Final Presentation

FNU Rhythm

STAT 639, Texas A&M University, College Station, TX

April 21, 2020

Section 1

1. **Supervised Learning**
   - Approach
   - Results
   - Summary
2. **Unsupervised Learning**
   - Approach
   - Results

Data

1. More predictors than observations, so some variables may be **collinear**.
2. More priority should be given to variables with more influence on the class.
3. Variables were selected using **Lasso**.
   - FD -Full data set that contains all 500 features
   - RD -Reduced data set that contains 21 features obtained using the lambda for the minimum Misclassification Rate (MSR) from lasso regularization
4. The performance of both datasets was compared by fitting different classifiers.
   - Logistic Regression
   - Boosting
   - Neural Networks
   - Decision Trees
   - Random Forest
   - Support Vector Machines (SVM)
   - Linear Discriminant Analysis (LDA)

Random Forest

Random Forest was tuned by varying the number of variables *(mtry)* sampled at each split point and comparing the estimated 10-fold Cross Validation errors.

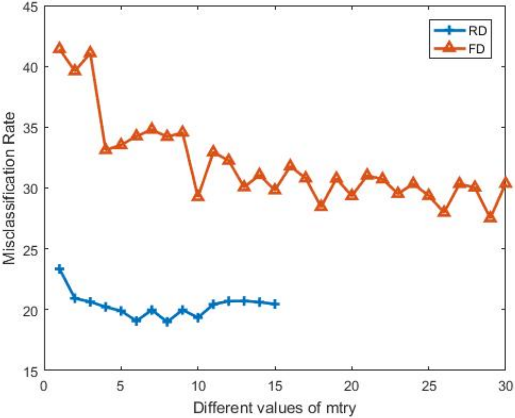

Figure: Comparison of best performance for the entire dataset vs the reduced dataset

Summary

The best parameters were used and 10-fold Cross Validation was used on the same data. The estimated testing MSR recored is given below:

Classifier	RD	FD
Linear Discriminant Analysis	23.64%	27.52%
Support Vector Machine	25%	36.5%
Decision Trees	27.06%	25.97%
Random Forest	**19%**	27.52%
Boosting	21.22%	25.69%

Table: MSRs of various classifiers after tuning parameters

Hence the predicted best classifier is a Random Forest model of the observations but taking into account only 21 of the 500 variables as suggested by doing Lasso regularization.

Section 2

1. Supervised Learning
   - Approach
   - Results
   - Summary

2. **Unsupervised Learning**
   - Approach
   - Results

Approach

1. Using Principal Component Analysis (PCA) create two new data sets
    - $y90$ - principal components that account for 90% of the variance in y
    - $y95$ - principal components that account for 95% of the variance in y
2. Using the following methods, determine the optimal number of clusters, k, for a given clustering method
    - AIC
    - BIC
    - NbClust[?]
    - WSS
    - Silhouette
    - Gap Statistic
3. Find the best k for the following clustering methods
    - K-means
    - Hierarchical
    - Gaussian Mixture Model
    - Density-Based
4. Determine optimal combination of clustering method and k

Results

1. K-means appears to be the best method for classification of this data set
2. The optimal number of clusters was determined to be $k = 2$
3. The results of two methods for determining the optimal number of clusters will be presented
    - **Silhouette:** the optimal number of clusters maximizes the average silhouette width
    - **NbClust:** package in R which contains 30 indices used to determine the optimal number of clusters and the optimal partition of the data. 26 of the 30 indices were tested[?]

Results: Average Silhouette Width

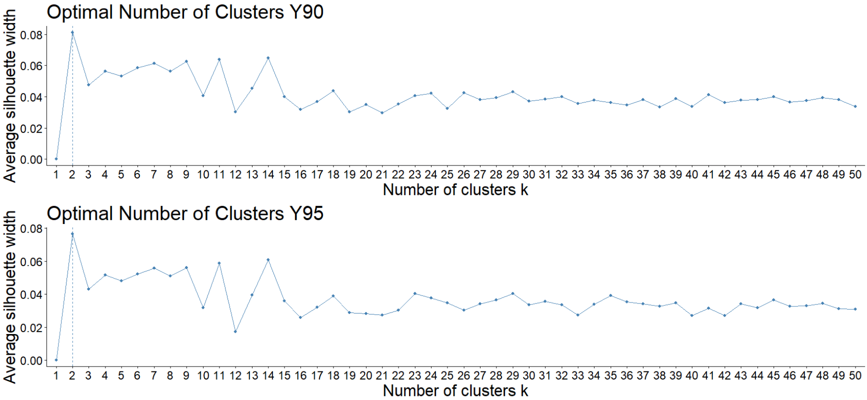

Figure: Using the silhouette method, the optimal number of clusters is $k = 2$ for both data sets

Results: NbClust

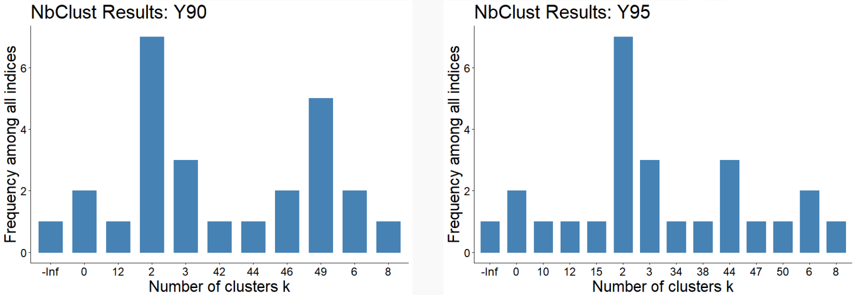

Figure: 7 of the 26 indices used in NbClust proposed $k = 2$ as the optimal number of clusters for both data sets

Results: Final Clusters

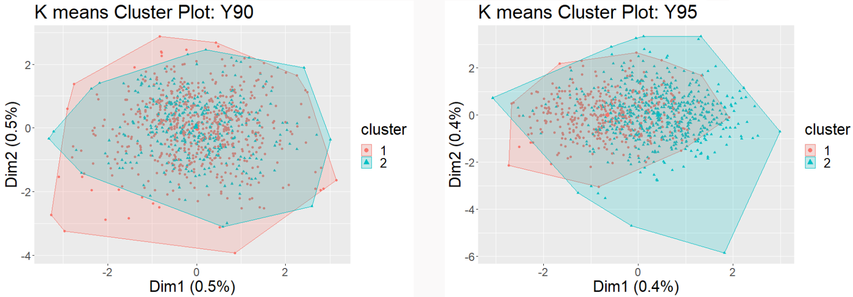

Figure: The optimal partition of the data into 2 proposed by NbClust. Both data sets have clusters with 395 and 605 members.

Conclusion: Unsupervised Learning

1. The determination of the best method and number of clusters to use in unsupervised learning remains a challenge
2. Principal Component Analysis (PCA) was used to reduce the data set into two smaller data sets, $y90$ and $y95$, which accounted for 90% and 95% of the variance in the data, respectively
3. A number of methods for clustering the data and determining the optimal number of clusters were explored
4. Results of this analysis indicated that the optimal number of clusters for this data set is $k = 2$

References

[1] alika Charrad, Nadia Ghazzali, Veronique Boiteau, Azam Niknafs (2014) NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software* 61(6), 1–36.

[2] Trevor Hastie, Robert Tibshirani, Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer.

[3] Gareth James, Daniela Watson, Trevor Hastie, Robert Tibshirani (2013). *An Introduction to Statistical Learning with Applications in R*. New York: Springer.

Thank You!