Group Project

Bryce Robinette, et al.

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.

This highly unnecessary since for our GLM model we need our data to represent 0s or 1s. So, personally I would like to convert everything to 0s and 1s in the first place.

It might be real shitpile for Kursten if we have to convert based on the medians of independently generated data sets for the GLM model. Poor Kursten...

It is outlined below in the GLM section, but with properly randomly generated data sets, KNN is the only model that can be relied on for any meaningful result.

```
data.generate = function(mu1, mu2, s, n, p){
  y = as.factor(c(rep(1,n), rep(2,n)))
  M = matrix(NA, nrow = 2*n, ncol = p, byrow = 1)
  df = data.frame(y)
  for (i in c(1:p)){
    M[,i] = c(rnorm(n,mu1,s), rnorm(n,mu2,s))
  }
  df = cbind(y,M)
  df = data.frame(df)
  return(df)
}
```

Sampling Testing and Training Data

We obviously sampled our data into training and testing data. I mean, duh... right?

Performances

In this section we outline the performance of each model under the given scenarios.

KNN

Since you and I are both smart, we both can collectively nod our heads at one another and say, "I know what KNN does." Indeed, from what we know, as the variance of our data increases, our reliable predictability will diminish.

David is krankin out mad knowledge as we speak to discover the underlying properties.

LDA

LDA and QDA are going to be similar in any of our cases since our data is randomly generated without a specific linear or quadratic function to dictate the generation process.

QDA

...See LDA section above....

GLM

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

Furhtermore, 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?

1 Conclusion

You should see me shirtless. Its been getting better since I decided to get in shape.

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... Bitches... I am the only reference you need.

JK seriously give yer references..