STA4241 Interactive Lab Week 5: Differences between classification algorithms

Today we are going to learn more about the similarities and differences between LDA and logistic regression for classification. Logistic regression is rarely used when the number of classes is greater than 2, so we will focus today on the setting where there are only two classes.

(1) First we will show that in the two class setting, the LDA approach implies a logistic model that is linear in x, which shows that LDA and logistic regression are very similar. The key difference between the two is the manner in which the logistic regression coefficients are estimated.

(2) Let's now highlight some differences between LDA and logistic regression. In particular we will focus on the assumptions that LDA makes. The two assumptions that we will evaluate are the normality assumption, and the assumption of common variances. To do this we will run a simulation study with the following steps:

1. Generate one large testing data set
2. Generate 100 training data sets
3. Fit both the logistic and LDA models on the training data sets
4. Predict the outcome for the testing data set using the fits of the two models
5. Compare error rates

To evaulate the assumptions inherent to LDA, we will perform this simulation study on four different types of data:

1. Covariate is normally distributed
2. Covariate follows a t-distribution with 2 degrees of freedom
3. Covariate follows a gamma distribution
4. Covariate follows a normal distribution, but the variance differs in the two groups

Situations 2-4 break the assumptions of LDA and it is not clear how well it will perform in these situations. The simulation will provide us with some insight into the robustness of LDA to these assumptions.

(3) Given these results, when would we use LDA over logistic regression?

(4) Now let's look at a case with data that is separated and see how logistic regression performs

.