STA4241 Homework 2, Fall 2021

Please turn in your own work, though you may discuss these problems with your classmates, professor, and TA. The assignment is due on Wednesday, September 29th at midnight.

(1) Assume that our outcome Y is binary and that we have only covariate x. Show that quadratic discriminant analysis (QDA) implies a logistic regression model of the form

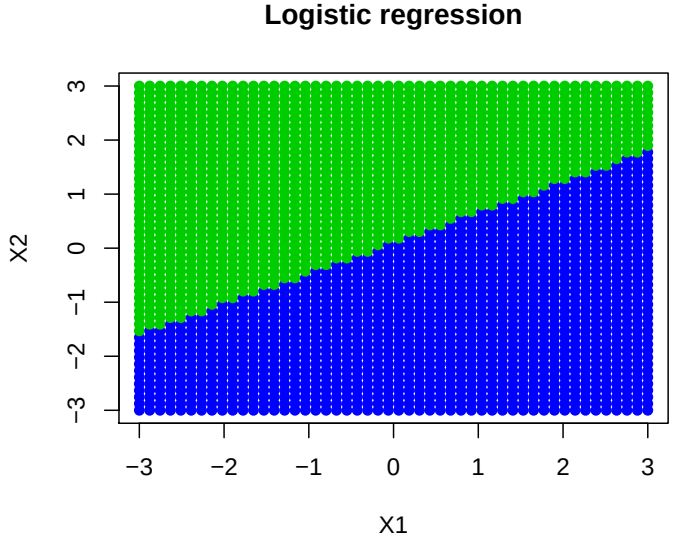
$$\log\left(\frac{P(Y=1|X=x)}{1-P(Y=1|X=x)}\right) = \beta_0 + \beta_1 x + \beta_2 x^2$$

and find the values of β_0, β_1 and β_2.

(2) Read in Problem2.csv off of the course website. There is a binary outcome Y and two covariates X_1 and X_2. I want you to fit four distinct models to this data to try and predict the outcome:

1. Logistic regression with linear terms for the covariates
2. Logistic regression that includes squared terms for X_1 and X_2
3. Linear discriminant analysis
4. Quadratic discriminant analysis

(i) For each of the four approaches, produce a plot that highlights the boundary region that separates the two classes. It should look something like the following, where green denotes areas that we would classify as $Y = 1$ and blue denotes $Y = 0$



(ii) Comment on any differences you see between the approaches. Do you think any of the approaches are overfit to the data? Do you have any way of knowing this without additional information?

(iii) Now read in the testing data set called Problem2test.csv and calculate the error rates on the testing set for each of the approaches. Comment on your findings.

(3) Read in Problem3.csv off of the course website. Now the outcome has four classes, i.e. $Y \in \{0, 1, 2, 3\}$. We still have two covariates in the data set. For this problem you only need to fit the LDA and QDA approaches.

(i) For each of the two approaches, make a plot that highlights the decision boundary that is analogous to the one done on the previous problem. The one difference, is that there should now be four regions or colors in your plot representing the four outcome classes.

(ii) Read in Problem3test.csv off of the course website and find the testing error rates for both approaches. Comment on your findings.

(iii) Should it concern you that your error rates are greater than 50%?

(iv) Do you think that your QDA model is an improvement on random guessing? By random guessing I mean randomly picking a class category with equal probability for each class.

(4) In this question I want you to run a simulation study to compare the performance of LDA and QDA. A simulation study requires you to generate a large number of data sets, then apply both LDA and QDA to each data set, and compare the performance across all data sets to evaluate which method outperforms the other. I want you to run a simulation study that highlights a situation in which QDA would outperform LDA in terms of prediction performance. For each data set, generate two covariates X_1 and X_2 from a standard normal distribution. Generate a binary outcome Y any way you like, but it should be done in a way that QDA will be expected to outperform LDA on average.

(i) Explain exactly how you generated data in your simulation study. It should be explained in sufficient detail so that I could replicate your results.

(ii) Explain why you think QDA will outperform LDA in this situation.

(iii) Either through a plot or a table, visualize your results comparing LDA and QDA in your simulation study.

(iv) Does the relative performance of LDA/QDA depend on the sample size in your situation? Ideally to answer this question, you would plot your results as a function of sample size to see how the results vary with n.