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

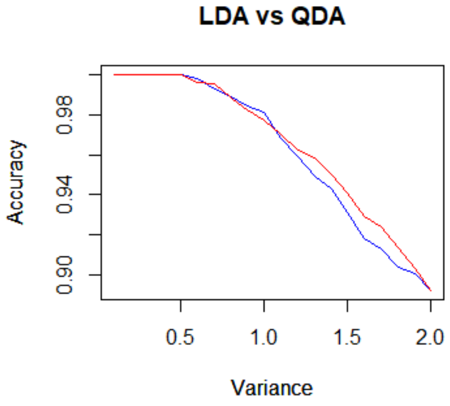
To begin, we perform each model with the parameters outlined in the four given scenarios. This gives us our starting point for our analysis. Then, for each scenario, we run each model against a number of data sets of identical properties but with increasing variance in our randomly generated data.
Indeed, we then evaluate the accuracy of each model as a function of the variance of the data sets in each scenario. This methodology should yield the generalized accuracy of the model given the variance of the data. Furthermore, we compare the accuracy of the models to gain insight into which model should likely be used given a set of data of normal distribution.

Data Generation

We generate our data such that the response variable y has two classes. That is, $y = 1$, or $y = 2$. Then randomly generate values 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 sample an approximately equal amount of data with each value of y.

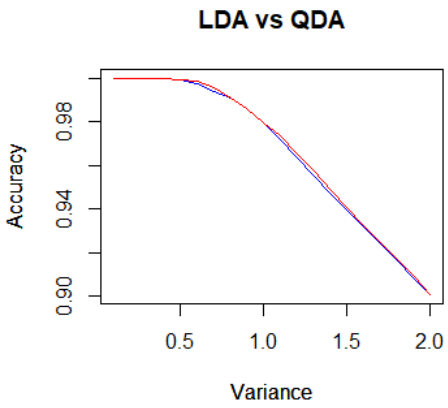
Scenario 1



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

I haven't compared the KNN model yet but I have compared the LDA and QDA. As you can see from the resulting figure, both models diminish in accuracy as the variance increases and at a very similar rate. This is to be expected since our data was generated randomly without any underlying function dictating the data generating process. Indeed, LDA assumes linear relationship between the response variable and the predictors while QDA assumes a nonlinear relationship. In our particular set of data, we do not have the luxury of these assumptions. So, since there is no discernible relationship between the response y and our parameters, then these two models will behave similar to one another, making one no more reliable than the other in predictive accuracy.

Scenario 2

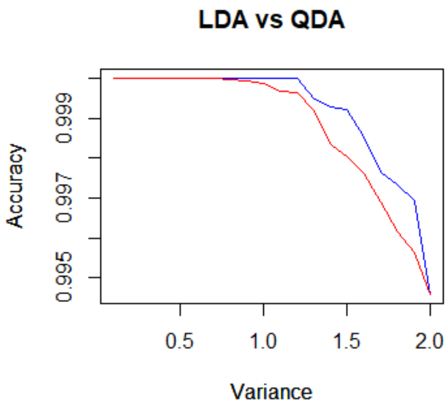


In this scenario, we take $N = 500$, parameters $p = 2$, and we normalize the data.

I can write more about this later, but as we can see form the resulting plot, both LDA and QDA did better in their predictability averages but still follow an almost identical pattern of decline in accuracy as the variance in data increases. Again, this is to be expected given the nature of our data.

Again, I will incorporate the other two models as their information becomes available.

Scenario 3



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

From the resulting plot, we imediatly conclude that an increase in our predictors p has yielded higher model accuracy over increasing variance of the data.

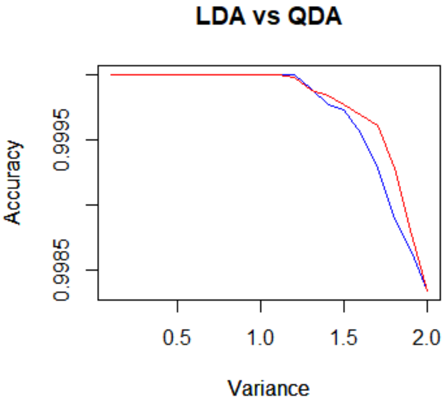
I suppose that one may conclude in cases where data variance is high, LDA and QDA may perform better with a larger number of predictors. However, we should be wearing of increasing the number of predictors too high that we start over-fitting and incorporating too much noise in our models.

Comparisons of KNN and GLM are to come as their information is made available.

Scenario 3

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

Scaling the data has further increased the average accuracy of the LDA and QDA models. Again, in a not so surprising way, LDA and QDA still follow highly similar paths of accuracy decline as the variance increases.



More on the other two models as their information becomes available...

GLM (to be rewritten later as part of the scenario analysis)

Logistic regression is going to have a different set of standards than our other models (except LDA). It is assumed that the reader recollects that Logistic Regression is parametric, while KNN is non-parametric.

Use your brain for a moment.... You then realize that LDA and GLM require linearity for any meaningful result. QDA is a decent compromise between KNN and any linear model, (such as GLM), since QDA also assumes non-linearity.

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....

In this regard, I have to say at this particular moment, that *Logistic Regression* is a largely unnecessary model when dealing with data generated with a known normal distribution without a further informative function dictating the outcome or response variable.

Like, for reel check it motha fuggas..... GLM and LDA are based on identical assumptions of linearity and statistical regression, making one no more informative than the other in our particular cases with randomly generated data derived from $N(\mu, \sigma^2)$.

Furthermore, with our particular method of data generation, QDA falls on its face as well. QDA assumes a non-linear result from our response variable when functioned with our parameters. Again, when we randomly generate our data of the distribution $N(\mu, \sigma^2)$, we circumvent this potentiality and render QDA almost obsolete. Therefore, KNN is the only model that has any reliable capability of prediction. No doubt, with a sufficient amount a data sets, we will have some that follow linearity and some that follow a quadratic nature; however, if we were to ascertain the reliability of a model in ***any*** case, then KNN would have to be the test first assumed by the statistician.

Discussion

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 Conclusion

2 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..