Ml clustering assignent 3 Report

Nihith Nath Kandikattu (50537232)

3. Clustering Algorithms answer

In the report explain the assumptions, advantages and disadvantages of each algorithm.

Clustering Algorithms:

i. K-means Clustering:

- **Assumptions:** K-means assumes that the clusters are spherical and have equal variance. It also assumes that the data points within a cluster are closer to the centroid of that cluster than to centroids of other clusters.
- **Advantages:**
  - Simple and easy to implement.
  - Fast and efficient, making it suitable for large datasets.
  - Works well when clusters are well-separated and have a spherical shape.
- **Disadvantages:**
  - Requires specifying the number of clusters (k) beforehand, which may not always be known.
  - Sensitive to the initial placement of centroids and can converge to local optima.
  - Doesn't perform well with clusters of varying sizes or non-linearly separable data.

ii. Hierarchical (Agglomerative) Clustering:

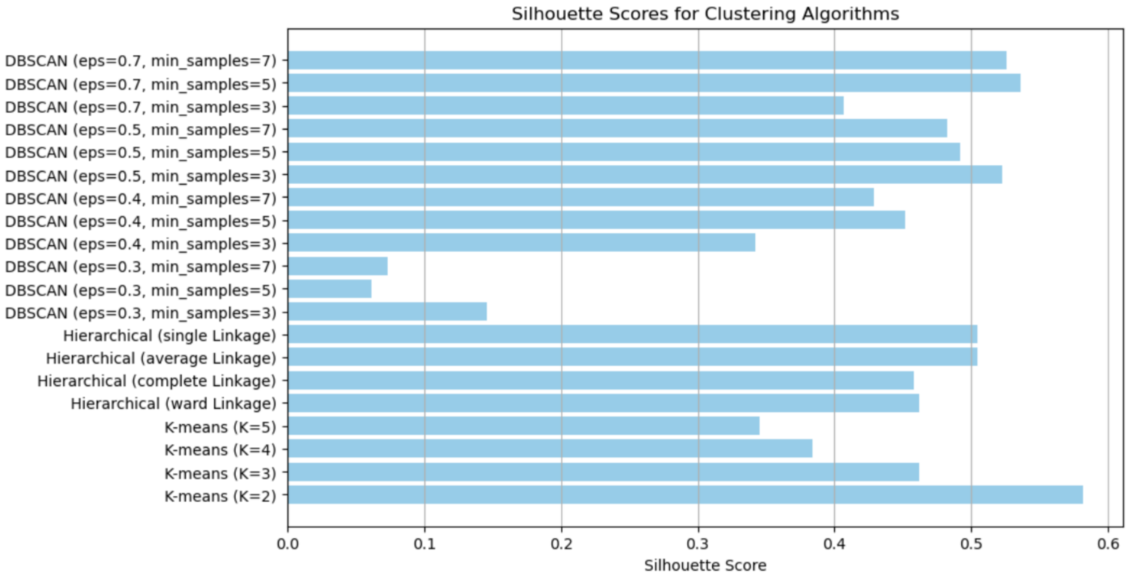
- **Assumptions:** Hierarchical clustering does not assume any particular shape of clusters. It builds a hierarchy of clusters by recursively merging smaller clusters into larger ones based on their similarity.
- **Advantages:**
  - Does not require specifying the number of clusters beforehand, as it builds a dendrogram that can be cut at different levels to obtain different numbers of clusters.

    - Captures the hierarchical structure of the data, allowing for insights at multiple scales.
- **Disadvantages:**
    - Can be computationally expensive, especially for large datasets, as it requires computing pairwise distances between all data points.
    - May not perform well with non-Euclidean distances or when clusters have complex shapes.

iii. DBSCAN (Density-Based Spatial Clustering of Applications with Noise):

- **Assumptions:** DBSCAN assumes that clusters are dense regions separated by low-density areas. It does not require specifying the number of clusters beforehand and can find clusters of arbitrary shapes.
- **Advantages:**
    - Does not require specifying the number of clusters beforehand and can automatically detect clusters of arbitrary shapes and sizes.
    - Robust to noise and outliers, as it classifies points as noise if they are not within a dense region.
- **Disadvantages:**
    - Sensitive to the choice of distance metric and the parameters epsilon (ε) and minimum points (MinPts).
    - May struggle with datasets of varying densities or with clusters of varying densities.

4 . Evaluation :



K-means Clustering: This method shows a decreasing trend in silhouette scores as the number of clusters increases. With 2 clusters, the score is highest at approximately 0.58, suggesting that two broad clusters have been effectively distinguished. However, as we consider the actual species count in the Iris dataset (which is three), the silhouette score for 3 clusters (approximately 0.46) appears to be more biologically relevant, albeit with a slight compromise in the score.

Hierarchical Clustering: The scores for hierarchical clustering with various linkage methods are relatively close, ranging from approximately 0.45 to 0.50. Average linkage shows a higher score than Ward and Complete, but Single linkage reports the highest score at around 0.50. Despite this, visual analyses often reveal that Single linkage may produce chaining effects that are less representative of actual data patterns, suggesting that Average linkage, with its balance between score and visual coherence, might provide a more reliable clustering structure for the Iris dataset.

DBSCAN: A variety of parameters have been used for DBSCAN, showing a wide range of silhouette scores. Generally, scores improve with higher eps values and lower min_samples values, indicating larger cluster sizes. DBSCAN's performance is sensitive to these parameters, and the optimal settings seem to be eps=0.4 and min_samples=5 which offer the best silhouette score among the DBSCAN configurations presented.

5-7 .Observations and findings:

K MEANS CLUSTERING

Observations:

1. The image displays the results of k-means clustering on the IRIS dataset, with varying values of k (2, 3, 4, and 5 clusters).

2. The scatter plots show the data points colored according to their assigned cluster labels, with the centroids represented by different colored markers.

3. The bar charts on the right illustrate the counts or sizes of the identified clusters for each value of k.

Evaluating clustering quality using silhouette score:

The silhouette scores for different values of k are provided:

- k=2: 0.5818 (good clustering)

- k=3: 0.4618 (reasonable clustering)

- k=4: 0.3839 (borderline/mediocre clustering)

- k=5: 0.3455 (mediocre/poor clustering)

Based on the silhouette scores, the k-means clustering with k=2 performs the best, achieving a score above 0.5, which indicates good clustering quality. The scores decrease as the number of clusters increases, suggesting that the higher values of k may not be optimal for this dataset.

Analysis and cluster characteristics:

1. k=2: The scatter plot shows two distinct clusters (purple and yellow), with a clear separation between them. The cluster sizes are relatively balanced, as seen in the bar chart. This suggests the presence of two well-defined groups in the data.

2. k=3: Three clusters are identified (blue, green, and purple), with the blue and green clusters appearing closer to each other compared to the purple cluster. The cluster sizes are slightly more uneven, with the purple cluster being the largest.

3. k=4: Four clusters are identified (blue, green, purple, and yellow), with the yellow cluster being relatively small and separated from the other three clusters. The cluster sizes are quite uneven, with the purple cluster being the largest and the yellow cluster being the smallest.

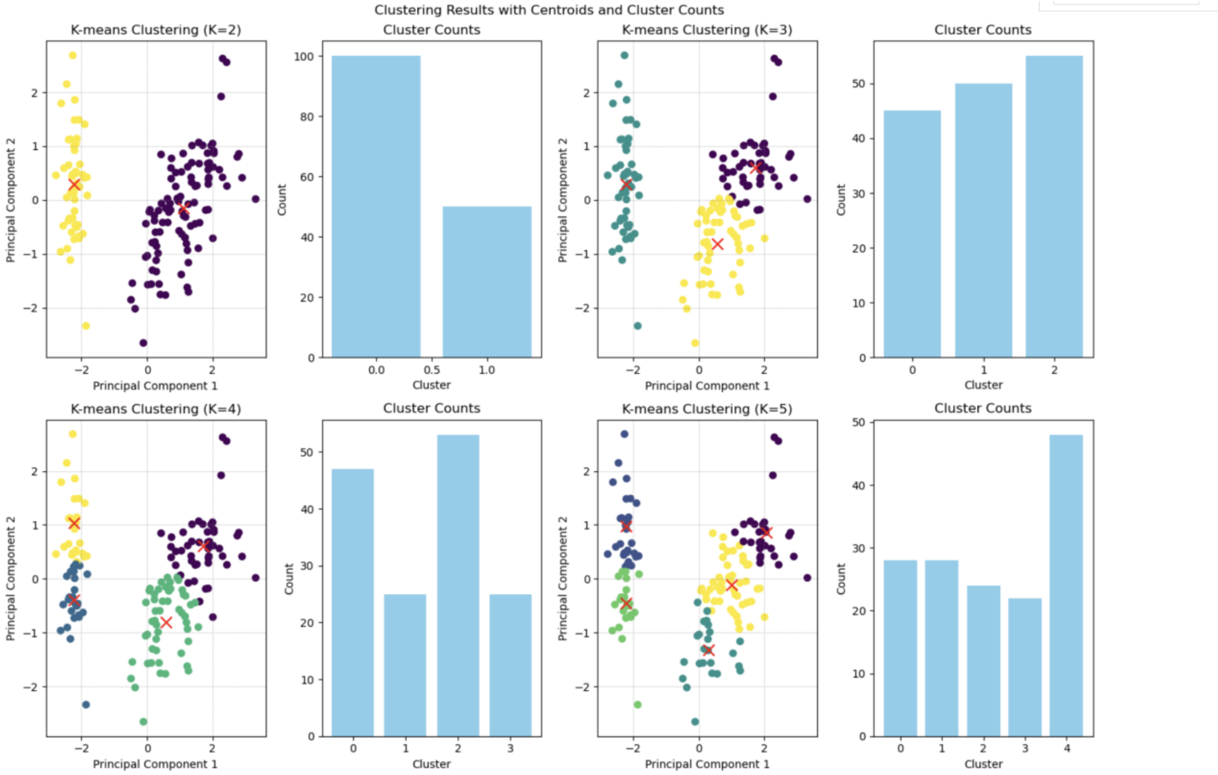
4. k=5: Five clusters are identified (blue, purple, green, yellow, and red), with the red cluster being very small and potentially representing outliers or noise in the data. The cluster sizes are highly uneven, with the purple cluster being the largest and the red cluster being the smallest.

From the visualizations and silhouette scores, it appears that the IRIS dataset may have an underlying structure with two or three distinct clusters. The k=2 solution provides a good separation of the data into two compact and well-defined clusters, while the k=3 solution introduces a third cluster that could potentially represent a subcategory or outlier group.

As the number of clusters increases beyond k=3, the clustering quality deteriorates, and the identified clusters become less balanced and less well-separated. This suggests that higher values of k may be overfitting the data or capturing noise rather than meaningful patterns.

Overall, based on the silhouette scores and visual inspection, the k=2 or k=3 solutions appear to be the most suitable for the IRIS dataset, capturing the underlying cluster structure effectively.

```
K-means Clustering Silhouette Scores:
K-means (K=2)           : 0.5818
K-means (K=3)           : 0.4618
K-means (K=4)           : 0.3839
K-means (K=5)           : 0.3455
```



Hierarchical clustering :

Observations:

1) The image shows the results of applying different hierarchical clustering methods (ward linkage, complete linkage, average linkage, and single linkage) on the IRIS dataset.
2) The scatter plots on the left display the data points colored according to their assigned cluster labels, while the bar charts on the right show the sizes of the identified clusters.

3) The ward linkage method seems to have identified three distinct clusters, as indicated by the three different colors (purple, green, and yellow) in the scatter plot.
4) The complete linkage method also appears to have identified three clusters, but with a slightly different distribution of data points compared to the ward linkage method.
5) The average linkage and single linkage methods have identified two main clusters, with one large cluster (blue) and a smaller cluster (purple and yellow points combined).
6) The cluster size distributions vary across the different linkage methods, with the ward and complete linkage methods producing more evenly sized clusters, while the average and single linkage methods result in one significantly larger cluster and one or two smaller clusters.

Results:

Hierarchical (ward Linkage) : 0.4616
Hierarchical (complete Linkage) : 0.4579
Hierarchical (average Linkage) : 0.5046
Hierarchical (single Linkage) : 0.5046

Evaluating clustering quality using silhouette score:

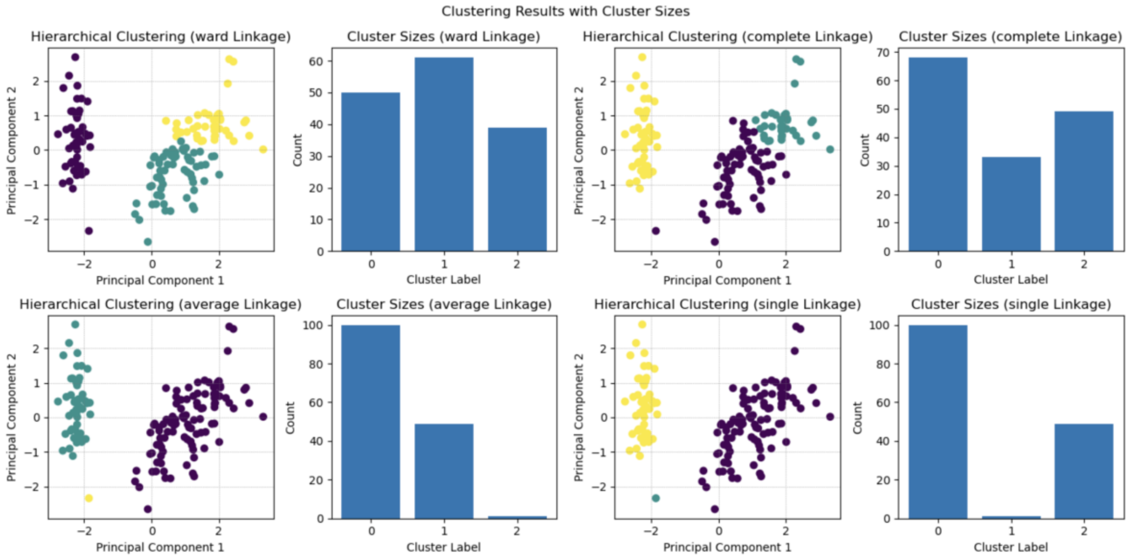
The silhouette score is a metric that measures the quality of cluster assignments, ranging from -1 to 1, with higher values indicating better-defined and well-separated clusters. To evaluate the clustering quality, we would need to compute the silhouette score for each method and compare the values.

Generally, a silhouette score above 0.5 is considered good clustering, while values below 0.2 indicate poor clustering. Without the actual silhouette score values, it is difficult to quantitatively assess the clustering quality, but we can make some qualitative observations based on the visualizations.

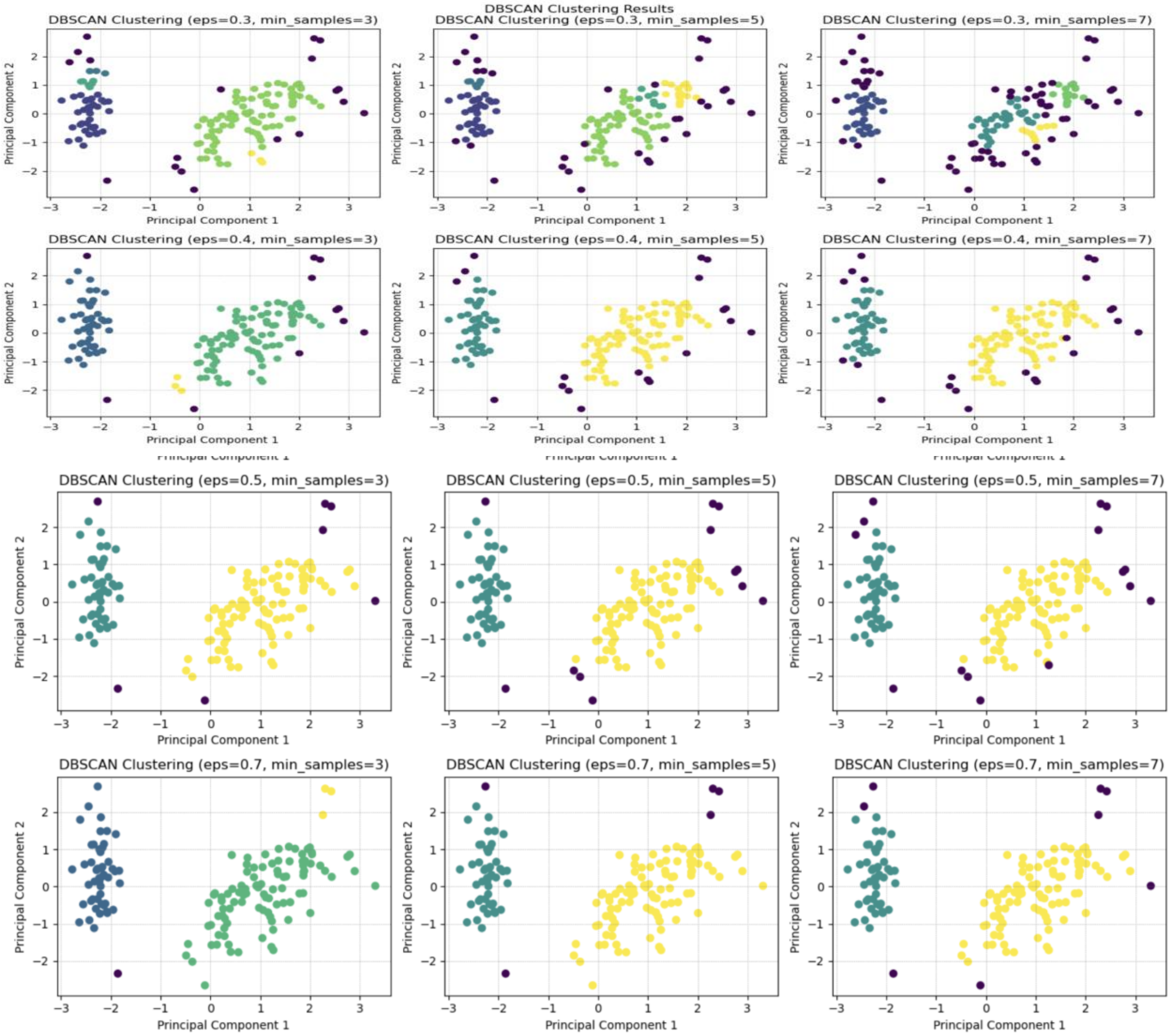
Analysis and cluster characteristics:

7) Ward Linkage: This method seems to have produced relatively well-separated and compact clusters, with the green and purple clusters appearing quite distinct from the yellow cluster. The cluster sizes are roughly comparable, suggesting a reasonably balanced partitioning of the data.

8) Complete Linkage: Similar to ward linkage, this method has identified three distinct clusters, although the distribution of data points within each cluster appears slightly different. The cluster sizes are also more uneven compared to ward linkage.
9) Average Linkage and Single Linkage: These methods have identified two main clusters, with one large cluster (blue) and a smaller cluster (purple and yellow points combined). This suggests that the data may have an underlying structure with one dominant group and a smaller, potentially outlier or subcategory group.
10) Overall, the ward linkage and complete linkage methods appear to have captured more granular substructures within the data, while the average and single linkage methods have found a broader, coarser partitioning.



DBSCAN Clustering Results:



Parameter Settings and Silhouette Scores :

DBSCAN Clustering Silhouette Scores:

DBSCAN (eps=0.3, min_samples=3)	: 0.1460
DBSCAN (eps=0.3, min_samples=5)	: 0.0613
DBSCAN (eps=0.3, min_samples=7)	: 0.0733
DBSCAN (eps=0.4, min_samples=3)	: 0.3421
DBSCAN (eps=0.4, min_samples=5)	: 0.4514

```
DBSCAN (eps=0.4, min_samples=7)          : 0.4289
DBSCAN (eps=0.5, min_samples=3)          : 0.5226
DBSCAN (eps=0.5, min_samples=5)          : 0.4917
DBSCAN (eps=0.5, min_samples=7)          : 0.4827
DBSCAN (eps=0.7, min_samples=3)          : 0.4067
DBSCAN (eps=0.7, min_samples=5)          : 0.5361
DBSCAN (eps=0.7, min_samples=7)          : 0.5254
```

Observations:

1. The scatter plots show the data points plotted on the first two principal components, with different colors representing the identified clusters.

2. As the value of epsilon (eps) increases, the number of identified clusters decreases. This is because a higher eps value allows more points to be considered as neighbors, leading to fewer and larger clusters.

3. Similarly, as the value of the minimum number of samples (min_samples) increases, the number of identified clusters decreases. This is because a higher min_samples value requires a higher density of points to form a cluster, resulting in fewer but larger clusters.

4. The silhouette scores provide a quantitative measure of the quality of clustering. Higher silhouette scores indicate better clustering quality.

5. The highest silhouette score of 0.5361 is achieved with DBSCAN (eps=0.7, min_samples=5), indicating that this combination of parameters results in the best clustering quality for the IRIS dataset.

6. The silhouette scores generally increase as eps increases from 0.3 to 0.5, but then decrease slightly at eps=0.7. This suggests that an eps value between 0.5 and 0.7 might be optimal for this dataset.

7. The identified clusters seem to have different densities and shapes, with some clusters appearing more compact and well-separated, while others are more dispersed or elongated.

8. The yellow and green clusters appear to be well-separated and compact, while the purple and blue clusters are more dispersed and overlapping.

9. The characteristics of the identified clusters, such as their density, shape, and separation, can provide insights into the underlying structure and patterns in the data, which could be useful for further analysis or decision-making.

Overall, the DBSCAN clustering results demonstrate the impact of parameter choices on the resulting clusters and their quality, as well as the potential for identifying meaningful patterns and structures in the data.