DATA 558
SPRING QUARTER 2022

Homework # 2
Due Via Online Submission to Canvas: Wed, April 27 at 5pm

Instructions: You may discuss the homework problems in small groups, but you must write up the final solutions and code yourself. Please turn in your code for the problems that involve coding. However, code without written answers will receive no credit. To receive credit, you must explain your answers and show your work. All plots should be appropriately labeled and legible, with axis labels, legends, etc., as needed.

1. In this problem, we will make use of the `Auto` data set, which is part of the `ISLR2` package.

   (a) Fit a least squares linear model to the data, in order to predict `mpg` using all of the other predictors except for `name`. Present your results in the form of a table. Be sure to indicate clearly how any qualitative variables should be interpreted.

   (b) What is the (training set) mean squared error of this model?

   (c) What gas mileage do you predict for a Japanese car with three cylinders, displacement 100, horsepower of 85, weight of 3000, acceleration of 20, built in the year 1980?

   (d) On average, holding all other covariates fixed, what is the difference between the `mpg` of a Japanese car and the `mpg` of an American car?

   (e) On average, holding all other covariates fixed, what is the change in `mpg` associated with a 10-unit change in horsepower?

2. Consider using only the `origin` variable to predict `mpg` on the `Auto` data set. In this problem, we will explore the coding of this qualitative variable.

   (a) First, code the `origin` variable using two dummy (indicator) variables, with `Japanese` as the default value. Write out an equation like (3.30) in the textbook, and report the coefficient estimates. What is the predicted `mpg` for a Japanese car? for an American car? for a European car?

(b) Now, code the `origin` variable using two dummy (indicator) variables, with `American` as the default. Write out an equation like (3.30) in the textbook, and report the coefficient estimates. What is the predicted `mpg` for a Japanese car? for an American car? for a European car?

(c) Now, code the `origin` variable using two variables that take on values of $+1$ or -1. Write out an equation like (3.30) in the textbook, and report the coefficient estimates. What is the predicted `mpg` for a Japanese car? for an American car? for a European car?

(d) Finally, code the `origin` variable using a single variable that takes on values of 0 for Japanese, 1 for American, and 2 for European. Write out an equation like (3.30) in the textbook, and report the coefficient estimates. What is the predicted `mpg` for a Japanese car? for an American car? for a European car?

(e) Comment on your results in (a)-(d).

3. Fit a model to predict `mpg` on the `Auto` dataset using `origin` and `horsepower`, as well as an interaction between `origin` and `horsepower`. Present your results, and write out an equation like (3.35) in the textbook. On average, how much does the `mpg` of a Japanese car change with a one-unit increase in horsepower? How about the `mpg` of an American car? a European car?

4. Consider using least squares linear regression to predict weight (Y) using height.

   (a) Suppose that you measure height in inches (X_1), fit the model

   $$Y = \beta_0 + \beta_1 X_1 + \epsilon,$$

   and obtain the coefficient estimates $\hat{\beta}_0 = -165.1$ and $\hat{\beta}_1 = 4.8$. What weight will you predict for an individual who is 64 inches tall?

   (b) Now suppose that you want to measure height in feet (X_2) instead of inches. (There are 12 inches to a foot.) You fit the model

   $$Y = \beta_0^* + \beta_1^* X_2 + \epsilon.$$

   What are the coefficient estimates? What weight will you predict for an individual who is 64 inches tall (i.e. 5.333 feet)?

   (c) Now suppose you fit the model

   $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon,$$

   which contains both height in inches and height in feet as predictors. Provide a general expression for the least squares coefficient estimates for this model.

   (d) How do the (training set) mean squared errors compare for three models fit in (a)–(c)?

5. Suppose we wish to perform classification of a binary response in a setting with $p = 1$: that is, $X \in \mathbb{R}$, and $Y \in \{1, 2\}$. We assume that the observations in Class 1 are drawn from a $N(\mu, \sigma^2)$ distribution, and that the observations in Class 2 are drawn from an Uniform$[-2, 2]$ distribution.

   (a) Derive an expression for the Bayes decision boundary: that is, for the set of x such that $P(Y = 1 \mid X = x) = P(Y = 2 \mid X = x)$. Write it out as simply as you can.

   (b) Suppose (for this sub-problem only) that $\mu = 0$, $\sigma = 1$, $\pi_1 = 0.45$ (here, π_1 is the prior probability that an observation belongs to Class 1). Describe the Bayes classifier in this case: what range of x values will get assigned to Class 1, and what range of x values will get assigned to Class 2? Write out your answer as simply as you can. Draw a picture showing the set of x values assigned to Class 1 and the set of x values assigned to Class 2.

   (c) Now suppose we observe n training observations $(x_1, y_1), \ldots, (x_n, y_n)$. Explain how you could use these observations to *estimate* μ, σ, and π_1 (instead of using the values that were given in part (b)).

   (d) Given a test observation $X = x_0$, provide an estimate of

$$P(Y = 1 \mid X = x_0).$$

   Your answer should involve *only* the training observations $(x_1, y_1), \ldots, (x_n, y_n)$ and the test observation x_0, and *no* unknown parameters.

6. This problem has to do with logistic regression.

   (a) Suppose you fit a logistic regression to some data and find that for a given observation $x = (x_1, \ldots, x_p)^T$, the estimated log-odds equals 0.7. What is $P(Y = 1 \mid X = x)$?

   (b) In the same setting as (a), suppose you are now interested in the observation $x^* = (x_1 + 1, x_2 - 1, x_3 + 2, x_4, \ldots, x_p)^T$. In other words, $x_1^* = x_1 + 1$, $x_2^* = x_2 - 1$, $x_3^* = x_3 + 2$, and $x_j^* = x_j$ for $j = 4, \ldots, p$. Write out a simple expression for $P(Y = 1 \mid X = x^*)$. Your answer will involve the estimated coefficients in the logistic regression model, as well as the number 0.7.

7. In this problem, you will generate data with $p = 2$ features and a qualitative response with $K = 3$ classes, and $n = 50$ observations per class. You will then apply linear discriminant analysis to the data.

   (a) Generate data such that the distribution of an observation in the kth class follows a $N(\mu_k, \Sigma)$ distribution, for $k = 1, \ldots, K$. That is, the data follow a bivariate normal distribution with a mean vector μ_k that is specific to the kth class, and a covariance matrix Σ that is shared across the K classes. Choose Σ and $\mu_1, \ldots, \mu_K$ such that there is some overlap between the K classes, i.e. no linear decision boundary is able to perfectly separate the training data. Specify your choices for Σ and $\mu_1, \ldots, \mu_K$.

(b) Plot the data, with the observations in each class displayed in a different color. Compute and display the Bayes decision boundary (or Bayes decision boundaries) on this plot. This plot should look something like the right-hand panel of Figure 4.6 in the textbook (although no need to worry about shading the background, and also you don't need to display the LDA decision boundary for this sub-problem — you will do that in the next sub-problem). Be sure to label which region(s) of the plot correspond to each class.

(c) Fit a linear discriminant analysis model to the data, and make a plot that displays the observations as well as the decision boundary (or boundaries) corresponding to this fitted model. How does the LDA decision boundary (which can be viewed as an estimate of the Bayes decision boundary) compare to the Bayes decision boundary that you computed and plotted in (b)?

(d) Report the $K \times K$ confusion matrix for the LDA model on the training data. The rows of this confusion matrix represent the predicted class labels, and the columns represent the true class labels. (See Table 4.4 in the textbook for an example of a confusion matrix.) Also, report the training error (i.e. the proportion of training observations that are misclassified).

(e) Generate $n = 50$ test observations in each of the K classes, using the bivariate normal distributions from (a). Report the $K \times K$ confusion matrix, as well as the test error, that results from applying the model fit to the training data in (c) to your test data.

(f) Compare your results from (d) and (e), and comment on your findings.

8. In this problem, you will apply quadratic discriminant analysis to the data from Q7.

(a) Fit a quadratic discriminant analysis model to the training data from Q7, and make a plot that displays the observations as well as the QDA decision boundary (or boundaries) corresponding to this fitted model. Be sure to label which region(s) of the plot correspond to each class. How does the QDA decision boundary compare to the Bayes decision boundary that you computed in Q7(b)?

(b) Report the $K \times K$ confusion matrix for the QDA model on the training data, as well as the training error.

(c) Repeat (b), but this time using the test data generated in Q7. (That is, apply the model fit to the training data in (a) in order to calculate the test error.)

(d) Compare your results in (b) and (c), and comment on your findings.

(e) Which method had smaller *training error* in this example: LDA or QDA? Comment on your findings.

(f) Which method had smaller *test error* in this example: LDA or QDA? Comment on your findings.

9. **EXTRA CREDIT.** We have seen in class that the least squares regression estimator involves finding the coefficients $\beta_0, \beta_1, \ldots, \beta_p$ that minimize the quantity

$$\sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})\right)^2 .$$

By contrast, the *ridge regression estimator* (which we will discuss in Chapter 6) involves finding the coefficients that minimize

$$\sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})\right)^2 + \lambda(\beta_1^2 + \ldots + \beta_p^2)$$

for some positive constant λ. For simplicity, assume that $\beta_0 = 0$. Derive an expression for the ridge regression estimator, i.e. for the coefficient estimates $\hat{\beta}_1, \ldots, \hat{\beta}_p$.