Part 2 Results and Discussion

Evan Casper, Dallas Landry, and Parker Edwards

Dr. Jose Ramignoli

4/8/2022

Clustering attempts to group data points based on some measure of similarity, where data points in the same cluster are more similar than data points in another cluster. The first clustering method investigated is kmeans. Unlike some other methods, the user can specify the number of clusters for data to be sorted into, which are based on a mean value. After that, the data is sorted based on how close it is to the nearest mean cluster value. A complimentary algorithm known as kmean++ can be used to give better initial guesses to the kmeans algorithm, causing it to converge much quicker. The last clustering algorithm used, DBSCAN, groups points into clusters based on their physical distance from each other, where areas of high density are a cluster. What defines “high density” is captured in a user specified value epsilon, where too low of a value results in every point being its own cluster, and too high results in everything being put into one cluster. Points that fall outside of the high density areas are deemed outliers, or noise, and are not sorted into dedicated clusters.

Dimensionality reduction (DR) is a useful tool in data preprocessing as it reduces the number of dimensions, or input variables, in a data set. The less dimensions in a data set, the less difficult it becomes to model via unsupervised learning, however, it is important for the reduced data to still retain enough variables to accurately capture the information present in the full data set. The first DR method we looked at was factor analysis. This can lower dimensionality by looking at what variables are not directly measured and instead inferred from other variables, called latent variables. Other methods used were linear discriminate analysis (LDA), principal component analysis (PCA), random forests, and **t-distributed stochastic neighbor embedding (t-SNE).**

When comparing clustering methods to the ground truth labels, we decided to use a metric known as the adjusted Rand index. This index generally falls between 0 and 1, where 0 signifies that none of the labels are the same when comparing the ground truth and the predicted, and 1 signifies that the clustering labels are identical. Of our clustering methods, kmeans gave us the highest adjusted Rand index (ARI) at a value of .51 when the specified number of clusters was 13. The optimal number of clusters was found by using a trial-and-error method to maximize the ARI. In comparison, DBSCAN gave a maximum ARI of .47 after varying the epsilon value to find a maximum, which occurred at an epsilon of 2.6.

In conclusion, while dimensionality reduction gave issues in our code, the most effective clustering method used was kmeans. This was expected due to kmeans generally working well with larger datasets and non-noisy data, which our data fits.