Project 3: Unsupervised Learning

Andrew Scott  
ascott97@gatech.edu

***Abstract—***This report examines several different methods of clustering and dimensionality reduction on classification data sets.

# Introduction

This report details the analysis done on two different clustering methods and four different dimensionality reduction methods. The clustering methods include K-Means and Expectation Maximization and the dimensionality reduction methods included Principal Component Analysis, Independent Component Analysis, Random Projection and blerggg. Each of these methods are analyzed in detail in the sections below for two separate classification datasets. Additionally, all of the methods are used to transform the datasets and then analyzed their performance training a neural network to classify the data for one of the datasets.

# Data Sets

## Wine Dataset

The Wine data set has 1599 data points which corresponds to different wines and the features for each data points includes 11 different measurements that describe the physical and chemical properties of the wine. The labels for this data set is a quality score given by a group of wine tasters on a scale of 1-7, with 7 being the highest quality and 1 being the lowest.

## Breast Cancer Dataset

The breast cancer data set has 682 data points which corresponds to different patient data. There are 9 different features for this dataset that are attributes for lumps/tumors observed for each patient. The labels for this dataset are either a positive or negative for a cancer diagnosis.

# Clustering

## K-Means

For K-Means the kMeans API from Sklearn was used to cluster the data from both the Wine and Cancer Data sets. The silhouette score and elbow method were used to select the optimal k value for K-Means while all the silhouette, completeness, and homogeneity scores on the clustered data. The plots below show the silhouette score and the within cluster sum of squared distance (WCSS) for both the wine and cancer datasets. These are plotted against the number of clusters (k values).

Chart, line chart

Description automatically generated

**Figure 1:** K-Means: K selection for Wine and Cancer Datasets

The process used to select the optimal k value was to look at both the silhouette score of the data in conjunction with using the elbow technique. The elbow technique involves identifying the point in the WCSS plot where the plot has the sharpest change in slope. This appears to occur at k = 7 for the wine data and k = 4 for the cancer data. The process for the silhouette score is simply to look at the values for k that have the highest score. While the silhouette score is not the highest for either k values for each of the data sets, the score is still relatively close to the highest score and it is prior to the score “dropping off” in both of the plots. The plots above seem to suggest that it might be useful to use both metrics to select the correct k value since just looking at either silhouette score or the elbow method in isolation might not yield the best information. In the case of the above datasets the elbow method seemed to be sufficient for selecting the k-value; however, the silhouette score still provides the opportunity for added confidence in the selection as a crosscheck.

## Expectation Maximization

For expectation maximization the Gaussian Mixtures API from Sklearn was used to cluster the data from both the Wine and Cancer Data sets. The metrics used to analyze the clustering performance were Alkaline Information Criteria (AIC) and Bayesian Information Criteria (BIC). The plots below show the results for the both the wine and cancer datasets.

Chart, line chart

Description automatically generated

**Figure 2:** GMM K Selection for Wine and Cancer Data

An interesting thing to note when comparing the plots between the wine and the cancer data sets is that the AIC/BIC curves align very closely for the cancer data set while the curves do match up as closely for the wine dataset. Since the BIC curve also takes into account the number of elements in the dataset this indicates that the maximum likelihood value calculated for the wine dataset is not enough to compensate for the increased model complexity for higher number of clusters. The method use to select the number of clusters to used involved looking at the lowest scores for both AIC and BIC. In the wine dataset the value of 7 was chosen since the BIC seems to be at a minimum at this point. The BIC and AIC scores were at a minimum at 9 clusters for the cancer dataset but there also appears to be another local minimum value at 7 so this point was chosen so as to avoid cluster overfitting.

The table below summarizes the homogeneity, completeness and v-measure scores for the wine and cancer datasets for both K-means and GMM for the selected number of clusters.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cluster Type | Dataset | Clusters | Homogeneity | Completeness | V-Measure |
| K-Means | Wine | 7 | 0.1238 | 0.0869 | 0.1021 |
| K-Means | Cancer | 6 | 0.8181 | 0.3420 | 0.4824 |
| GMM | Wine | 7 | 0.113 | 0.0750 | 0.0902 |
| GMM | Cancer | 7 | 0.7161 | 0.3367 | 0.4581 |

All three of the metrics seems to be fairly close when using K-means and GMM for each dataset; however, all of the score are higher for the K-Means clustering than they are for the GMM clustering. Explain why K-means does better here.

# Dimensionality Reduction

## Principal Components Analysis (PCA)

To determine how many principal components to use to reduce the dimensionality of the data the explained variance ratio was analyzed while running the PCA algorithms. The explained variance ratio is calculated by looking at the amount of variance that each component contributes. The value for each component corresponds to the eigenvalue for each component and essentially ranks how important that specific component is. The plot below shows a plot of the explained variance for the wine data starting with only one component all the way up until the number of components equals the number of features from the original data set. Here, the explained variance is a cumulative sum that is added each time a new component is included. The component values along the a-axis are ordered such that the first component is the most significant while the final component is least significant, which is why the curve in both plots starts out steeper and then shallows out.

Chart, line chart

Description automatically generated

**Figure 3:** PCA: Wine and Cancer Data - Explained Variance Ratio

There are several different heuristics that can be used to select the target number of components when using the PCA algorithm. In this case, the number of components was selected by looking at the point the explained variance is greater than 90%. This threshold value was chosen to preserve as much information in the data as possible while also attempting to avoid overfitting. In this case, that occurs once there are 7 components for the wine data and 6 components for the cancer data.

One interesting to note between the two datasets is that the wine data seems to have a more consistent decrease in contribution for each component which gives it a more parabolic shape. The first component of the cancer data set account for 65% of the variance, while most of other components equally contribute around 5-10% of the variance which is why the curve appears to be more linear.

After running the PCA algorithm on both datasets with the reduced number of seven components, the new datasets were both run through the K-Means and GMM clustering algorithms in order to assess and inspect the effect that the dimensionality reduction has on the clustering performance. The plot below shows the K-Means performance for both the wine and cancer datasets.

Chart, line chart

Description automatically generated

**Figure 4:** PCA - Wine and Cancer K-Means Metrics

For the wine dataset there seems to be an improvement in the clustering metrics particularly for the lower clusters (less than 6 clusters). This seems logical given that the original dataset was reduced from 11 original features to 7 components.

The plots below show the GMM metrics for both the wine and cancer data after running PCA with six components.

Chart, line chart

Description automatically generated

**Figure 5:** PCA - Wine and Cancer GMM Metrics

The GMM metrics show a noticeable and constant improvement across all the metrics for both the wine datasets with the reduce components.

## Independent Components Analysis (ICA)

The method used to choose the optimal number of components for ICA involved looking at the kurtosis of the dataset after applying ICA dimensionality reduction algorithm for various values of the number of components. Kurtosis is a measure of how closely a set of variables matches a Gaussian distribution. In the plots below the kurtosis value shown is the average kurtosis of all the datapoints for each component. Additionally, the kurtosis value is normalized to zero so that higher numbers correspond top higher average kurtosis among all the components.

Chart, line chart

Description automatically generated

**Figure 6:** ICA Wine and Cancer - Kurtosis

The plot for the wine data indicates a clear point, when number of components = 8, where the average kurtosis is highest. This implies that reducing the number of components down from 11 (original number of features) to 8 helps increase the non-gaussian-ity of the data and thus this seems to be a fairly useful heuristic for determining the appropriate dimensionality reduction of the wine data. The Cancer dataset is a bit tricker since the kurtosis seems to be highest when the number of components is 2. Although, the kurtosis does not seem to vary nearly as much with the cancer data as the wine dataset.

Kmeans overview

Chart, line chart

Description automatically generated

**Figure 7:** ICA - Wine and Cancer K-Means Metrics

Discuss kmeans

GMM overview

Chart, line chart

Description automatically generated

**Figure 8:** ICA - Wine and Cancer GMM Metrics

Discuss GMM

## Random Projections (RP)

For random projection the reconstruction error was initially used as a method for choosing the optimal number of components. However, this turned out to not be the best heuristic for choosing the number of components since the reconstruction error essentially constantly decreased as the number of components were increased. As an alternative the pairwise Euclidean distance was calculated for both the original and reconstructed data and for each component value the squared differences between the Euclidean distance matrices were averaged. This produced a more useful curve that gave more insight as to when the reconstruction of the original data seems to look more promising. The plots below show these curves for both the wine and cancer data.

Chart, line chart

Description automatically generated

**Figure 9:** RP Wine and Cancer - Euclidean Distance Comparison

Using a heuristic like the elbow technique used for k-means clustering the number of selected components was selected as 9 and 6 for the wine and cancer data respectively.

Chart, line chart

Description automatically generated

Figure : RP- Wine and Cancer K-Means Metrics

Analysis…

Chart, line chart

Description automatically generated

**Figure 11:** RP - Wine and Cancer GMM Metrics

## Variance Threshold (VT)

Variance Threshold is a relatively simple feature selection method that eliminates features that have a variance less than some threshold variance. For both the wine and cancer datasets a variance threshold value or “1” was used, meaning that any feature that has a variance less than 1 across all datapoints is eliminated. The value of 1 was selected since this seemed to be the only variance threshold that was successful in reducing the number of features without removing all of the features in both datasets. The VT method reduced the number of features in the wine dataset from 11 to 4 while the cancer dataset was only reduced from 9 to 6. It is interesting in this case that the wine dataset had more features eliminated than the cancer data set and is probably due to the number of similar values in each feature set in the wine dataset vs the cancer dataset.

# Nueral Network Analysis

# Conclusion

# Citations

<https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients>

<https://archive.ics.uci.edu/ml/datasets/Wine+Quality>