STATS 415 Homework 4

Due Thursday Oct 8, 2020

Turn in a pdf scan of your homework on Canvas. Please limit your answer to Q2 to 8 pages, organized into a coherent typed data analysis report. Answers to Q1 may be either typed or handwritten. Please clearly write your name, your UMID, and your GSI/lab number on the homework..

1. Suppose you have one continuous predictor X and a binary categorical response Y, which can take values 1 or 2. Suppose you collected training data from the two classes and obtained class-specific sample means $\hat{\mu}_1 = -1$ and $\hat{\mu}_2 = 3$, along with the pooled variance estimate over the two classes, $\hat{\sigma}^2 = 1$. (40pt total, 8pt for each question)

   (a) Assume equal class priors and derive the LDA classification rule for this problem. Sketch the estimated class-conditional densities and show your decision boundary on the plot. Make sure you label the axes and indicate the numerical value for the boundary; let's call it c.

   (b) Suppose the estimates were in fact obtained from 100 training points, among which 40 were from class 1 and 60 were from class 2. Suppose now you will estimate class priors from data, repeat all the calculations in part (a) and obtain a new boundary value, let's call it $\tilde{c}$. Without actually doing this, would you be able to tell whether $\tilde{c}$ will be the same as, less than, or greater than c, or is there no way to tell? Explain your answer without calculating $\tilde{c}$. Note: It's ok to recheck your answer once you have actually calculate $\tilde{c}$ in part (c), but your explanation must not involve the numerical value.

   (c) Now calculate the new boundary value $\tilde{c}$ described in part (b).

   (d) Suppose in addition to the pooled covariance value $\hat{\sigma}^2$ I now tell you the individual class specific covariances were estimated as $\hat{\sigma}_1^2 = 0.25$ and $\hat{\sigma}_2^2 = 1.5$. Based on this new information, would you recommend using LDA or QDA, and why?

   (e) Derive the QDA rule for part (d), assuming equal class priors.

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 in the `ISLR` package. (60pt total, 10pt each question)

   (a) Create a binary variable, `mpg01`, that is equal to 1 if the value of `mpg` for that car is above 25, 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 123, 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?