Group Project:
Evaluation of Statistical Model Accuracy

Bryce Robinette   Jacelyn Villalobos   Kursten Reznik

David Koster   Jacob Ruiz

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 normalization of the data.

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

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

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

Scenario 3
·Sample Size $N = 50$.
·Number of Inputs $\rightarrow 20$.
·No scaling or normalizing the data.

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

In each of these scenarios, we will look at the 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.

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. In this report, we assume variance s such that $0.1 \leq s \leq 2$. This methodology should yield the generalized accuracy of the models given the variance of the data.

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 $N(\mu, \sigma^2)1$ 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.

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 number of generated data sets with differing variances and chose the most common "best k" that was associated with these particular data sets. We took this common k to be our generalized k-value for our following analysis. In plotting the error rate versus the variance we found that the best value for our generated data was $k = 3$ for the majority of the different sets. We will keep this value constant throughout the different scenarios.

Do we want a KNN best k plot? I dunno... maybe?? It would look good but I dont have too much invested in having it.
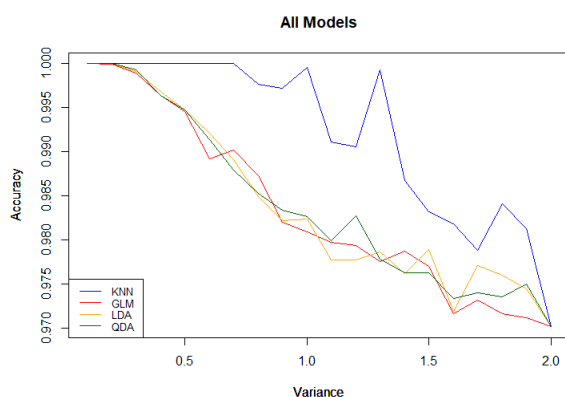
Scenario 1

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

From the resulting figure, 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 is non-parametric. QDA is a decent compromise between KNN and any linear model as it assumes a nonlinear relationship between the predictors and the response.

In this scenario, 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. In this regard, we might expect our KNN model to perform better under the assumed circumstances for scenario 1.
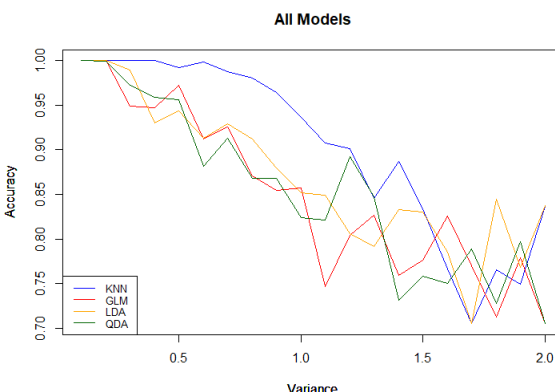
Scenario 2

In this scenario, 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.

Scenario 3

Now we take $N = 50$, and increase our predictors to $p = 20$. The data is not to be scaled in this scenario.

All Models
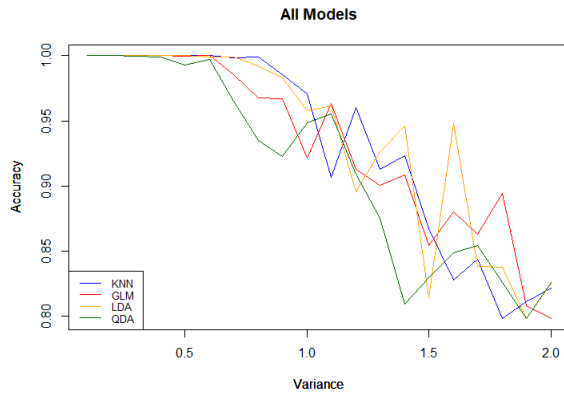
From the resulting plot, we immediately conclude that an increase in our predictors p has yielded higher model accuracy over increasing variance of the data in comparison to the results from scenario 1.

Since we have designated or 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.

• Maybe write something about scenario 3 having a problem with QDA. When we have too small a sample size to compare with our predictors, then shit gets whack.

NO... Dr, Chan. we must have more data. If you don't like it, its Davids fault

4

Scenario 4

We now increase our number of observations to $N = 500$, keep or 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. KNN's accuracy level 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 slightly increase in accuracy level but still does not outperform the other models.
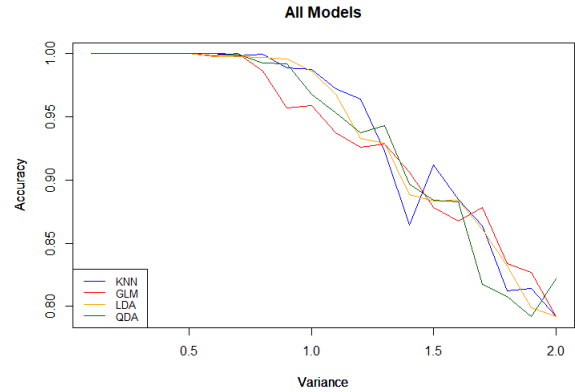
As in scenario 3, we need to be aware of the possibility of over-fitting due to larger number of predictors. However, since we know the details of our data and generating process, we need to worry about this less than we would when dealing with other data sets.

Conclusion

Overall, as the variance increases, the level of accuracy decreases throughout all of the different models. This is to be expected given the effect variance has on model accuracy. In each of the scenarios KNN outperformed the other models because it is the most flexible and does not require linear decision boundary, but this is not always the case, especially for scenario 4. As you can see, KNN's accuracy level reduced significantly because KNN struggles when the number of inputs is large.

Extensions?

Instead of normally distributed data, we could assess performance or accuracy of the models over other distribution sets, such as a uniform distribution.

1 Credits

McDonald's: Obviously.

Alcohol: for its continuing effort to keep me sane.

Sir David Attenborough: for the love of all that is good.

I dunno... maybe my cat?

References

ME... FOOLS... I am the only reference you need.

JK seriously give yer references..

Writing Notes

Use your eyes and read:
KNN will support non-linear solutions where GLM can only support linear solutions. Furthermore, KNN cannot yield confidence levels due to its selection nature where GLM can because it has the property of linearity.

Neighborhoods have meaningful weight....

The number of parameters needed to estimate QDA increases faster than LDA. The number of parameters estimated in LDA increases linearly with p while that of QDA increases quadratically with p. So, from the resulting analysis, we can see that when we have a larger data set and the number of predictors is large, the Quadratic Discriminant Analysis performed better than the linear models due to the fact that with a larger set if data(larger variance) in scenario 4 we are give a larger value of predictors, which in turn benefits the quadratic model to accept nonlinear relationships between the response variable and the the predictors.

as the number of predictors increases we will need to use a more inflexible model. Having a high number of predictors can lead to over fitting of the data.

QDA differs from LDA by assuming the variance or variance-covariance matrix of the feature(s)varies from class to class. This allows for more flexible and non-linear decision boundaries, but requires estimation of more parameters. As with all other models we've seen, estimating more parameters increases the likelihood of over-fitting and so should only be used when the number of observations is large relative to the number of features and classes.