Problem Set 3
CSCI 688

Problem 1 Assuming that we have training data $\{(x_n, y_n)\}_{n=1}^N$ where $x_n \in \mathbb{R}$ (i.e. $P = 1$) and y_n comes from one of two classes -1 or 1 encoded so that

$$y_n = \begin{cases} -1 & \text{if in class -1} \\ 1 & \text{if in class 1.} \end{cases}$$

Assume our training data has equal number of each class.

(a) Consider training an LDA model on this data. Given a newly observed x, show that LDA predicts class 1 if and only if

$$\hat{\alpha_0} + \hat{\alpha} x > 0$$

where

$$\hat{\alpha}_0 = -\frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_{-1})(\hat{\mu}_1 - \hat{\mu}_{-1})$$

and

$$\hat{\alpha} = \hat{\mu}_1 - \hat{\mu}_{-1}.$$

(b) Fit a linear regression model of the $y_n s$ onto the $x_n s$ so that

$$y_n = \hat{\beta}_0 + \hat{\beta} x_n.$$

Recall that for this type of simple linear regression

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}\bar{x}$$

and

$$\hat{\beta} = \frac{\sum_{n=1}^N x_n(y_n - \bar{y}}{\sum_{n=1}^N (x_n - \bar{x})^2}.$$

We can use this to build a classifier to classify points to class 1 if and only if

$$\hat{\beta}_0 + \hat{\beta} x > 0.$$

Show that this is equivalent to the LDA classifier. Hint: What is $\bar{y}$? What is $\bar{x}$ and $\sum_n x_n y_n$ in terms of $\hat{\mu}_1$ and $\hat{\mu}_{-1}$?

Problem 2 Consider the one-dimensional training data below.

x	-3	-2	0	1	-1	2	3	4	5
y	-1	-1	-1	-1	1	1	1	1	1

(a) What are the parameters needed to spcify LDA and QDA? Write code to calculate their estimated values from this training data. Hint: the pooled variance is just the weighted average of the variance of the two groups:

$$\hat{\sigma}^2_{pooled} = \frac{(N_1 - 1)\hat{\sigma}_1^2 + (N_{-1} - 1)\hat{\sigma}_{-1}^2}{N - 2}.$$

where N_1 and N_{-1} are the number in each group, resp., and N is the total number of training points.

(b) What are the discriminant functions? Write code that takes in a one-dimensional x and returns $\delta_1(x)$ and $\delta_{-1}(x)$.

(c) Classify the training data using the discriminant functions for both LDA and QDA. What is the training misclassification error rate?

(d) Given test pairs

x	-1.5	-1	0	1	.5	1	2.5	5
y	-1	-1	-1	-1	1	1	1	1

classify this test data using both LDA and QDA. What is the test error rate?

(e) Which is more suitable LDA or QDA?

Problem 3 In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the `Auto` data set in the `ISLR` package.

(a) Create a binary variable, mpg01, that contains a 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median. You can compute the median using the `median()` function. Note you may find it helpful to use the data.frame() function to create a single data set containing both mpg01 and the other Auto variables.

(b) Explore the data graphically in order to investigate the association between mpg01 and the other features. Which of the other features seem most likely to be useful in predicting mpg01? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

(c) Split the data into a training set and a test set.

(d) Perform LDA on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?

(e) Perform QDA on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?

(f) Perform logistic regression on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?

(g) Perform KNN on the training data, with several values of K, in order to predict mpg01. Use only the variables that seemed most associated with mpg01 in (b). What test errors do you obtain? Which value of K seems to perform the best on this data set?