Lachlan Sinclair

2/17/2020

Assignment 2 Reflections

1. In the lecture for linear discriminants and in the textbook, it is mentioned that the way logistic regression classifies the binary result inherently limits the constraints on the distribution of classes. This means there are certain situations where a single separatrix can't be used to effectively divide the classes. The textbook gives examples of datasets that have their classes distributed into smaller dispersed clumps rather than in two distinct clumps. In these cases, the repeatedly drawn division lines perpendicular to the axis's that represent how a tree model can divide the dataset clearly provides a better model. This was made abundantly clear by the churn example performed in this assignment, using a logistic regression function to predict the results was only 63% accurate which is not ideal. Using the same dataset in the previous weeks we saw that tree models were able to do a significantly better job. This difference was because the churn data set doesn't cleanly separate itself into two groups that can be split by a single linear line.

2. In this assignment we explored an example of trying to use logistic regression on the churn dataset. This example clearly showed that even though this is a larger dataset, when compared to the breast cancer dataset, the predictive ability of the model is clearly worse. This has to do with the fundamental way the model is designed and developed in relation to the way the classes are distributed in the dataset. Increasing the training dataset size will not make significant increase in the predictive ability of the model in this scenario. I think this same logic can be applied to a tree model as well, however that is not the topic of this question. The conclusion drawn from the churn example can be abstracted to say, the model design, not just the size of the dataset, plays a key role in ensuring the model is well suited for a particular dataset. Regardless of how well a given model is performing, increasing the training dataset size might not impact its performance. Logistic regression models are significantly impacted by a range of factors other than the training set size. These factors bind its performance regardless of changes made to the training set size. Not to mention other problems can be introduced when increasing the size of the dataset, such as overfitting.