Group Project:
Evaluation of Statistical Model Accuracy

Bryce Robinette Jacelyn Villalobos Kursten Reznik
David Koster Jacob Ruiz

10/17/2020

Contents

1 Introduction

In this paper, we compare the performance of four statistical models using randomly generated data with normal distribution $N(\mu, \sigma^2)$: The models we are investigating are: K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and the Logistic Regression model (GLM in R). That is, we analyze these models' accuracy with respect to sample size, number of input variables, the variance of the data, and data scaling.

We have been given four precise scenarios for which these models are to be tested. We present them here:

Scenario 1
·Sample Size $N = 50$.
·Number of Inputs $p = 2$.
·No scaling or normalizing the data.

Scenario 2
·Sample Size $N = 500$.
·Number of Inputs $p = 2$.
·Normalized Data.

Scenario 3
·Sample Size $N = 100$.
·Number of Inputs $p = 20$.
·No scaling or normalizing the data.

Scenario 4
·Sample Size $N = 500$.
·Number of Inputs $p = 20$.
·Normalized Data.

In each of these scenarios, we will investigate the average accuracy of each model as it relates to the variance of the data. It should also be noted that this paper assumes that the reader is familiar with these statistical models and the programming language R.

2 Materials and Methods

To begin, we perform each model with the parameters outlined in the four given scenarios. These four scenarios will give us our blueprint for analysis. Indeed, for each scenario, we run each model against a number of data sets of identical properties but with increasing variance in our randomly generated data. Each of these data sets are also sampled many times in order to return each model's average accuracy for that data's given variance.
Moreover, we then evaluate the accuracy of each model as a function of the variance of the data sets in each scenario and plot the results. This methodology should yield the generalized accuracy of the models given the variance of the data. It should be noted that for this report, we choose our variance s such that $0.1 \leq s \leq 2$.

2.1 Data Generation

We generate our data such that the response variable y has two classes. That is, $y = 1$, or $y = 2$ with class means $\mu_1 = 1$ and $\mu_2 = 0$, respectively. Then we randomly generate values with distribution $N(\mu, \sigma^2)$ for our predictors.

It should also be noted that when we have a smaller amount of observations N, there is a chance that when sampling the data into testing and training data that we could end up with an imbalanced sample of our response value. To mitigate this, we could sample an approximately equal amount of data with each value of y; however, since we are running a multitude of simulations on the data, we can rely on the central limit theorem to obtain our average accuracy.

2.2 Choosing our best k for KNN

In this section we present our best k. The very best k. In fact, it is the best k in the history of k's. So good it will blow your mind. Here is how we did it....

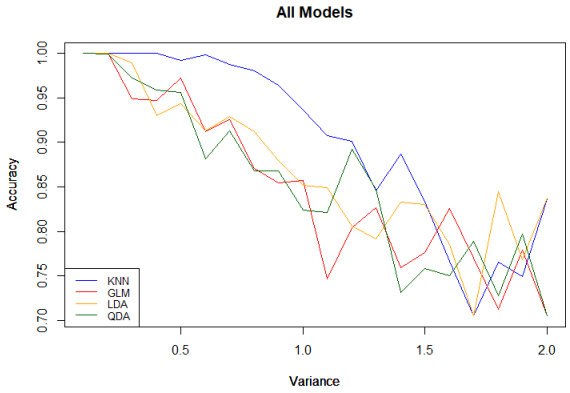
We simulated our KNN model over a thousand generated data sets with differing variances and chose the most common "best k" that was associated with the data sets. We took this common k to be our generalized k-value for our following analysis. We found that the best value of k for our generated data sets was $k = 3$. We keep this value constant throughout the different scenarios.

3 Scenario Analysis

3.1 Scenario 1

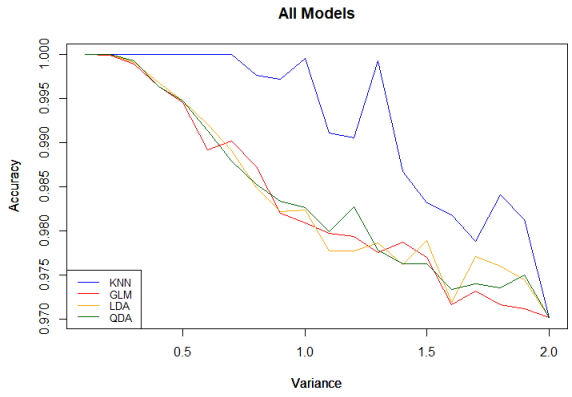
In this scenario, we compare our models with $N = 50$ observations, $p = 2$ parameters, and without scaling the data.

From the resulting plot, we can conclude that all models diminish in accuracy as the variance increases. This is to be expected given that our data was generated randomly without any underlying function dictating the data generating process. Indeed, LDA and Logistic Regression assume a linear relationship between the response variable and the predictors, while KNN does not make any assumptions about the underlying data. QDA is a decent compromise between KNN and any linear model as it assumes a nonlinear relationship between the predictors and the response.



As is evident from the resulting plot, the KNN model performs the best. This is to be expected due to the fact that we have chosen a relatively small k for our analysis which makes our particular KNN model more flexible than it would be with a larger k-value. Furthermore, KNN performs best with smaller data sets, which is what we have here. In this regard, we might expect our KNN model to perform better under the assumed circumstances for *Scenario 1*.

3.2 Scenario 2



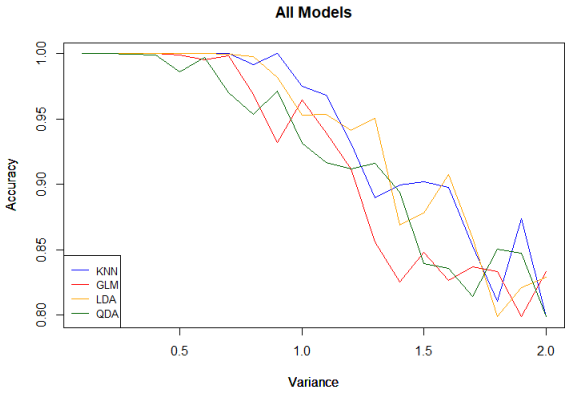
For scenario 2, we take $N = 500$, parameters $p = 2$, and we scale the data.

We immediately see that the KNN model performs the best. Indeed, as a general rule, when we are presented with a large number of observations and the value of our predictors is small, we want to choose a more flexible model. Due to the large sample size, we are less likely to over-fit, even with a more flexible model. Since we have chosen a relatively low value for k, our KNN model has the greatest flexibility and hence, the greatest accuracy over the different variances.

We must note that all of the models performed with greater accuracy when the data was scaled. Scaling the data is going to reduce the overall variance in the data. If we have a data set with wide variation, scaling can become a necessary step in order to preserve the accuracy of statistical models. So in this particular scenario, all of our models seem to be doing well; however, the flexible nature of the KNN model makes it a superior predictor in this case.

3.3 Scenario 3

Now we take $N = 100$, and increase our predictors to $p = 20$. The data is not to be scaled in this evaluation.
Note that we take a larger value for N than the other scenarios since each level, or predictor, needs some amount of observations. If there are not enough observations in a group, then QDA in R will throw an error.

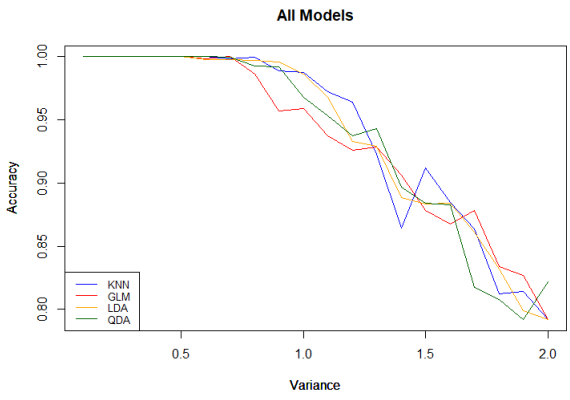


From the resulting plot, we immediately conclude that increasing our predictors p has yielded higher model accuracy over larger variance of the data in comparison to the results from *Scenario 1*.
Since we have defined the number of predictors to be $p = 20$, it may result that LDA and QDA perform better since we have increased our parameters. That is, QDA (in general) needs a larger amount of predictors, however since we are not allowing the models the convenience of *"best predictors"*, if the predictors have little significance, then they will not significantly contribute to the overall model accuracy and may lead to an over-fit and little change in predictive accuracy.

It could be concluded in cases where data variance is high, LDA and QDA may perform better with a larger number of predictors. However, we should be weary of increasing the number of predictors too high such that we start over-fitting and incorporating too much noise in our models. Furthermore, KNN now performs similar to the other models. KNN begins to struggle when the number of inputs is large.

3.4 Scenario 4

We now increase our number of observations to $N = 500$, keep our parameters $p = 20$, and scale our data.



As we expected from the increased number of input variables, KNN does not outperform any of the other models. The larger number of input variables will help LDA, QDA, and Logistic Regression due to their parametric nature, but diminishes the average accuracy the KNN since it may incorporate "neighbors" that have no relationship to the desired response. Indeed, KNN's accuracy reduced significantly because KNN struggles when there is a large data set and a

large amount of predictors. Now, when the data is scaled, the KNN will increase in accuracy but still does not outperform the other models.

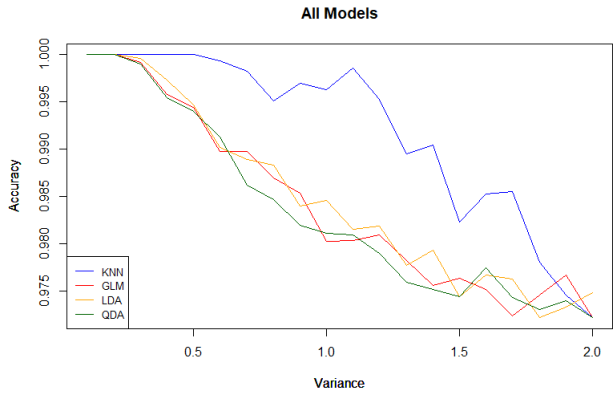
As in *Scenario 3*, we need to be aware of the possibility of over-fitting due to larger number of predictors. However, since we know the details of our data and the generating process, we need to worry about this less than we would when dealing with other data sets of a more natural origin.

4 Conclusion

Overall, as the variance in our data increases, the average accuracy decreases throughout all of the different models. This is to be expected given the effect variance has on model accuracy.

It seems to be apparent that when given a smaller amount of observations and predictors, our k-nearest neighbors is likely to perform the best under these circumstances. Indeed, we may conclude that given a smaller number of predictors, we may rely on the KNN model to perform with highest accuracy under our given circumstances.

The following graph yields some insight in to this statement.



The graph that is displayed above is the result of *Scenario 2* but without scaling the data. It is presented to show that even without scaling our data, a smaller number of predictors still allows KNN to outperform the other parametric models.

This follows from what we may assume about the other models being of a parametric nature. Since LDA, QDA, and Logistic Regression are all parametric, we should see their average accuracy increase when introduced with more parameters. Of course, as was stated in some of the previous sections, when we start introducing more parameters, we should always be

aware of the possibility of over-fitting our models. Again, however in the particular case of this report, we know the nature of our data and the generating process. Hence we worry less about over-fitting or including too much noise in our data.

This goes further as we look at the scenarios that have a larger number of observations and predictors. Without being weary of over-fitting, we see that the average accuracy of the parametric models performing better while our non-parametric KNN becomes less accurate as it struggles with large amounts of observations and predictors.

We must also author a note about the values we chose for our variance, s. There may be more insight to glean if we increase the range of variance in the data. However, in the purview of this report, we conclude that for smaller data sets and a lower number of predictors, we rely on the KNN model to outperform the others. While we see an almost identical degradation in accuracy in all four models when our data set is larger with a larger number of predictors, in reality we may see that the parametric models will perform better as long as the increased number of parameters has statistical significance to the response.

Without further speculation, and given the restrictions of this report, we are confident in the conclusion thus far; however, we can not confidently conclude a *best model* in a generalized sense. In spite of a non-conclusory "best model", it is hoped that the information yielded by this report further informs us on the nature of these statistical models, allowing us to be more discerning in our future endeavors when presented with real-world data.