STAT 639 Project

Natalie Coleman (635003935), Rohan Singh Wilkho (227009504), Chiawei Hsu (630008717)

April 29, 2021

1 Supervised Learning (Classification Problem)

Objective

The objective is to train and analyze different classification models using the provided training covariates (x), which consisted of 500 features and 400 observations, and the response (y), which was divided into $[0, 1]$.

Methodology

The methodology generated models with the goal of minimizing the estimated testing predicted error as a proxy for the testing error of new responses:

1. *Scale the covariates (x)*: If features are not in the same scale, this may lead to inaccurate predictions. Although information about the units of the 500 features was not available, it is still necessary to scale them first.
2. *Train-test split*: The data was split into an 80% train subset, which was used to experiment and train different models, and 20% test subset, which was used to predict the testing error.
3. *Feature selection*: The prediction capability of any machine learning model depends upon the quality of features, or the best feature subset to represent the data. To accomplish this, four different feature selection techniques and two pathways were considered. First, highly correlated features were removed using a correlation threshold ($> 0.90, 0.75$). Second, statistically insignificant features were removed using ANOVA p-value threshold ($< 0.05, 0.01$). Third, the features were reduced either by using Step-Wise Regression or Recursive Feature Elimination. Each classifier had 8 distinct feature selection combinations.
4. *Hyperparameter Tuning*: After the best performing feature subset was selected, hyper-

parameter tuning was performed for different classifiers. Control Grids for the values of hyperparameters were created to be checked and be repeated with cross-validation. Again, all hyperparameter tuning only considered the training subset.

To begin, Logistic Regression considered the interaction between remaining variables and Lasso elimination. K-Nearest Neighbors depended on parameter K. Both Linear Discriminant Analysis and Quadratic Discriminant had no specific hyperparameters. Naïve Bayes used laplace smoothing. Support Vector Machine used C (cost of constraints violation) and sigma (kernel parameter). Finally, Random Forest used mtry (random samples) and ntress (trees grown).

5. *Model training*: Then, different classifiers were trained over the entire train set. Here, the classifiers were initialized with the tuned hyperparameter values previously detected.
6. *Estimating testing error*: The final step was to predict the responses corresponding to the features in the test set and calculate the misclassification error rate, which is the estimated testing misclassification reported. The model with the least misclassification error rate was finally selected.

Results

Each classifier had 8 features combinations which gave 8 different estimated testing errors. Out of these 8, the combination which gave the least estimated misclassification error rate was chosen for each model and the bar graph demonstrates the best results (Fig. 1).

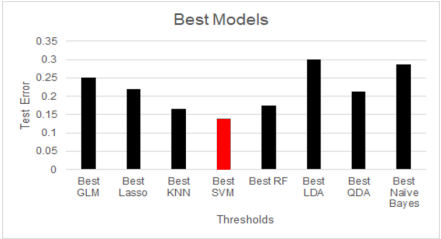


Figure 1: Results from Different Classification Methods

The best performance was given by the Support Vector Machine (i.e. estimated test error of **0.1375**). This method focuses on maximizing the marginal distance between data

points for better classification. Feature selection was done with a correlation threshold of 0.75, ANOVA p-value of 0.05, and RFE selection for 7 remaining features. The tuned hyperparameters were calculated as $C = 12.9$ and sigma= 0.0774.

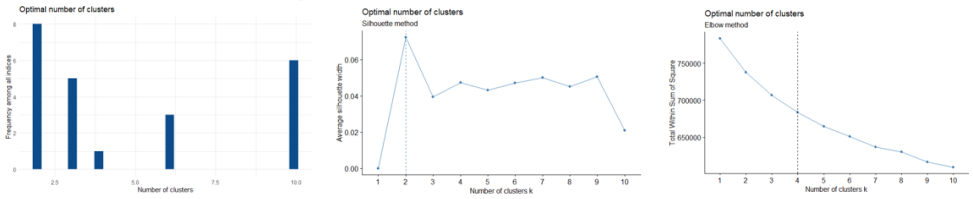
2 Unsupervised Learning (Clustering Problem)

Objective

To cluster observations into an unknown number of 'K' clusters with a dataset of 1000 observations with 784 variables. The methodology explored and analyzed different clustering techniques (1) K-Means, (2) K-Medoids, (3) DBSCAN, (4) Hierarchical Clustering.

Experiments and Results

1. *K-Means*: As shown in Fig. 2, the optimal K chosen was based on NbClust (**K=2**), Silhouette (**K=2**), and Elbow (**K=4**). NbClust uses 30 indices and determines the one with the best frequency; Silhouette bases it on the point similarity to the cluster; while Elbow determines the bend in the graphic to explain graph variation.

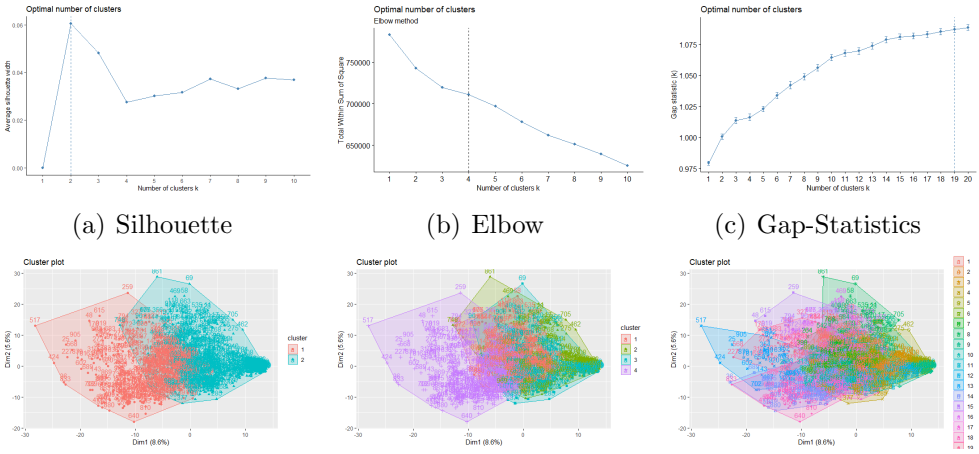


(a) NbClust (b) Silhouette (c) Elbow

(d) Method(a)/(b): K=2 (e) Elbow: K=4

Figure 2: K-Means Results

2. *K-Medoids*: As shown in Fig. 3, optimal K was chosen using Silhouette (**K=2**) and Elbow (**K=4**) in addition to Gap Statistic (**K=19**).



(a) Silhouette (b) Elbow (c) Gap-Statistics

(d) Silhouette: K=2 (e) Elbow: K=4 (f) Gap-Statistics: K=19

Figure 3: K-Medoids Results

3. *DBSCAN*: This technique uses KNNdistplot to find the optimal 'eps' for different MinPts based on the distance in the elbow. Fig 4 shows that the optimal 'eps' values remained to be 40 despite different 'MinPts' (1,7); the number of clusters were also found to be same, i.e. **K=1**.

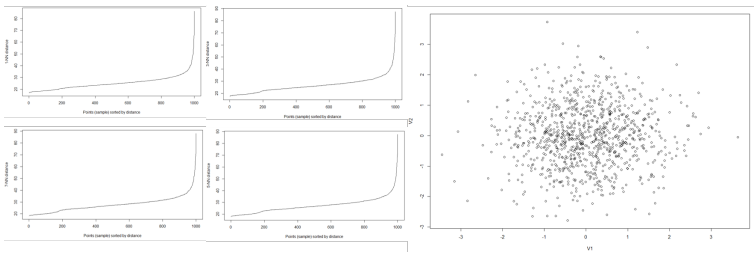


Figure 4: Results from kNNdistplot and Clustering with DBSCAN

4. *Hierarchical clustering*: The final technique attempted was using Divisive (DIANA) and Agglomerative (AGNES) hierarchical clustering for a final **K=1**.

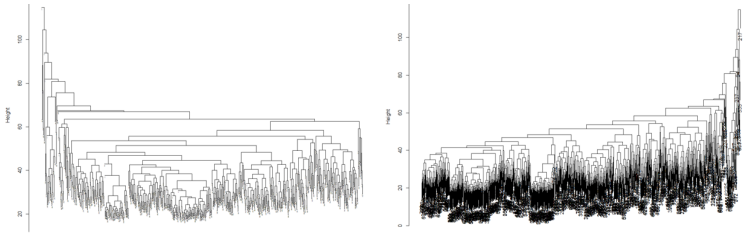


Figure 5: Hierarchical clustering results; Left:DIANA, Right:AGNES