Katarzyna Michalska, 244577IV

1 Clustering and feature selection

The objective of this task was to implement the Hopkins statistic to assess the clustering quality of datasets. A three-dimensional dataset was created, allowing for the calculation of the Hopkins statistic across all two-dimensional combinations of features.

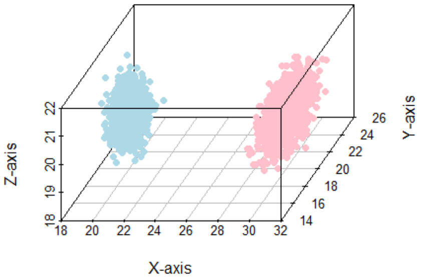

Figure 1: Plot of the dataset.

The plot below illustrate the distribution of data points based on selected feature pairs. The Hopkins statistic for the first pair (x-y) is 0.8555645.

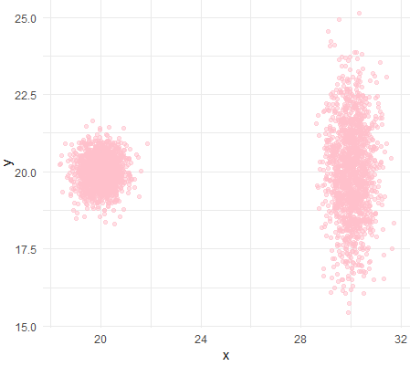

Figure 2: Plots of x-y.

For the x-z feature combination, the Hopkins statistic was 0.9841136, demonstrating significant clustering.

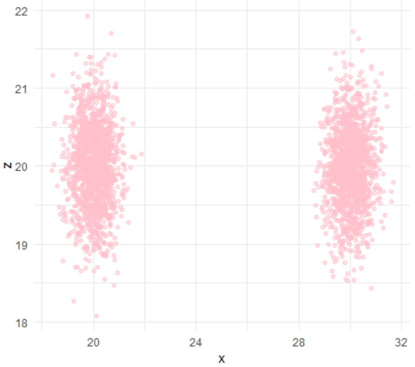

Figure 3: Plots of x-z.

Lastly, for the y-z feature pair, the Hopkins statistic reached 0.968184, indicating strong clustering.

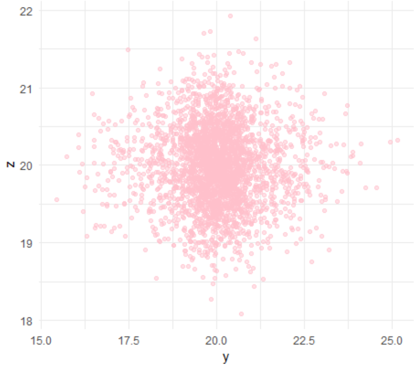

Figure 4: Plots of y-z.

To identify the optimal combination of features based on the Hopkins statistic, a simple feature selection algorithm was implemented. The algorithm determined that the x-z feature pair was the best combination.

2 EM algorithm

To perform the EM algorithm, two datasets were generated: one that is well-separated and another that is more complex and overlapping.

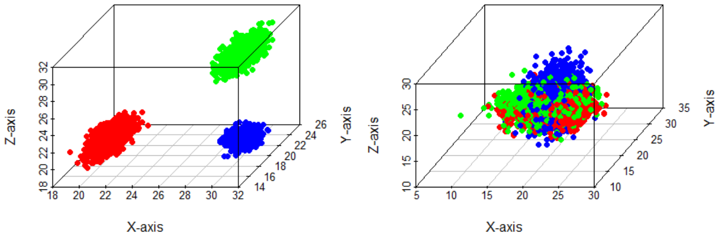

Figure 5: Plots of the datasets.

The K-Means and EM algorithms have been implemented based on the [1], and the data has been clustered. In the first case, we examine a well-separated dataset. The following plots illustrate the results of the clustering process using both K-Means (on the left) and the EM algorithm (on the right), with $k = 3$. As shown, both algorithms effectively cluster the data.

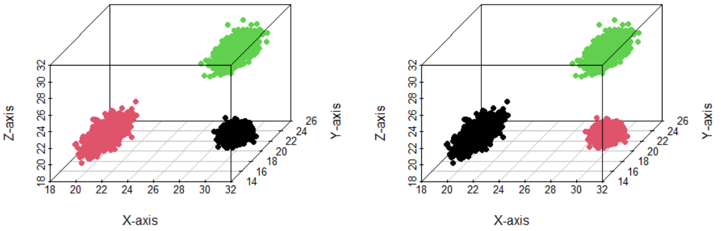

Figure 6: Clustered data with k = 3.

By increasing the value of k to 5, we can observe how the clusters change. In K-Means, clusters are delineated as distinct groups, while in the EM algorithm, we can see individual point assignments to different classes.

Next, we consider a more complex dataset. With $k = 3$, the data is clustered reasonably well; however, we can now observe

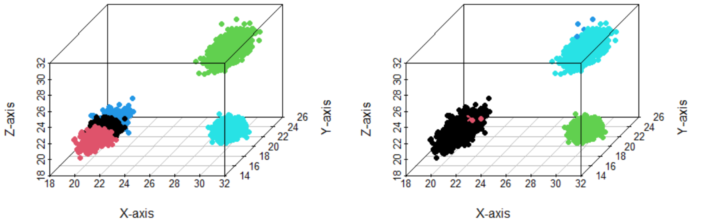

Figure 7: Clustered data with k = 5.

the differences between the two algorithms. The EM algorithm achieves nearly perfect clustering, while K-Means attempts to divide the points into separate groups. This means that individual points assigned to different classes within the same group cannot be distinguished. In contrast, the EM algorithm effectively handles this issue, demonstrating its superiority for our dataset.

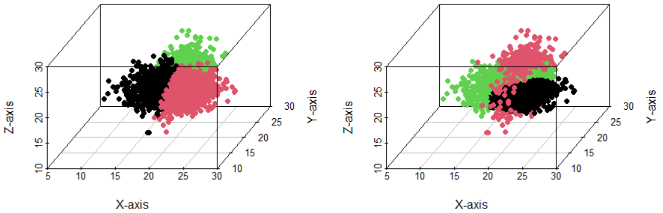

Figure 8: Clustered data with k = 3.

Finally, we examine how our algorithms perform with $k = 5$.

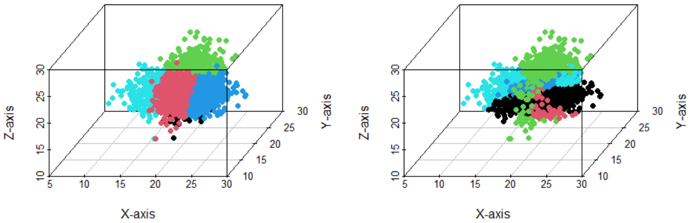

Figure 9: Clustered data with k = 5.

An illustration of convergence is shown in the attached GIFs.

3 Density based methods

To analyze the performance of the DB-SCAN algorithm, a two-dimensional dataset was generated and clustered using DB-SCAN with different combinations of hyperparameters. Initially, the combination of $\epsilon = 0.3$ and a minimum number of points set to 5 was used, as shown in the plot on the left. We observed a significant number of noisy points due to the low value of ϵ. Next, the parameters were adjusted to $\epsilon = 2$ and a minimum number of points set to 3, as shown in the plot on the right. In this case, we can observe that the clusters are now correctly distinguished.

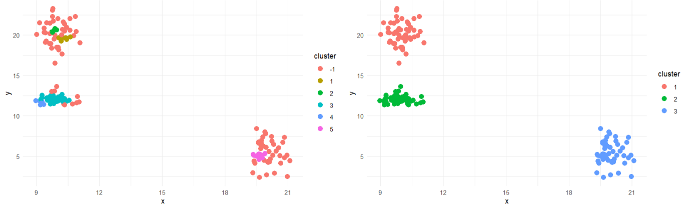

Figure 10: Clusters distinguish by DB-SCAN

4 Transformation projection

The half-moon dataset was generated, and a logistic regression model was trained using the 2D dataset. As seen in the plot below, the decision boundary does not perfectly separate the two classes, reflecting the inherent non-linearity of the dataset. The model achieved an accuracy of 0.89, along with other metrics indicating strong performance, but it is not ideal.

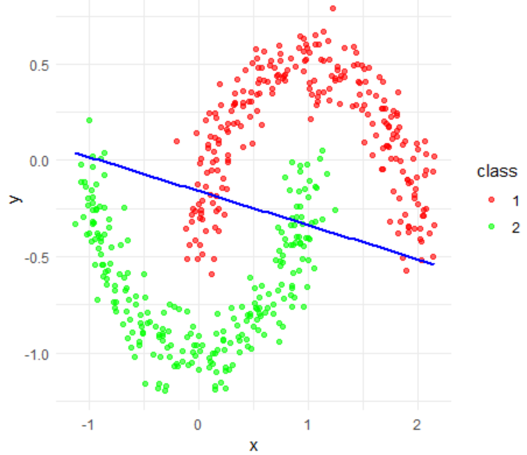

Figure 11: 2D visualization of the half-moon dataset with logistic regression decision boundary

To achieve a perfect logistic regression separator, a transformation into 3D space was applied. A logistic regression model was then trained using the 3D dataset. As seen in the plot, the hyperplane now perfectly separates the two classes. The model's performance metrics indicate an accuracy of 1, along with other optimal results.

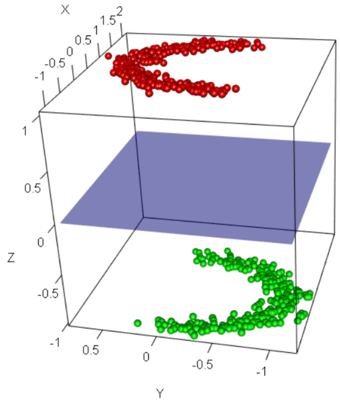

Figure 12: 3D visualization of the half-moon dataset with logistic regression decision boundary

References

[1] Haojun Zhu. EM Algorithm Implementation. https://rpubs.com/H_Zhu/246450, 2016.