Group 5
Colton Townsend / cxt180021 (Person 5)
3/22/2023
CS 4375.004

Searching for Similarity Narrative Document

In part 1 & 2, we found a medium-size data set and ran logistic regression, kNN, and decision tree algorithms on it. In the case of kNN classification, we can define a hyperparameter called k that the algorithm uses to identify the k-nearest observations from the training set and assign it a class based on the majority class of the k-nearest neighbors. Similarly; for kNN regression, the kNN algorithm identifies the k-nearest observations from the training set and assigns a value to the observation based on the average value of the k-nearest neighbors. The distance metric used to determine the nearest neighbors for classification and regression was Euclidean distance, though we have identified that other distance metrics were also viable.

Similar to kNN algorithms, we have identified that decision tree algorithms are also suitable for classification and regression. For classification, decision tree algorithms are different because they create a hierarchical model that partitions the data into regions corresponding to different class labels. This tree is grown by binary recursive partitioning, splitting the data into regions until the leaf nodes are too small or too few to be split. For regression, the decision tree is similar to the classification tree but we predict a continuous variable instead of a categorical variable. Each leaf node in the tree corresponds to a predicted value, which is the average value of the nodes that belong to that region. In our experience, kNN performed the best on our dataset because decision trees tend to sacrifice accuracy for increased interpretability.

For part 3, we utilized 3 clustering methods: k-Means clustering, hierarchical clustering, and model-based clustering. k-Means clustering works by grouping data points into k-number of predefined clusters The algorithm iteratively assigns each data point to the nearest cluster center, which is calculated as a mean of all the points in the cluster. Once all points have been assigned, the algorithm recalculates the cluster centers and reassigns data points until the cluster assignments converge.

Hierarchical clustering works by starting each data point in its own cluster and iteratively merges the closest clusters together until all data points have merged into a single cluster. The distance between clusters was calculated by the Euclidean distance metric, though we have identified that other distance metrics were also viable.

Model-based clustering works by assuming the data points were generated from a probabilistic model. The algorithm first estimates the parameters of the model, such as the number of clusters and the distribution between each cluster, then assigns each data point to the most likely cluster based on the model.

In part 4, we performed PLCA and LDA dimensionality reduction, and then tried classification and reduction on the reduced data. PCA achieves dimensionality reduction by maximizing the variance in the data. First, PCA calculates the covariance matrix of the data and then finds the eigenvectors and eigenvalues of the matrix. By projecting a subset of the principal components that explain some of the high variance in the data on the principal components the, we reduce the dimensionality of the data while retaining most of the important information.

LDA achieves dimensionality reduction by maximizing class separability. LDA works by calculating the mean and covariance matrix of each class. Furthermore, LDA calculates the eigenvectors and eigenvalues in order to find a projection of the data that we can apply to a new set of features that maximizes the ratio of the between-class variance to the within-class variance. PCA and LDA are both powerful techniques that are great for simplifying the analysis of a large dataset, especially when some of the features of the dataset may not have been relevant to the analysis.