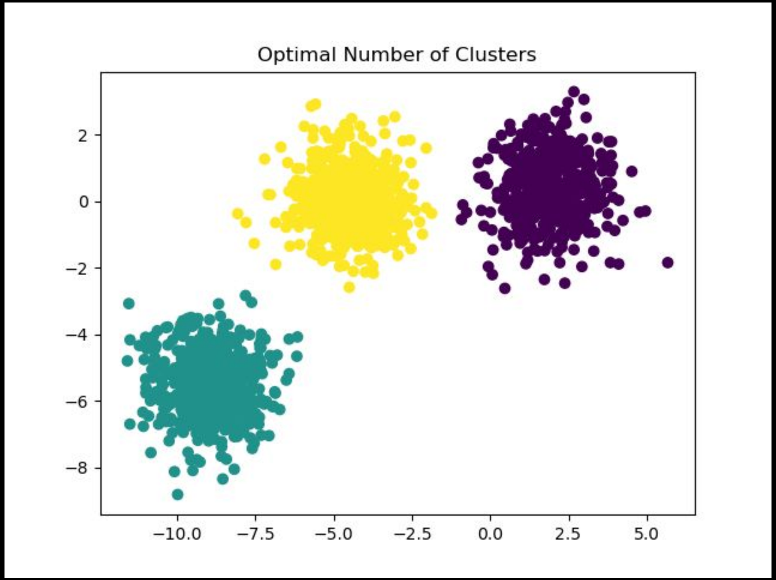
K-Means Clustering: Unveiling Hidden Patterns



Introduction to Unsupervised Learning

What is Unsupervised Learning?

- Learning from unlabeled data.
- Discovering hidden patterns and structures.
- No target variable to predict.

What is K-Means Clustering?

Definition:

- An iterative algorithm that partitions data into K distinct clusters.
- Each data point belongs to the cluster with the nearest mean (centroid).

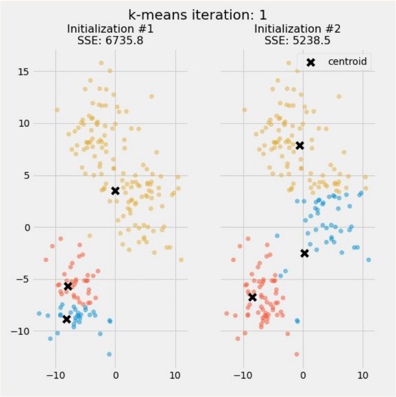
Goal:

- Minimize the within-cluster variance (WCSS) or Sum of Squared Errors (SSE) that is defined as the sum of the squared Euclidean distances.

The K-Means Algorithm: Step-by-Step

Algorithm 1 k-means algorithm

Specify the number k of clusters to assign.
Randomly initialize k centroids.
repeat
  expectation: Assign each point to its closest centroid.
  maximization: Compute the new centroid (mean) of each cluster.
until The centroid positions do not change.



Mathematical Formulation

Distance Metric:

- Typically Euclidean distance: $d(x, \mu_k) = \sqrt{\sum_{i=1}^{n} (x_i - \mu_{ki})^2}$
- Where x is a data point, μ_k is the k-th centroid, and n is the number of dimensions.

Objective Function (WCSS or SSE):

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} d(x_i, \mu_k)^2$$

- Where C_k is the k-th cluster.

Centroid Update:

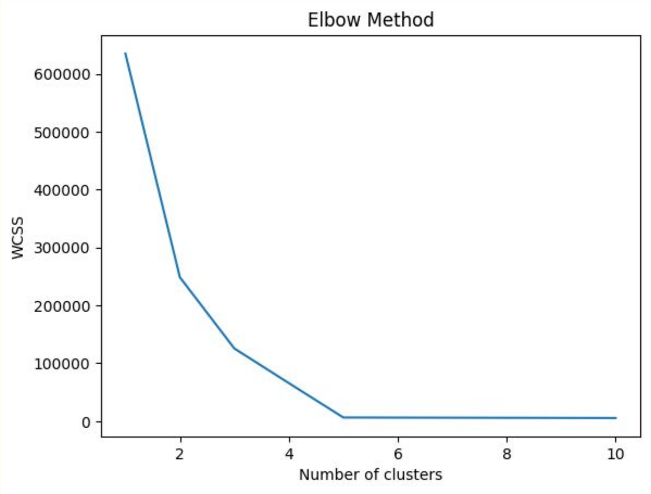
$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i$$

- Where $|C_k|$ is the number of points in cluster C_k.

Determining the Optimal Number of Clusters (K)

Elbow Method:

- Plot the WCSS (inertia) for different values of K.
- Identify the 'elbow' point where the rate of decrease in WCSS slows down.
- This point represents the optimal K.



Performance Metrics

Within-Cluster Sum of Squares (WCSS or SSE) / Inertia:

- Measures the compactness of clusters.
- Lower WCSS indicates tighter clusters.

Silhouette Score:

- Measures how similar an object is to its own cluster compared to other clusters.
- Ranges from -1 to 1: 1 indicates well-separated clusters, -1 indicates misclassification.

Adjusted Rand Index (ARI):

- Unlike the silhouette coefficient, the ARI uses true cluster assignments to measure the similarity between true and predicted labels.
- The ARI output values range between -1 and 1. A score close to 0 indicates random assignments, and a score close to 1 indicates perfectly labeled clusters.

Hyperparameters

Number of Clusters (K):

- The most critical parameter.
- Determines the number of clusters to form.
- Must be chosen carefully, using the elbow method, or silhouette score

Initialization Method (k-means++, random):

- k-means++ helps to select better initial centroids, leading to faster convergence and better results.

Maximum Iterations (max_iter):

- Limits the number of iterations to avoid infinite loops.

Random State (random_state):

- Controls the randomness of the initial centroid choice. By using a constant random state, the results can be reproduced."

Real-World Applications

Customer Segmentation:

- Grouping customers based on purchasing behavior, demographics, etc.

Anomaly Detection:

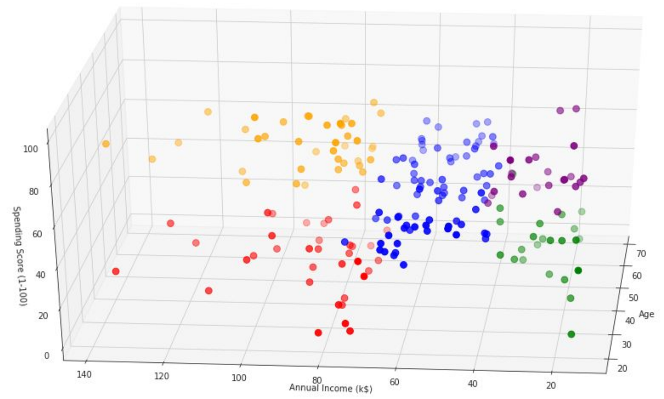
- Identifying unusual data points that deviate from normal patterns.
- Fraud detection, network security.

Document Clustering:

- Grouping similar documents based on their content.
- Topic modeling, information retrieval.

Genetics:

- Grouping genes with similar expression patterns.
- Disease classification.



Advantages and Disadvantages

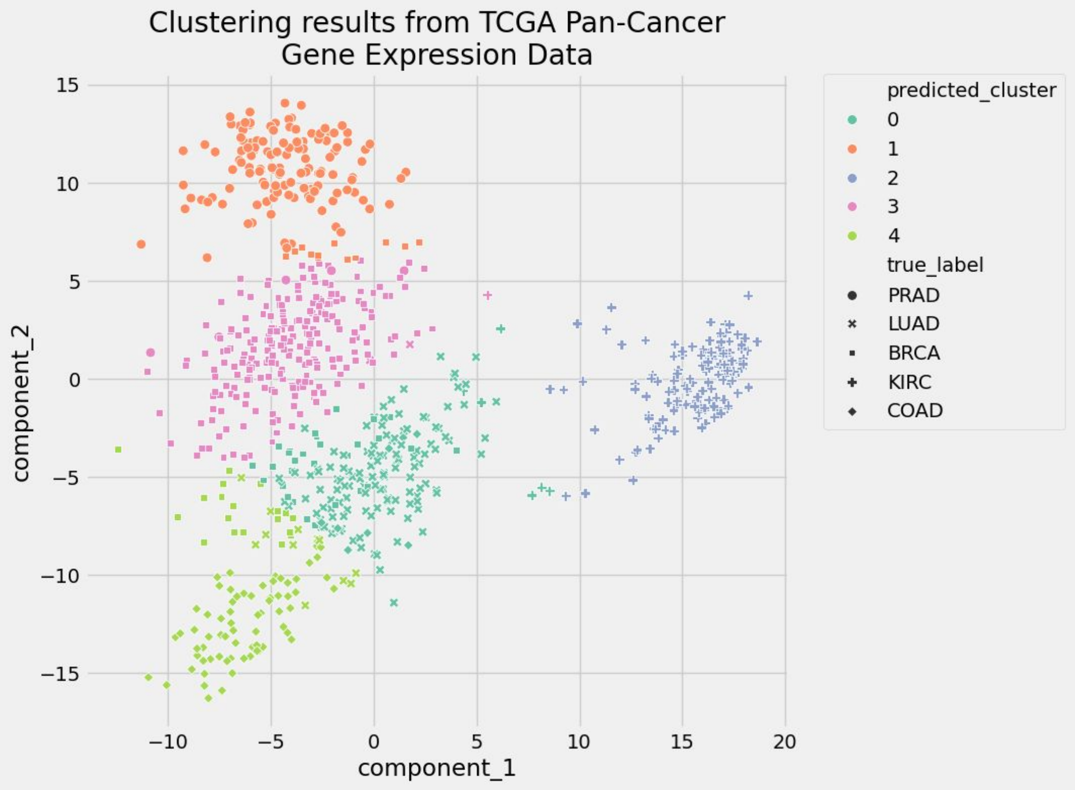
Advantages:

- Simple and easy to implement.
- Efficient for large datasets.
- Relatively scalable

Disadvantages:

- Sensitive to initial centroid selection.
- Assumes spherical clusters.
- Requires pre-specifying K.
- Sensitive to outliers

Coding



[k_means_clustering_programacion_ii.ipynb]()