1 Introduction

This project investigates algorithms in unsupervised learning: PCA (Principal Component Analysis), K-means clustering and GMMs (Gaussian Mixture Models).

1.1 Experiments

First, two datasets are chosen. In this experiment, the datasets are the Mushrooms dataset from Kaggle and the Pima Indians Diabetes dataset from Kaggle.

Then, two clustering algorithms (K-means clustering and GMMs via expectation maximization) are run.

Then, the datasets are reduced into fewer dimensions using Principal Component Analysis. The clustering algorithms are re-performed on the reduced dataset.

Then, the original results of the PCA algorithm on the Pima Indians Diabetes dataset are taken, and the neural network is trained on this reduced dataset. To evaluate the accuracy of the neural network, both the training and testing datasets are projected with the dimensionality function that had been generated before.

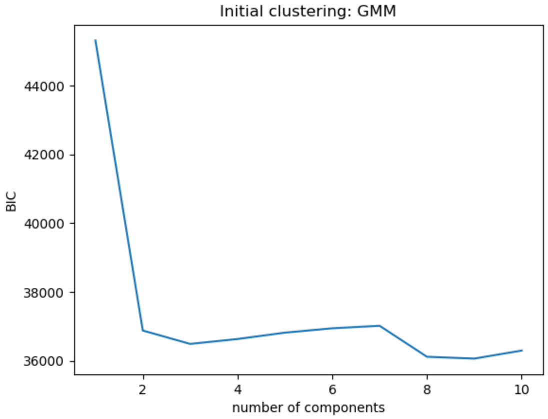
Finally, the training dataset of the Pima Indians Diabetes dataset is clustered using the k-Means and GMM algorithms and reduced to fewer dimensions using the two algorithms. In these specific experiements, the k-Means algorithm reduces the data to the cluster-distance space (sklearn's `transform` function), while the GMM reduces each datapoint to the probabilities of the datapoint being in each cluster (sklearn's `predict_proba` function). Then, with the clustering algorithms used on the training set, the test set is also reduced to those same features. A neural network is then trained on the training datasets produced by the two clustering algorithms.

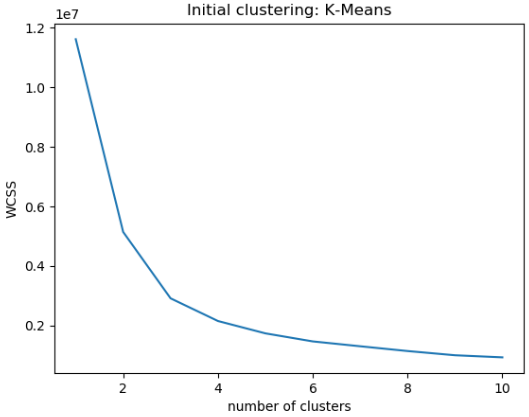
2 Results

2.1 Initial clustering

For each of the two datasets, the GMM and K-means clustering were both run on the training set. To determine the optimal number of cluster, some form of a metric was plotted over the number of clusters. For K-means, this metric was WCSS (within-cluster sums of squares), while metric for GMMs was BIC(Bayesian information criterion). Here are the plots generated:

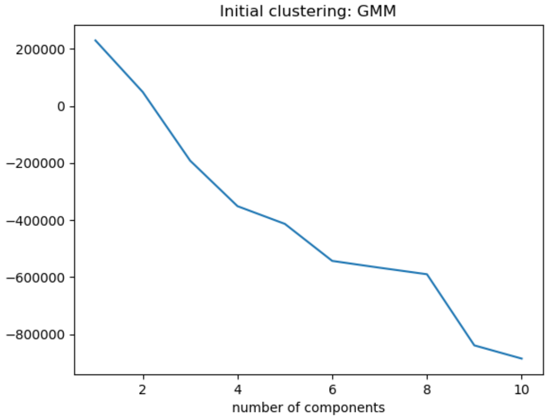
2.1.1 Initial clustering: Pima Indians Diabetes dataset

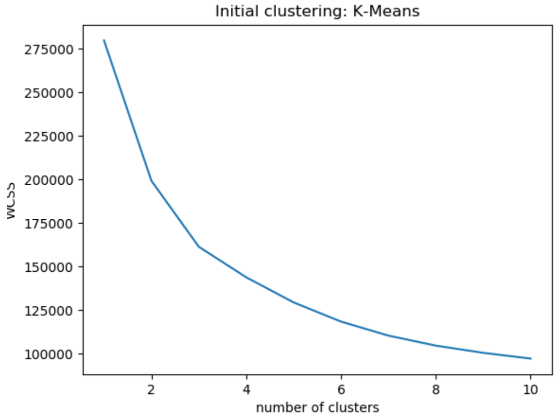




In this dataset, a clear ”elbow” can be seen at two clusters for the GMM, and there is a slight ”elbow” at three clusters for the K-means. This problem is a binary classification problem (predicts whether someone has diabetes or not), so it is reasonable for the number of clusters to be close to two.

2.1.2 Initial clustering: Mushrooms dataset

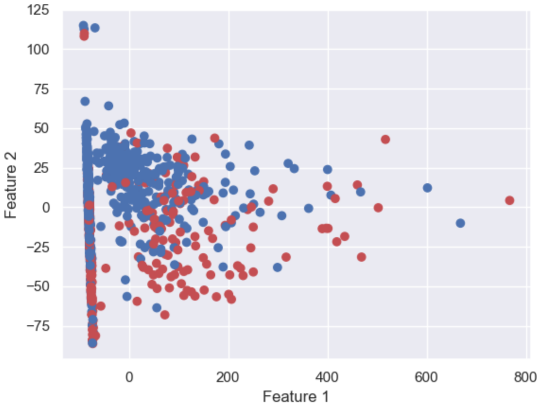




2.2 PCA

PCA was run on a couple of the datasets. To make the datasets easy to visualize, the PCA was run such that the data was to be split into two components. Here are the results of the PCA that was performed:

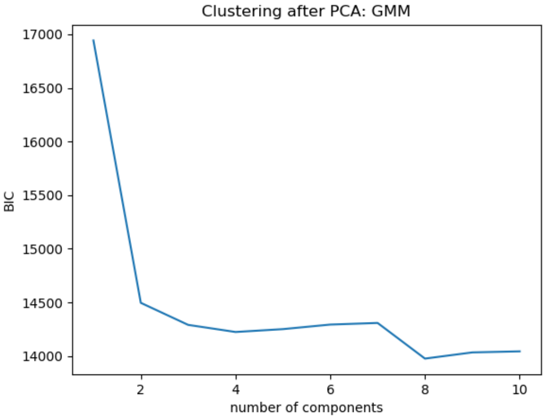
2.2.1 PCA: Pima Indians Diabetes dataset

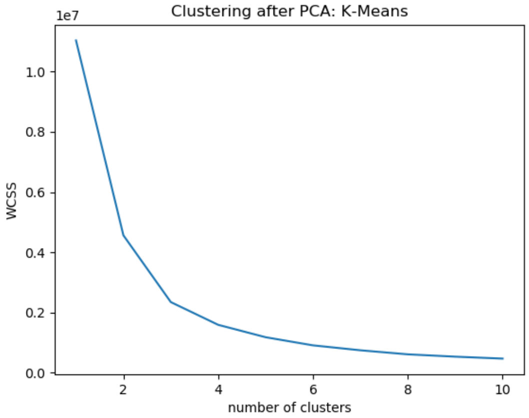


2.3 Clustering after PCA

The two datasets were clustered with the two algorithms after the PCA dimensionality reduction algorithms were run on them.

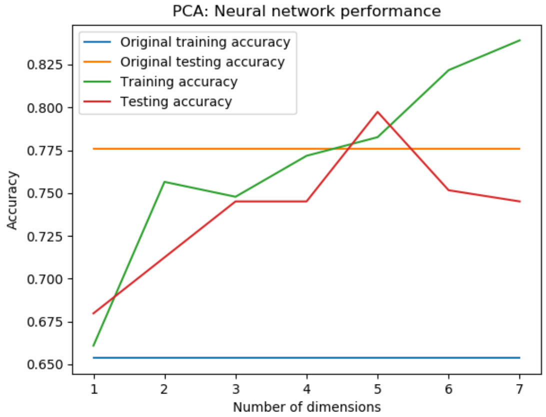
2.3.1 Clustering after PCA: Pima Indians Diabetes dataset





2.4 Training a neural network after PCA

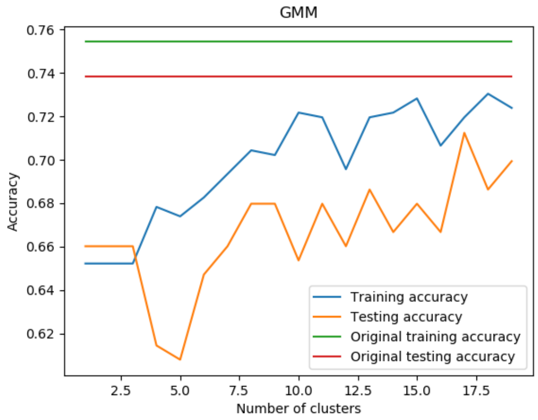
The Pima Indians dataset was split such that 60 percent of it was training data and 20 percent was testing data. Here is a plot of the accuracy of the neural network (training and testing) with respect to the number of principal components generated in the PCA algorithm:

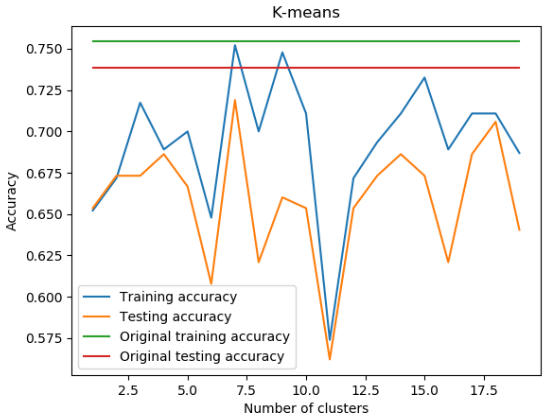


2.5 Clustering for dimensionality reduction

The Pima Indians diabetes was split into a training and testing set (60 percent training and 20 percent testing).

Then, the GMM and the k-Means algorithm were each used to reduce the dimensions of the data for the Pima Indian Diabetes dataset. For each number of clusters, a neural network was trained to produce a mapping from the reduced data to labels. Here are the results for the training and testing accuracies:





3 Sources

[1] https://www.kaggle.com/uciml/pima-indians-diabetes-database