learning.
- ✓ We introduced the applications of the unsupervised learning.
- ✓ We discussed K-means algorithm steps with example.
- ✓ We understood K-medoid algorithm.
- ✓ We introduces three performance metrics and the elbow method to choose number of the clusters.