Assignment 2: Clustering algorithms

Peformance

The algorithms both performed about the same, with Gaussian Mixture clustering performing slightly better than KMeans. Gaussian Mix clustering has an error rate of 7.18%, while KMeans has an error rate of 10.53% for the chosen seed. The reason for this is quite obvious, as Gaussian Mixture is a generalization of KMeans and thus behaves quite similarly. Performance is good for both since the data is separated quite well naturally, and are available in semi-circular clusters (at least after dimensionality reduction).

Attempted changes to the algorithms

Since it was given in the dataset that there are three available clusters, there was no need to attempt multiple different cluster counts, and we already knew what the optimal cluster count is. It also let us verify the performance of the algorithm as if it was a supervised algorithm (by matching the clusters to the correct label). However, we also evaluated the algorithm as if we didn't have the ground truth, using the silhouette score, though we didn't rely on this value when tweaking the algorithm.

We ended using nearly the default settings for both algorithms, as that gave the best results. For the Gaussian Mixture algorithm: Increasing the convergence threshold above the default decreased performance, and decreasing it didn't improve it. However, we slightly increased the regularization, which gave gave better results results. Increasing it more caused the algorithm to perform worse again. Changing the covariance type caused the program to either crash or give very bad results. For the KMeans algorithm we tried changing the convergence threshold as well, which didn't make any difference in performance.

As can be seen when enabling extra plots, most of the features have very little correlation (except feature 1 and 2). We decided against removing any of the features before training because of this, and because our results were quite good.

Legend

As noted on the plots circles display the dimensionality-reduced ground truth data labels for each point. The triangles show the label assigned by the clustering algorithm. The large ellipses show the cluster centers (center of ellipse is the cluster center).

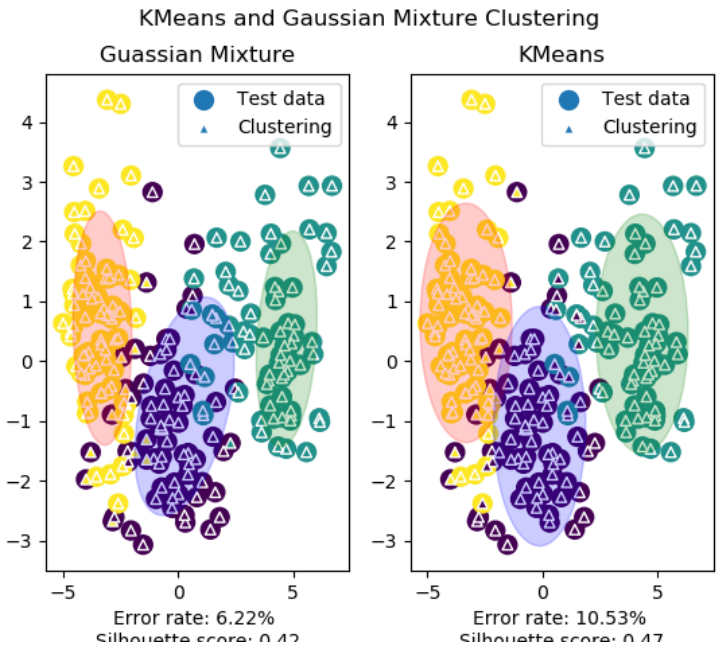

Figure 1: Visualization of algorithms (two-axis PCA applied to data). Silhouette scores slightly cut off