# STATS 415 Homework 4

Due Thursday Feb 8, 2018

**Please include your name, uniqname, and lab section (number or time or GSI). A point will be taken off homework without the section info.** Turn in a printout of your homework in the lecture or in your GSI's mailbox across room 305A West Hall, no later than 5pm on the due date.

1. Consider the (training) data in the table below, with one predictor $x$.

| $x$ | -4 | -1 | 0 | 1 | -1 | 2 | 3 | 4 | 7 |
|---|---|---|---|---|---|---|---|---|---|
| $y$ | -1 | -1 | -1 | -1 | 1 | 1 | 1 | 1 | 1 |

   (a) What are the class-specific parameters needed to specify the LDA and QDA classifiers, respectively? What are their estimated values from these training data? (Note: you can use R to compute means and variances, and generally to do arithmetic. You cannot answer this question by calling the lda() function).

   (b) Write down the discriminant functions for both LDA and QDA, with numerical values for coefficients. State the rule each method uses to assign the value of class variable $y$ given a specific value of $x$.

   (c) Compute the training errors for both LDA and QDA.

   (d) Compute the test errors for both LDA and QDA using the following test set:

| $x$ | -1.5 | -1 | 0 | 1 | 0.5 | 1 | 2.5 | 5 |
|---|---|---|---|---|---|---|---|---|
| $y$ | -1 | -1 | -1 | -1 | 1 | 1 | 1 | 1 |

   (e) Which do you think is more suitable for this data set, LDA or QDA? Justify your answer.

2. In this problem, you will develop a model to predict whether a given car will be classified as having high or low gas mileage based on the `Auto` data set.

   (a) Create a binary variable, `mpg01`, that is equal to 1 if the value of `mpg` for that car is above the median `mpg`, and 0 otherwise. You may then want to use the `data.frame()` function to create a single data set containing both `mpg01` and the other `Auto` variables.

   (b) Make some exploratory plots to investigate the association between `mpg01` and other variables. Describe your findings. Which of the features seem most likely to be useful in predicting `mpg01`? Scatterplots and boxplots may be useful tools to answer this question. (Note: do not use the `mpg` variable that was used to create `mpg1`).

   (c) Split the data into a training set and a test set: fix the random seed to the value 12345, and randomly select 80% of the observations (round down to the nearest integer) from *each* class to be the training data. Use the rest as test data.

   (d) Perform LDA on the training data in order to predict `mpg01` using four quantitative variables that seem most associated with `mpg01` based on (b). Report the training and test errors. Make a plot of the training data points, using two variables which appear to be most associated with the class as your axes. Using different colors to show the true values of `mpg01`, and different plotting symbols to show predicted values.

   (e) Perform QDA on the training data in order to predict `mpg01` using the same variables you used for LDA. Report the training and test errors. Make a plot analogous to the one you made for LDA.

   (f) Compare and contrast the performance of LDA and QDA. What do your results suggest about the class-specific covariances?

Please limit your solution to at most 7 pages.