ME8813

Homework 4

Patrick Gardocki

1. Scree-Plot
   1. High dimensional data is problematic for various reasons. This type of data can lend to overfitting as the model becomes very complex and has noise that makes it difficult for any patterns or trends to be found during analysis. Also, highly dimensional data is difficult to visualize which also impacts the ability to understand the data and its trends. The complexity of the data is usually addressed by attempts to reduce the dimensionality of the data.
   2. The Scree Plot shows that the first for principle components contribute most to the variance of the dataset.
   3. Chart, line chart

      Description automatically generated
2. K-Means
   1. K-Means was used to cluster the dataset. Random restarts and SSE were used. The optimal number of clusters found was 3 because that is where the elbow plot ‘bends’ and starts a linear trend as more clusters are added. K-Means has a Silhouette score of 0.539. This is neither good or bad, as these scores range from -1 to 1. 1 is the best score and -1 is the worst score.
   2. Chart, line chart

      Description automatically generated
3. GMM
   1. Both K-Means and GMM are clustering methods but K-Means uses a deterministic approach while GMM uses a probabilistic approach. K-Means assumes that the clusters are spherical and have equal variance, which can limit the complexity of data it can analyze effectively. GMM can create clusters of any shape or size since each cluster is a mixture of Gaussian distributions. This allows GMM to better handle noisy, or missing data but at the expense of computational speed. Also, GMM benefits from knowing the number of clusters beforehand. K-Means, on the other hand, is comparatively faster and more efficient. Regarding the given dataset, GMM has a Silhouette score of -0.009. This is neither good or bad, as these scores range from -1 to 1. 1 is the best score and -1 is the worst score. K-Means had a score of 0.539, showing that is performed better with the given dataset.
4. Autoencoders
   1. The implemented autoencoder reduces the number of dimensions from 15 down to 4. The figure below shows the prediction loss as the mean absolute error between the actual and reconstructed dataset. It follows a normal distribution centered around about 0.18.
   2. Chart, histogram

      Description automatically generated
   3. PCA is a linear transformation method to find a lower dimension representation of the data. It finds directions with the most variance, which become Principal Components. The results can be used for visualization or more analysis. Autoencoders are neural networks that have an encoder and decoder. The input data is mapped onto a latent space and then reconstructed back to the original space. The autoencoder aims to minimize the difference between the original and reconstructed data. While PCA is linear, autoencoders can capture nonlinear behavior more accurately. Autoencoders create a lower dimensioned representation of the data while PCA makes a set of principle components that can be used to reduce the dimensionality of the original data.