Salma Chaaban – 301216551

COMP257 - Assignment 3 – Analysis Report

1. The Olivetti dataset loaded like the previous assignment.
2. The data was split into training, testing, and validation, like the previous assignment.
3. An SVM classifier was trained on the training set like the previous assignment.
4. Using Agglomerative Hierarchical Clustering with the centroid-based clustering rule to reduce the dimensionality of the set, three different similarity measures were used: Euclidean Distance, Minkowski Distance, and Cosine Similarity. For Euclidean, the ward linkage was used, with 40 clusters (before the silhouette score, as 40 is a reasonable number since there are 40 different faces). For Minkowski, the average linkage, and the metric used was “precomputed” as unlike Euclidean or Cosine, Minkowski is not a predefined metric; therefore, it was computed using the pairwise distances between all samples in the dataset, with the Minkowski metric and a distance value of 3 (p=3), which means it is using the Minkowski distance of order 3. Cosine was also used directly like Euclidean, with average linkage.
5. Although all metrics had the same number of clusters (40), their clusters did not look the same. The clusters for each were plotted using TSNE using the first 2 components to see any discrepancies between the metrics. Euclidean had identified some meaningful clusters as they would be identified when they’re seen in the plot, most of them looked reasonable in size, but not all. Minkowski’s clusters did not look as good, as some looked too big. Cosine looked the least reasonable as it had some clusters that were big, and some clusters that looked like they had been broken into more clusters.

The silhouette score approach was used to find the best number of clusters for each distance metric. A range from 5-90 with an increment of 5 was set, with a loop through each distance metric. Different ranges were applied; however, this range was the best, 90 might seem like a high number, but the silhouette score was the best with this range. With higher scores the silhouette score was a little higher, but not that high as to increase the range; it would not make sense to have too many clusters. The best silhouette scores were printed with a plot to show the different numbers of clusters and their silhouette score. The best number of clusters based on silhouette score for Euclidean was 80 and 85 for Minkowski and Cosine; therefore, these were the values used, then the silhouette score was printed for them, which were 0.21 for Euclidean, 0.14 for Minkowski, and 0.13 for Cosine. Again, these scores are not ideal, as they are very close to 0, but they were the best without unreasonably increasing the number of clusters, even with increasing the number of clusters, the silhouette scores would not have a significant increase. Then again, the results were plotted, and Euclidean looked the best.

1. An SVM classifier with k-fold cross-validation (k=5) and with the best parameters found after applying Grid Search CV was used to train a classifier with the Euclidean set (as it had the best results). The average cross-validation accuracy was 0.89, but the validation accuracy was 0.016, which is not ideal; however, it could not be improved; even after trying a different number of clusters, different parameters, or different distance metrics. In conclusion, although hierarchical clustering is a useful method, it does not seem to be the best for the Olivetti faces dataset.