Arushi Gupta, Frazier Baker, Jeremiah Greer
Adv. ML HW4

Assignment 4: Clustering Multidimensional Data

Data was generated by drawing, in equal portions, trapezoidal, triangular, interval, and scalar samples from two normal distributions, centered at -5 and 5, respectively. Both normal distributions had a standard deviation of 1. For an entry of dimension *n*, *n* values were drawn from the distribution and sorted least to greatest to represent the entry in the correct dimension. Hausdorff distances were computed between the data entries and each entry was embedded into a four-dimensional real space. Euclidean distances were computed between the embeddings. K-medoids clustering was performed using the Hausdorff and Embedding approaches and silhouettes and accuracy were evaluated. Accuracy was assessed by assigning labels based on cluster majority. True labels assigned based on the parent distribution from which they were generated (-5 or +5).

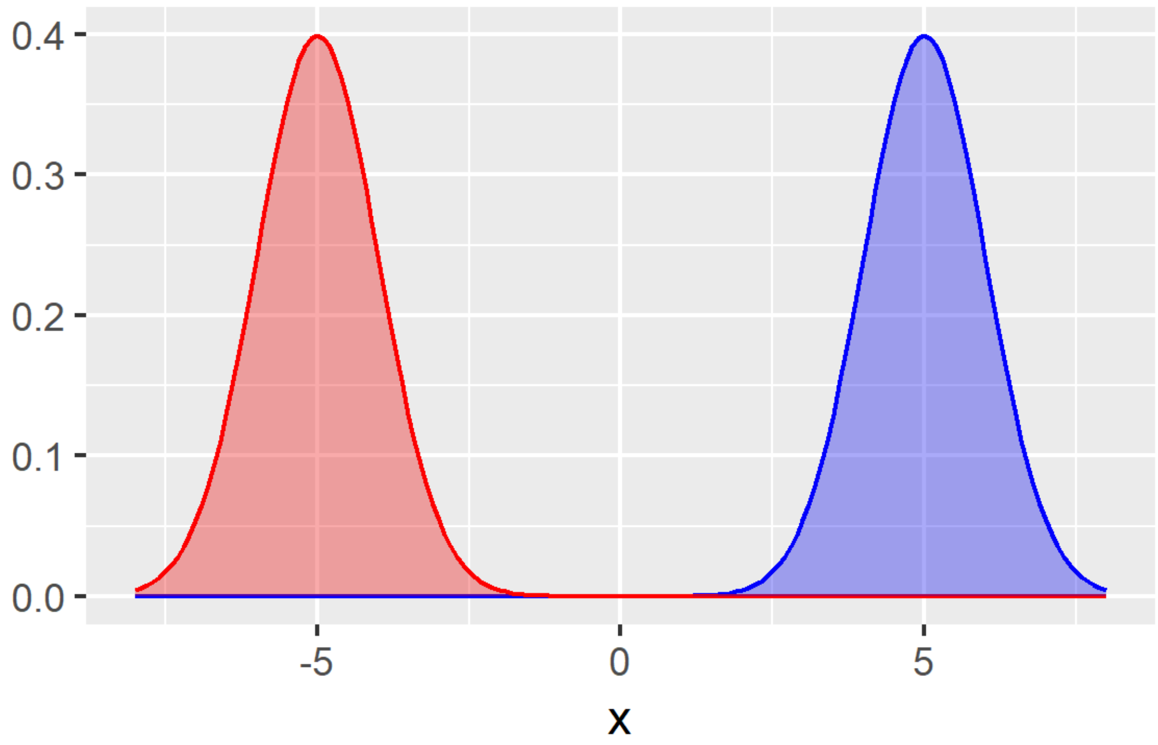


Figure 0. Distributions from which data is generated.

Hausdorff distance was computed and used in k-medoids clustering. It is evident from the silhouettes below that this distance was sufficient for creating clusters that corresponded to the distributions from which the data were generated. It is worth noting that there is more of a curve in the silhouettes for Hausdorff, suggesting that the distances vary more than those of the embeddings, which might be expected with different types of data across a range of values. The distances based on embedding all data into a four dimensional space results in a nearly ideal silhouette plot. Both distance metrics provide a clear-cut clustering with low intra-cluster distance, embeddings just provided a tighter clustering.

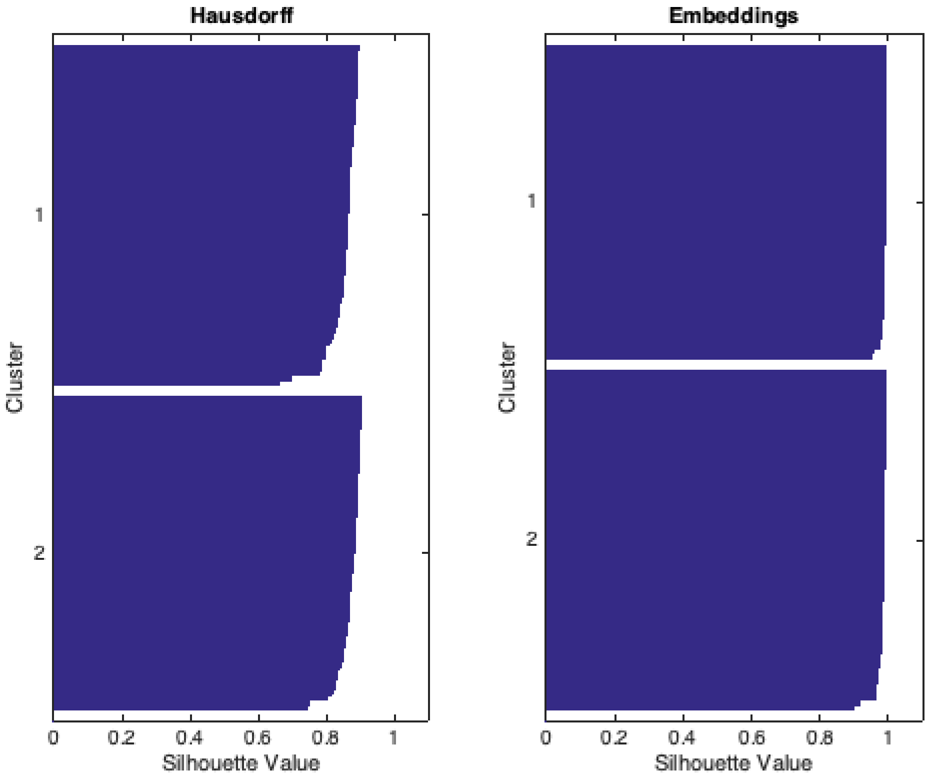
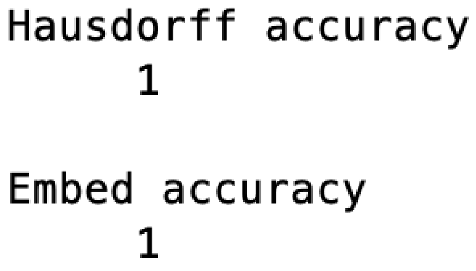


Figure 1. Silhouette Plots for data generated from normal distributions with means of -5 and 5.

The accuracy of this clustering, if the clustering is used as a classification based on the majority of the data points within each cluster, is 1 for both methods of distance. Figure 2, below, confirms this numerically. Of course, in this case the original data is easily separable very low probability of overlap, therefore it makes sense that the clusters should be so easily separable. This serves as a nice proof-of-concept.



```
Hausdorff accuracy
     1


Embed accuracy
     1
```

Figure 2. Accuracies of clusterings using hausdorff distance vs embedding the data using the dataset with means of -5 and 5. Clusters created using k-medoids with k=2.

Real world data often do overlap, however, so this experiment was repeated with means of -1 and 1. With a standard deviation of 1, these normal distributions are likely to generate some overlapping data. Figure 3 shows the new distributions in red and blue with an obvious overlap centered on 0.

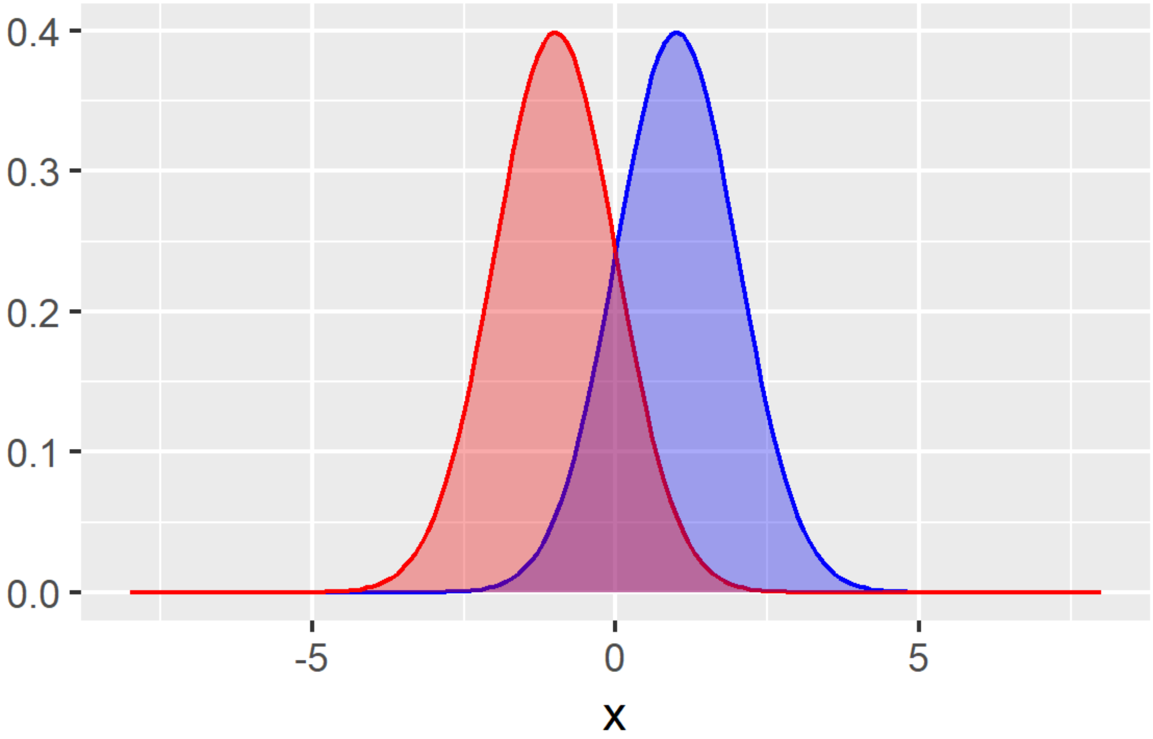


Figure 3. Distributions from which data was drawn for the second experiment.

Samples were taken as before from these new distributions, embeddings with euclidean distance and hausdorff distances were computed, and k-medoids clustering with k=2 was performed. As can be seen in Figure 4, the clustering was still very successful for each of the distance metrics, with only some mixed clustering using Hausdorff distance and no mixed clustering using the embedding approach. This time, the embedding silhouettes are decidedly more curved, indicating higher intra-cluster distance. The silhouette of the Hausdorff distance, which was already slightly curved in the previous example, shows significantly lower values, including negative values. This indicates that some points might have been placed into the wrong cluster based on this metric.

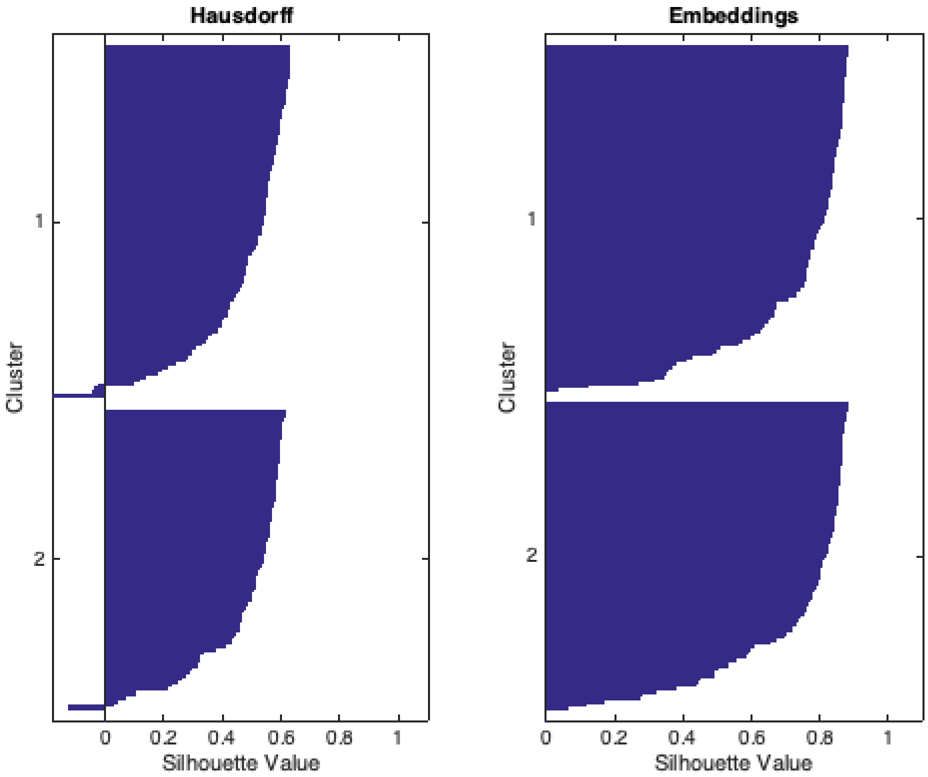
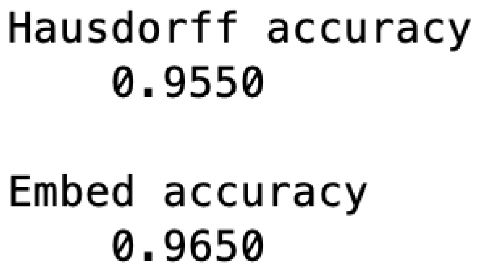


Figure 4. Silhouette Plots for data generated from normal distributions with means of -1 and 1.

The accuracy decreases for both metrics due to this overlap, as can be seen below in Figure 5. This makes sense, as the overlap may cause some data points to be closer to points generated by the other distribution. Overall, though, the embedding approach leads to tighter clusters and higher accuracy.



```
Hausdorff accuracy
    0.9550

Embed accuracy
    0.9650
```

Figure 5. Accuracies of clusterings using hausdorff distance vs embedding the data using the dataset with means of -1 and 1. Clusters created using k-medoids.

In conclusion, the embedding approach had better accuracy and lower intra cluster distance (as seen in the silhouette) when compared to the Hausdorff distance approach. This may have occurred because the Hausdorff distance is based on a maximum whereas the Euclidean distance is an aggregate method which is less susceptible to changes in one value of the set.