Oscar Olaya

## Final Project Report

***Methods*:** To complete the supervised learning task, we utilized SVM with linear, polynomial kernel, radial basis function kernel, logistic regression, regression tree, and random forest. To complete the unsupervised learning task, we used PCA and K-means. Furthermore, we used caret package for parameter tuning. The parameter Kappa was the test statistic.

***Supervised Learning*:** We used the following R libraries: ISLR, MASS, caret, glmnet, tidyverse, elasticnet. The task involved classifying 400 observations (n) with 500 predictors (p) each into two classes. Since p > n, certain features could have existed that were not truly associated with the response; if we were to include such features when applying a classifier, this would lead to reduction in the quality of the fitted model and hence, an increased test error**.** In addition, many of these predictors were correlated with each other, which would increase the variance of their coefficients and lead to multicollinearity problems. First, we used **lasso** regularization as a means of variable selection method to prioritize most influential variables.

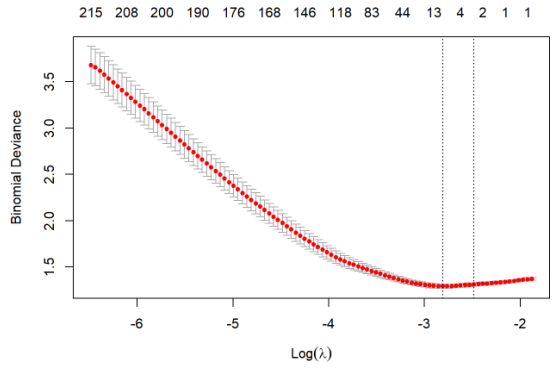


Figure 1

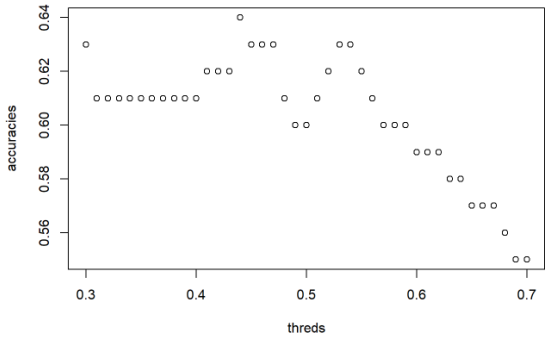


Figure 2

*Lasso*: In Figure 1, the first dashed line is the lambda with the smallest cross validation error. However, the lambda on the second dashed line was selected ($\lambda$=9). This lambda gives a binomial deviance of 0.06.

*Logistic regression*: In Figure 2, lasso was used to shrink the variables. With respect to the accuracy of the reduced model, results show it does not perform well. As shown in Figure 2, at 0.5 threads, the accuracy of the model is 0.6, which is not desirable in this case for 2 classes. It is relevant to mention that even the highest accuracy rate (0.64) is not desirable either.

*ElesticNet*: Next, **elasticnet** was used with the full dataset and 10-fold cross validated resampling. However, no desirable outcome was found by tuning different combinations of parameters. The accuracy rates for different combinations were not higher than 0.66. Accuracy was used to select the optimal model using the largest value. The final values used for the model were alpha = 0.9 and lambda = 0.07.

*SVM with a linear kernel*: Here, we used only 9 predictors and the full data set of 400 samples and resampled the dataset using 5-fold cross validation. The final value used for the model was C = 5, which gave an accuracy of 68.26% and a kappa value of 0.35.

*SVM with polynomial kernel*: Here, we also used a full dataset as input and 9 predictors. We resampled using 5-fold cross validation. We observed 2 to 4 degree polynomial kernels. The tuning parameter 'scale' was held constant at a value of 1. The final values used for the model were degree = 2, scale = 1 and C = 1. Highest value accuracy was used to fix the model parameters. Largest accuracy was 66.02%.

*SVM with radial basis function kernel*: For the full size dataset and 9 predictors, 5-fold cross validated resampling was used. C values were used in range of 20 to 100 with 10 increments, and sigma value used was in range 0.05 to 0.25 with increments of 0.01. The final values used for the

model were sigma = 0.05 and C = 20. For this model, the highest accuracy we obtained was 69.71%, and the kappa value was 0.38.

*Regression tree*: We used 9 predictors on a full dataset and 5-fold cross validated resampling. We tuned the cp value from 0.001 to 0.3 with an increment of 0.01. The final value used for the model was cp = 0.251 and accuracy was 69%. Next, we used the same tree model with all 500 predictors. The highest accuracy we obtained was 70.51% for a cp value of 0.071.

*Random forest*: Random forest was used on the full dataset with 500 predictors and 5-fold cross-validated sampling. The range of **mtry** used was 1 to 499 with increments of 20. The final value used for the model was mtry = 201, which gave the highest accuracy of 70.74%. Next, we reduced the number of predictors using the previous methods and used a random forest with 9 predictors. The largest accuracy we acquired was 71% for a mtry value of 5.

***Unsupervised learning methods*:** We tried clustering methods PCA and K-means for this purpose. We used the following R libraries: factoextra, purrr, cluster, and NbClust, gridExtra, ggplot2, and clValid. We used the dataset cluster_data.RData, which has 1000 input observations and 784 predictors.

*Principal component analysis*: Due the size of the dataset, PCA was conducted to reduce the original variables into principal components explaining most of the variance in the original variables. By performing this, we found that for the No.1 Principle component could only explain a proportion of variance of 8% (see Figure 3). In order to achieve a high cumulative variance, we would need to use 200 PC to reach 90% of total variance (see Figure 4). Thus, we picked three numbers of PCs to use, 100, 250 and 400.

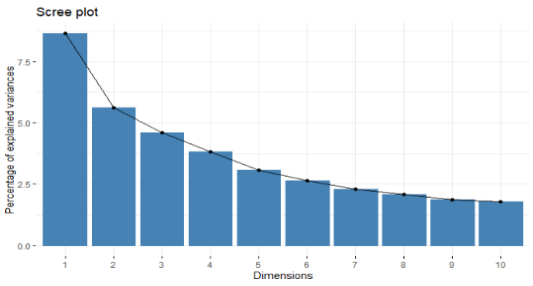


Figure 3

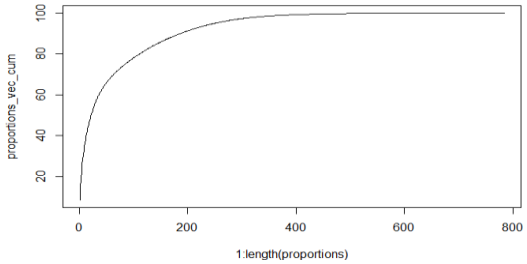


Figure 4

*Determining k*: Multiple methods were proposed to determine the k-value such as NbClust, community detection with Girvan-Newman process, and community detection with Louvain; however, none of them were successful in determining a desirable k. We then decide to use the method called Average Silhouette Method (ASM), which measures the quality of the clustering. A high average silhouette width indicates good clustering. Using this method, we used three PCA reduced models, with 100, 250, and 400 principle components. We used the nstart parameter = 25 and conducted 25 iterations. Then, based on the result of ASM, we suggest that there could be either 3 clusters, as shown in Figure 5, or 3 or 6 clusters, as shown in Figure 6. Based on the original dataset using K-mean, we know that k=6 is a promising result for this cluster; however, by using the PCA method and 250 principle components and performing the same ASM, our result indicates that k=3. The choice between 3 and 6 will be determined through the future use of this cluster, as 3 or 6 will perform differently in different situations. Thus, the optimal number could be determined through further clustering analysis, which would reveal the different strength on predictions or statistical inferences.

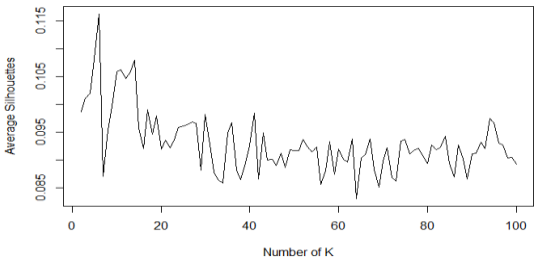


Figure 5

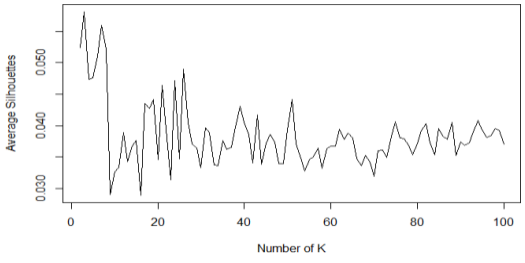


Figure 6