STAT 639 - 600 Grace Atama & 526005850

Group 13 Charan Jirra & 132005206

Jiahong Zhou & 928000924

Final Project: Classification and Clustering

Introduction and Methods

This project consists of two parts: the first part is a classification problem, and the second is a clustering problem. The methods used for each task are discussed directly below, and the corresponding results are reported in the subsequent section.

For the classification task, various supervised machine learning techniques are used to perform the classification of the data set and include Logistic Regression (LR), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Random Forest, Boosting, Bagging and Naive Bayes. Since the data is high-dimensional (500 features with only 400 observations), we can either directly use methods that deal with high-dimensional data, or we can perform feature selection first, then apply classification algorithms to the lower-dimension data. To estimate the test error, we use cross-validation. If there are tuning parameters that need to be estimated, nested cross-validation is applied to ensure that the estimation of parameters and test errors are independent of each other (Varma and Simon, 2006).

For the clustering problem, we first perform dimensionality reduction using PCA and utilize the following techniques to determine k, the number of clusters: Elbow method, Silhouette method, and Gap Statistic (Kassambara, 2018). The Elbow method consists in computing the within-cluster Sum of Square (WCSS) for different numbers of clusters, which is the sum of squared distance between each observation and the centroid in a cluster, and then

plotting the WCSS against the number of clusters. The WCSS decreases with increasing k values creating an elbow shape. The optimal number of clusters is the "elbow point" on the graph.

The Silhouette method consists in computing the silhouette coefficient, which is a measure of how close each point in one cluster is to points in the neighboring clusters, and plotting it against the number of clusters. The optimal number of clusters is the k value corresponding to the largest silhouette coefficient. Lastly, the gap statistic consists in comparing the total intracluster variation for different values of k with their expected values under the null reference distribution of the data and plotting the total intracluster variation against k. The optimal number of clusters is the k value corresponding to the largest total intracluster variation.

The clustering algorithms considered include K-means, PAM, CLARA, and Hierarchical Clustering.

Discussion and Results

For the classification problem, the results are presented in Table 1. Several feature selection techniques used are: 1) PCA to select the number of features that explain 80% of the total variance. 2) Two-sample t-test is used to select features with large absolute values of t statistic (Vabalas et al., 2019). 30 features are selected in this case. 3) Stepwise logistic regression selects the best subset of features based on the AIC criterion. 4) Boruta, which is a wrapper built around Random Forest that tries to capture all the important features in the dataset with respect to the outcome variable. Although estimated test errors are not obtained by cross-validation for all methods, for example, Random Forest returns an out-of-bag (OOB) error which gives an appropriate estimate of the test error. We will use the term "CV error" in the tables to denote the estimated test errors. For methods that do use cross-validation to estimate test errors, 10-fold cross-validation was employed.

Table 1. Estimate of misclassification test error for different supervised classification algorithms performed after various feature selection techniques.

Method	Penalized LR	Penalized LDA	Random Forest	Bagging	Naïve Bayes
CV Error	0.353	0.315	0.305	0.28	0.35
Method	PCA followed by LDA	T-test followed by LR	T-test followed by LDA	T-test followed by RF	T-test followed by penalized LR
CV Error	0.338	0.32	0.315	0.303	0.308
Method	T-test followed by penalized LDA	T-test followed by Naïve Bayes	Stepwise LR followed by Penalized LR	Boruta followed by SVM	Boruta followed by Boosting
CV Error	0.313	0.308	0.163	0.16	0.24

For the clustering task, we present k selected by each method in the table below. Since the elbow method is subjective, we will not include results of this method here.

Table 2. Estimated number of clusters

Clustering algorithm	**Model for Selecting k**	**Selected k**
K-means	Silhouette method	9
	Gap Statistic	10
PAM	Silhouette method	2
	Gap Statistic	10
Clara	Silhouette method	3
	Gap Statistic	2
Hierarchical Clustering	Silhouette method	8
	Gap Statistic	10

Conclusions

For the classification task, applying Boruta selection followed by radial SVM yields the smallest estimated test error of 0.16 with optimum parameters C = 1 and Sigma = 0.1129. Stepwise logistic regression followed by penalized logistic regression gives a comparable error

of 0.163. However, the problems with stepwise methods have been discussed in previous studies (Flom, 2018) and possible drawbacks include R2 values being highly biased, the F-statistics do not have the claimed distribution, and the standard errors of the parameter estimates being too small, and so on. Thus, we will not consider this method as one of the best approaches.

For the clustering task, both the results from PAM and Clara clusterings are discarded as they significantly differ from the K-means and Hierarchical clustering results. The optimal number of clusters was chosen to be 8 because K-means is restricted to convex data, and the silhouette method yielded a more conservative estimation of the number of clusters.

References

Flom, Peter. "Stopping stepwise: Why stepwise selection is bad and what you should use instead." *Towards Data Science*, 22 September 2018, https://towardsdatascience.com/stopping-stepwise-why-stepwise-selection-is-bad-and-what-you-should-use-instead-90818b3f52df

Kassambara, Alboukadel. "Determining The Optimal Number Of Clusters: 3 Must Know Methods." *Datanovia*, 2018, https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/#:~:text=The%20optimal%20number%20of%20clusters%20can%20be%20defined%20as%20follow,sum%20of%20square%20(wss)

Vabalas, Andrius, et al. "Machine learning algorithm validation with a limited sample size." *PLOS One*, 2019, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0224365

Varma, Sudhir, and Richard Simon. "Bias in error estimation when using cross-validation for model selection - BMC Bioinformatics." *BMC Bioinformatics*, 2006, https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-91