Indian Institute of Technology Patna

CS564 - Foundation of Machine Learning

Assignment #4: K-Means with elbow method,
Dunn index, Davies Bouldin index,
Silhouette index.
Submission Date: 04th May 2024

Baskar Natarajan - 2403res19(IITP001799)
Jyotisman Kar – 2403res35(IITP001751)
SEMESTER-1
MTECH AI & DSC
INDIAN INSTITUTE OF TECHNOLOGY PATNA.

1. Problem Definition:

Perform k-means clustering on the given dataset “iris.csv”.

1. Apply the elbow method to find out the optimal number of clusters
2. Apply cluster validity Dunn index to find the better cluster.
3. Apply cluster validity Davies Bouldin index find the better cluster.
4. Applying the Silhouette index to find the similarity between the clusters.
5. Plot the cluster distributions for each feature (in 2D form)

2. Introduction:

Clustering analysis is a fundamental technique in unsupervised machine learning used to group data points into clusters based on their similarity. In this documentation, we will explore a clustering analysis performed on the Iris dataset, a classic benchmark dataset in machine learning. We are using K-Means clustering.

3. How K-Means Clustering works?

a. What it does:

- Groups similar data points together based on their features (characteristics).
- Aims to minimize the within-cluster variance (spread of data points) while maximizing the between-cluster variance (separation of clusters).

b. How it works:

7. **Initialization:** Choose the desired number of clusters (k). K-Means randomly selects k data points as initial cluster centroids (centers of the clusters).
8. **Assignment:** Assign each data point to the closest cluster centroid based on a distance metric (e.g., Euclidean distance).
9. **Recalculate Centroids:** Re-compute the centroids by averaging the positions of all data points assigned to each cluster.
10. **Repeat:** Repeat steps 2 and 3 until the centroids no longer change significantly (convergence) or a maximum number of iterations is reached.

c. Key Points:

- **Unsupervised Learning:** K-means doesn't require labeled data (data points with predefined categories). It discovers patterns in the data itself.

- **Predefined Number of Clusters (k):** You need to specify the number of clusters (k) beforehand. Choosing the optimal k is crucial for meaningful results. The elbow method is a common technique to help determine this.
- **Distance Metric:** K-means uses a distance metric (like Euclidean distance) to measure the similarity between data points and centroids.
- **Convergence:** The algorithm iterates until the centroids stabilize, indicating that the clusters have (hopefully) converged to a local optimum.

d. Applications:

- K-means clustering has a wide range of applications, including:
  - Customer segmentation in marketing
  - Image segmentation in computer vision
  - Anomaly detection in fraud analysis
  - Document clustering in information retrieval

e. Advantages:

- Simple and efficient algorithm, especially for large datasets.
- Easy to understand and implement.

f. Disadvantages:

- Requires predefining the number of clusters (k), which can be challenging.
- Sensitive to initial centroid placement. Different initializations can lead to different clustering.
- May not work well for non-spherical clusters (clusters that aren't round or blob-shaped).

4. What is Elbow Method:

- The Elbow Method is a simple but helpful way to get a starting point for choosing the number of clusters in k-means clustering.
- **Calculate WCSS** (Within-Cluster Sum of Squares): For various k values (number of clusters), it calculates the total squared distance between data points and their assigned cluster centroid (center). This measures how spread out the data points are within each cluster.
- **Plot WCSS vs. k**: The results are plotted on a graph with WCSS on the y-axis and k on the x-axis.

- **Identify the "Elbow"**: Look for an "elbow" shape in the curve. This point represents the value of k where the WCSS starts to decrease slowly, indicating that adding more clusters doesn't significantly improve the clustering

5. Cluster Validity Metrics

1. Cluster validity metrics are quantitative measures used to assess the quality of clustering results in unsupervised learning.
2. They help you evaluate how well your chosen number of clusters (k) and the resulting cluster assignments represent the inherent structure in your data.

1. Dunn Index:

a. Computes the Dunn Index, which measures the compactness of clusters and separation between them. Selects the optimal number of clusters based on the maximum Dunn Index.
b. The Dunn Index is calculated as the ratio of the minimum inter-cluster distance (smallest distance between any two cluster centroids) to the maximum intra-cluster distance (largest distance between any two data points within a single cluster).
c. `Dunn Index = (min inter-cluster distance) / (max intra-cluster distance)`
d. Higher Dunn Index: Indicates better clustering.
e. Lower Dunn Index: Suggests clusters might be overlapping or not very distinct.
f. The Dunn Index ranges from 0 to positive infinity.

2. Davies Bouldin Index:

a. Calculates the Davies Bouldin Index, which quantifies the average similarity between each cluster and its most similar cluster. Chooses the optimal number of clusters based on the minimum Davies Bouldin Index.
b. DBI calculates the ratio of the within-cluster scatter (spread) to the between-cluster distance.
c. It aims for:
    i. Low Within-Cluster Scatter: Data points within a cluster should be close to each other, indicating compactness.
    ii. High Between-Cluster Distance: Clusters should be well-separated from each other.
d. DBI = Average (Within-Cluster Scatter_i / Between-Cluster Distance_(i, nearest_cluster)) for all clusters i

e. Lower DBI Value: Indicates better clustering.
f. Higher DBI Value: Suggests potential issues with the clustering.

3. Silhouette Index:

a. Computes the Silhouette Index, a measure of how similar an object is to its own cluster compared to other clusters. Determines the optimal number of clusters based on the maximum Silhouette Index.
b. Assess how well each data point is assigned to its cluster. It considers both the cohesion (similarity) within a cluster and the separation between clusters.
c. **a(i):** Average distance between 'i' and other data points in its assigned cluster. This reflects the cohesion within the cluster.
d. **b(i):** Average distance between 'i' and the closest data points in a different cluster. This reflects the separation between clusters.
e. **Calculation:**
    i. The Silhouette Index (S(i)) for data point 'i' is calculated as:
    ii. `S(i) = (b(i) - a(i)) / max(a(i), b(i))`
f. **S(i) close to 1:** Excellent choice.
g. **S(i) close to 0:** Indicates potential issues.
h. **S(i) close to -1:** Bad assignment.

6. Libraries Used:

1. **Pandas**: Utilized for data manipulation and analysis, particularly for loading the dataset and preprocessing.
2. **NumPy**: Essential for numerical operations and array manipulations, used extensively for mathematical computations.
3. **Matplotlib** and **Seaborn**: These visualization libraries are crucial for generating plots and graphs to visualize the data and clustering results.
4. **Scikit-learn**: A powerful machine learning library providing various algorithms and tools for clustering, evaluation metrics, and preprocessing.
5. **SciPy**: Used for scientific computing and providing additional functionality for clustering analysis.

7. Pairplot Visualization:

- Generates pairplots for each clustering method to visually inspect the clusters formed in the feature space.

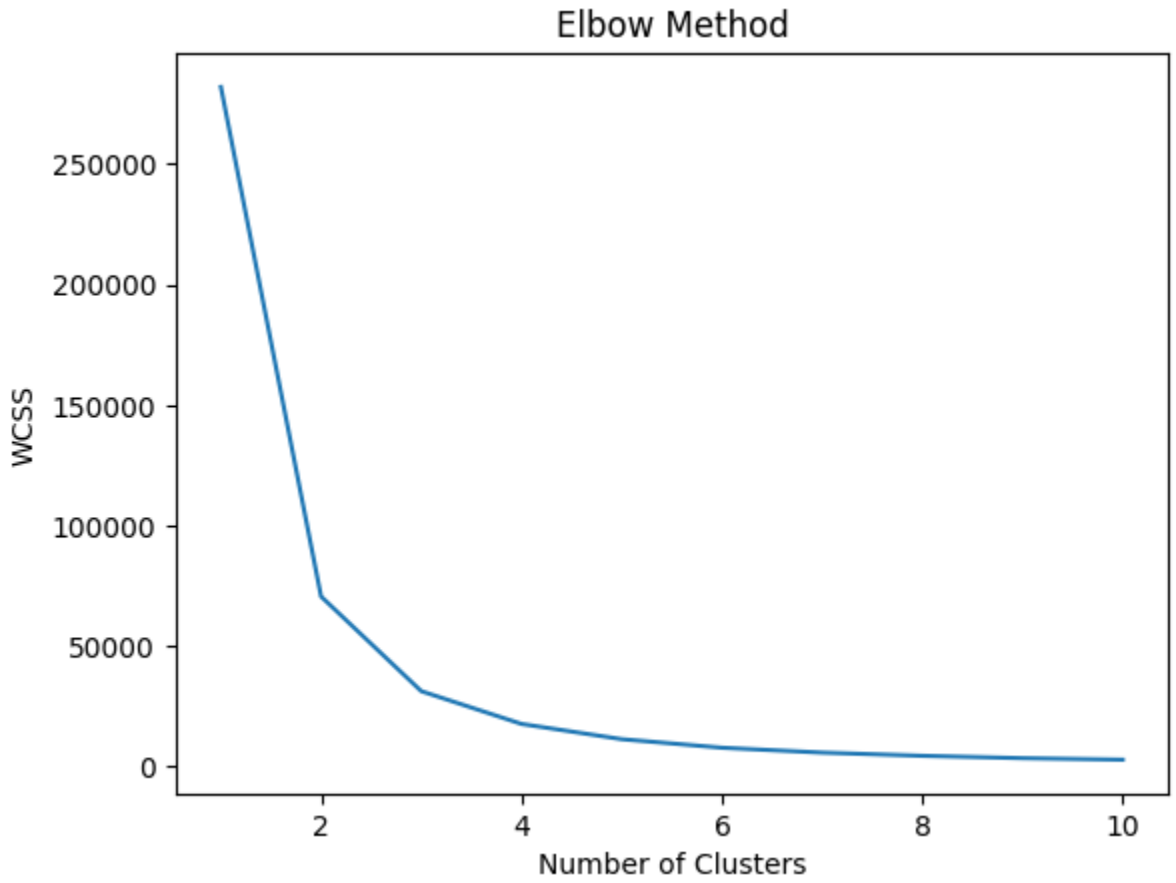
8. Best Clustering calculation Method:

- Compares silhouette scores of different clustering methods to determine the best one. Selects the method with the highest silhouette score and prints the optimal number of clusters.
- Conclusion:

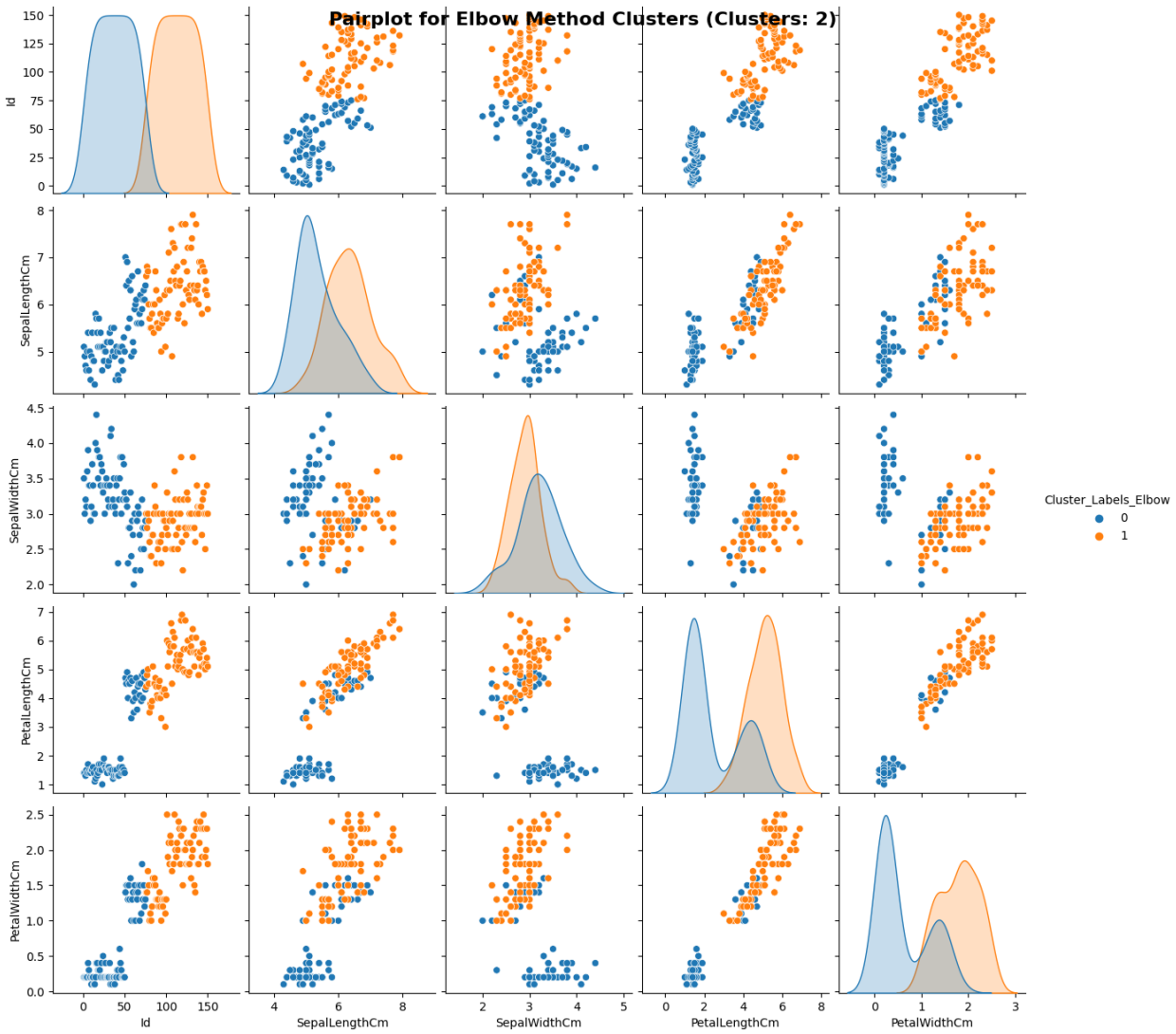
9. Conclusion

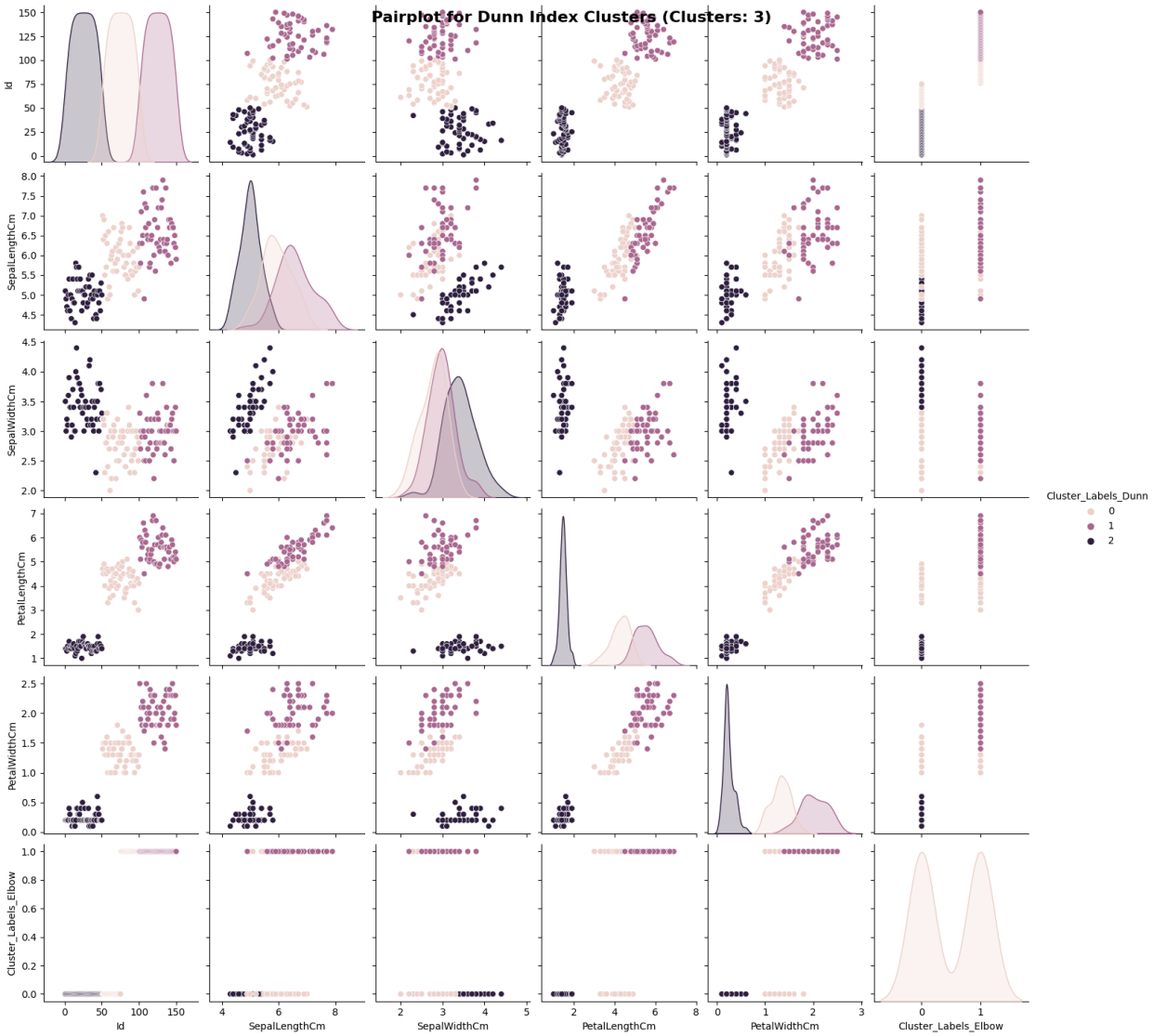
- clustering analysis on the Iris dataset demonstrates the application of various methods for determining the optimal number of clusters.
- Each method offers unique insights into the clustering structure of the data, allowing for a comprehensive evaluation of clustering quality.
- By comparing the results obtained from different methods, researchers and practitioners can make informed decisions when selecting the most suitable clustering approach for their specific dataset and objectives.

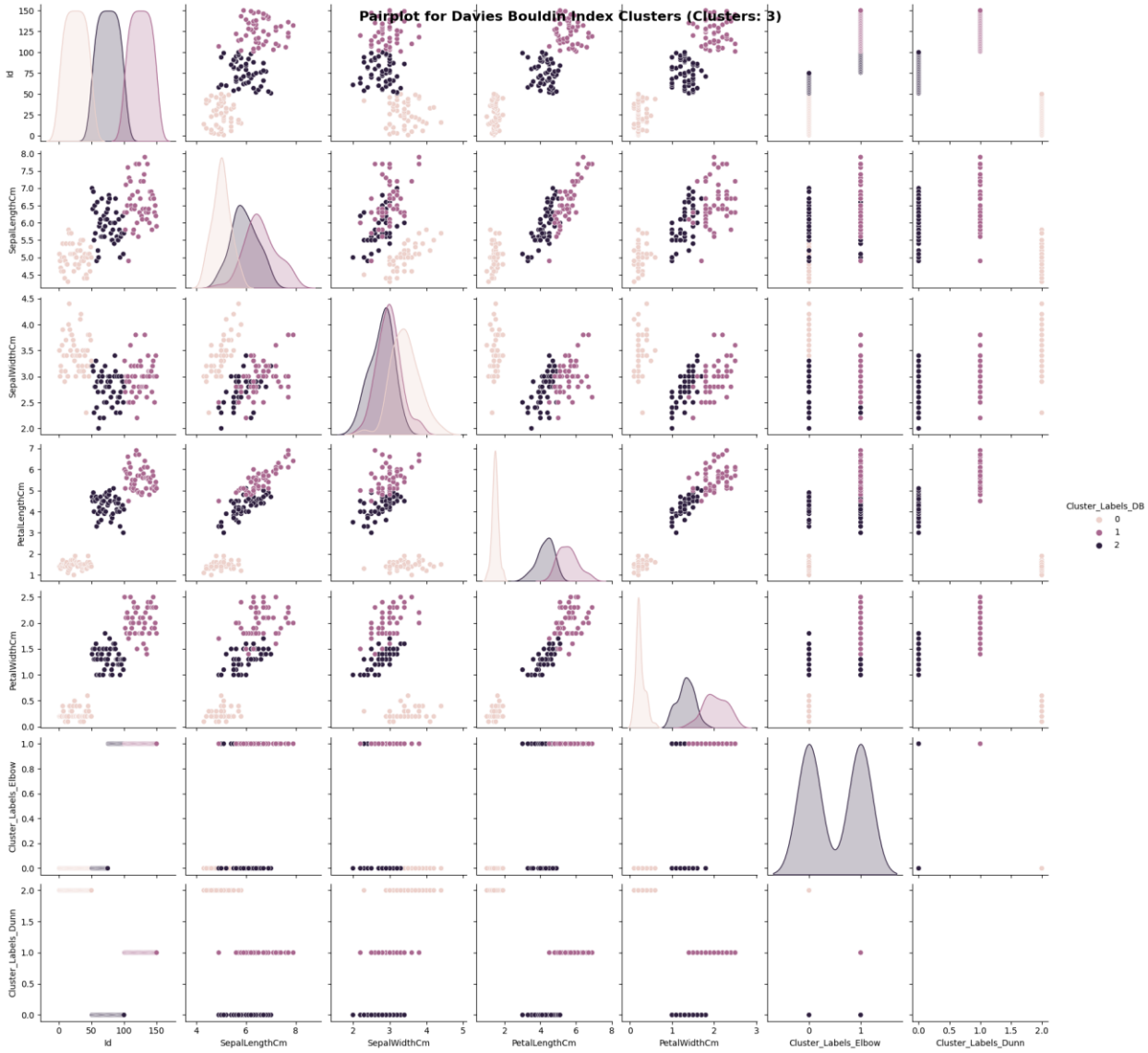
10. OUTPUT



Optimal clusters using Dunn Index: 3
Optimal clusters using Davies Bouldin Index: 3
Optimal clusters using Silhouette Score: 2


Pairplot for Elbow Method Clusters (Clusters: 2)
Id
SepalLengthCm
SepalWidthCm
PetalLengthCm
PetalWidthCm
Cluster_Labels_Elbow
0
1


Pairplot for Dunn Index Clusters (Clusters: 3)
Id
SepalLengthCm
SepalWidthCm
PetalLengthCm
PetalWidthCm
Cluster_Labels_Elbow
Cluster_Labels_Dunn
0
1
2



Silhouette Scores:

Elbow Method: 0.6204656046551029

Dunn Index: 0.5821934246576435

Davies Bouldin Index: 0.5821934246576435

Silhouette Index: 0.6204656046551029

Best Clustering Method: Elbow Method

Optimal Clusters for Best Method: 2