Comparing Clustering Methods and Distance Metrics

1. Introduction

In this analysis, I applied different clustering techniques to a binary dataset to identify the optimal clustering method and distance metric. I started with K-Means clustering, explored alternative methods such as Agglomerative Clustering, DBSCAN, and Gaussian Mixture Model (GMM), and used t-SNE for visualization. My goal was to evaluate the performance of each method and rationalize the observed commonalities and differences, especially in the context of binary data and appropriate distance metrics.

2. Initial Approach: K-Means Clustering

K-means clustering was the first method applied to the dataset. K-Means typically assume Euclidean distance as the default distance metric, which can be limiting for binary or sparse datasets. Binary datasets do not align well with the geometric assumptions that K-Means makes, leading to potential inaccuracies. I plotted the elbow curve, which helped determine the optimal number of clusters (optimal_k) based on the flattening of the second derivative of the inertia curve.

3. Exploring Alternative Clustering Methods

Agglomerative Clustering: Agglomerative Clustering supports multiple distance metrics and offers a more flexible approach, particularly for binary data. I experimented with Cosine and Jaccard distances, which are more suitable for binary features.

Gaussian Mixture Model (GMM): GMMs allow for more flexible cluster shapes compared to K-Means, as they assume that data can follow a Gaussian distribution. Although binary data do not necessarily fit a Gaussian distribution, GMM was still helpful in capturing complex cluster shapes.

DBSCAN: DBSCAN is well-suited for datasets with noise or irregular cluster shapes. I experimented with two configurations of DBSCAN using non-Euclidean distance metrics. However, in my dataset, DBSCAN did not perform as effectively:

- Density variations: Since my dataset did not exhibit clear density-based clusters, the resulting clusters were not distinct, and the visualization was less effective compared to Agglomerative Clustering and GMM.

4. t-SNE Visualization

To better visualize the clusters, I used t-SNE. The plot shows how the clusters are distributed in a 2D space, based on K-Means clustering results (with 4 clusters, as identified earlier). Points are colored according to their cluster labels, allowing for an intuitive assessment of how well-separated the clusters are.

The clusters appear well-separated and distinct, suggesting that the high-dimensional clustering captured meaningful structures in the dataset. However, it's important to note that t-SNE itself is not a clustering method, but a dimensionality reduction technique used to visualize high-dimensional data.

5. Conclusion

Overall, Agglomerative Clustering and Gaussian Mixture Models provided the best visual and analytical representations of my binary dataset. K-Means struggled due to its reliance on Euclidean distance, which is inappropriate for binary data. The choice of distance metric was crucial, and metrics like Jaccard proved more effective for binary clustering. The use of t-SNE for visualization helped clarify the clusters and validate the clustering results.