Final Report

Samuel Burge, Chad Austgen, Skylar Liu

April 19, 2022

Classification Task

Methodology

For the classification task we approached the problem with a spread of well-known classification algorithms. This included the Naive Bayes classifier, regularized logistic regression (specifically ridge, lasso and elastic net regularization), gradient boosted trees, random forests and support vector machines (using radial and polynomial kernels). For model selection and assessment, we utilized 10-fold cross-validation and, with some algorithms, another nested cross-validation to select tuning hyper-parameter using a grid search. Nested cross-validation is a well-known approach to handling both tuning and assessment of models with hyper-parameters, and we opted to employ this approach to avoid potential over-fitting and bias introduced when the same data used to validate the models is also included in hyper-parameter optimization (Cawley and Talbot 2010).

Logistic Regression

We first verified that our assumptions, which validates logistic regression as a methodology, were satisfied. Before the analysis we created large plot matrices and verified across all

predictors that X and Y were not well separated. However, we were not able to fit a logistic regression model outright since $p > n$. We utilized regularization methods to reduce the features in our analysis using lasso, ridge, and elastic net regularization methods (Zou and Hastie 2005).

```
##             Training Error Test Error
## Lasso            0.3201453  0.3509644
## Net              0.3201453  0.3862684
## Ridge            0.2016434  0.4178184
## Naive Bayes      0.3783479  0.4081633
```

The results of our analysis are depicted in the table above. Note the degree to which ridge logistic regression under-performed during the assessment despite a low training error. This is likely over-fitting, and we attribute this to high excess variance in the data set compared to bias. The lasso and elastic net eliminate this variance by performing variable selection to eliminate predictors.

Boosted Trees

Boosted classification trees yielded the best results of all the algorithms we employed. In order to generate our results we used nested cross-validation method to perform a grid search over various shrinkage rates, bag inclusion, and interaction depths. We then compared the models with a withheld validation set. We found that as is common with validation set approaches there was significant variance in the accuracy of the model with respect to the validation set. In order to identify higher performing models we resampled repeatedly and fed the data into the cross validation trees algorithm and measured against a corresponding validation set. We chose our model manually using a slight bias toward models which were simpler and had less variance.

Table 1: Results from 10-fold cross-validation to assess model generalization performance.

	Training Error	Est. Test Error
Lasso	0.0132565	0.0490566
Net	0.0132565	0.0509434
Ridge	0.0144092	0.0490566
Naive Bayes	0.1000000	0.0566038
Radial SVM	0.0426387	0.0584906
Poly. SVM	0.0414706	0.0547170
Sigmoid SVM	0.0584906	0.0000000
Random Forest	0.0000000	0.3125000
Boosted Trees	0.0005764	0.0000000

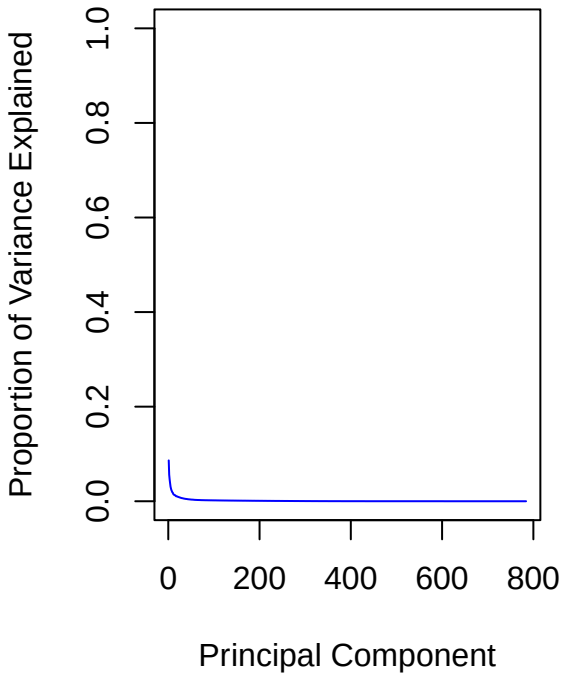
Results

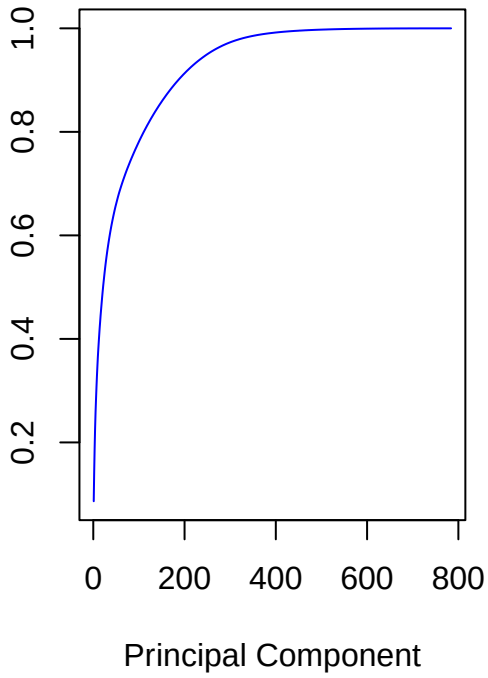
We can put a description of the results, like a table of the average training and cross-validation errors, in this section. I was also thinking we could include ROC curves.

Clustering Task

Pre-processing

For the clustering task, we initially looked at performing principal components analysis (PCA) to reduce the number of dimensions in the data set. The two scree plots below show that the number of principal components necessary to capture at least 90% of the variation in the data set was about 187, which did not seem beneficial enough to consider for the analysis. Therefore, we decided to retain all the original features in the data set.





Methodology

We opted to use k-means and hierarchical clustering, in part because the high-dimensional data is intractable with DBSCAN and due to computing restraints. Since the given data does not have any contextual basis for selecting the number of clusters K, we need one or more criteria to determine the number of clusters. Several well-known and widely used approaches include the elbow method, the silhouette method (Rousseeuw 1987), and more recently the use of the gap statistic (Tibshirani, Walther, and Hastie 2001).

References

Cawley, Gavin C., and Nicola L. C. Talbot. 2010. "On over-Fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation." *Journal of Machine Learning Research* 11 (70): 2079–2107. http://jmlr.org/papers/v11/cawley10a.html.

Rousseeuw, Peter J. 1987. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." *Journal of Computational and Applied Mathematics* Volume 20 (https://doi.org/10.1016/0377-0427(87)90125-7): 53–65.

Tibshirani, Robert, Guenther Walther, and Trevor Hastie. 2001. "Estimating the Number of Clusters in a Data Set via the Gap Statistic." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63 (2): 411–23. https://doi.org/https://doi.org/10.1111/1467-9868.00293.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67 (2): 301–20. http://www.jstor.org/stable/3647580.