CS 5565, LAB4(Cross Validation and Bootstrap) 120 pts.

Name _____

1. View the videos at the following URLs
   https://www.youtube.com/watch?v=6dSXlqHAoMk
   https://www.youtube.com/watch?v=YVSmsWoBKnA
   You may download the R Code for Labs and the Data Sets to use from the textbook website.
   http://www-bcf.usc.edu/~gareth/ISL/

2. (40 points total) In Chapter 4, we used logistic regression to predict the probability of `default` using `income` and `balance` on the `Default` data set. We will now estimate the test error of this logistic regression model using the validation set approach. Do not forget to set a random seed before beginning your analysis.

   (a) (10 points) Fit a logistic regression model that uses `income` and `balance` to predict `default`.

   (b) (10 points total) Using the validation set approach, estimate the test error of this model. In order to do this, you must perform the following steps:

      i. (2.5 points) Split the sample set into a training set and a validation set.
      ii. (2.5 points) Fit a multiple logistic regression model using only the training observations.
      iii. (2.5 points) Obtain a prediction of default status for each individual in the validation set by computing the posterior probability of default for that individual, and classifying the individual to the `default` category if the posterior probability is greater than 0.5.
      iv. (2.5 points) Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.

   (c) (10 points) Repeat the process in (b) three times, using three different splits of the observations into a training set and a validation set. Comment on the results obtained.

   (d) (10 points) Now consider a logistic regression model that predicts the probability of `default` using `income`, `balance`, and a dummy variable for `student`. Estimate the test error for this model using the validation set approach. Comment on whether or not including a dummy variable for `student` leads to a reduction in the test error rate.

3. (40 points) We continue to consider the use of a logistic regression model to predict the probability of `default` using `income` and `balance` on the `Default` data set. In particular, we will now compute estimates for the standard errors of the income and balance logistic regression coefficients in two different ways: (1) using the bootstrap, and (2) using the standard formula for computing the standard errors in the `glm()` function. Do not forget to set a random seed before beginning your analysis.

   (a) (10 points) Using the `summary()` and `glm()` functions, determine the estimated standard errors for the coefficients associated with `income` and `balance` in a multiple logistic regression model that uses both predictors.

(b) (10 points)Write a function, `boot.fn()`, that takes as input the `Default` data set as well as an index of the observations, and that outputs the coefficient estimates for `income` and `balance` in the multiple logistic regression model.

(c) (10 points) Use the `boot()` function together with your `boot.fn()` function to estimate the standard errors of the logistic regression coefficients for `income` and `balance`.

(d) (10 points) Comment on the estimated standard errors obtained using the `glm()` function and using your bootstrap function.

4. (40 points) We will now consider the `Boston` housing data set, from the `MASS` library.

(a) (5 points) Based on this data set, provide an estimate for the population mean of `medv`. Call this estimate $\hat{\mu}$.

(b) Provide an estimate of the standard error of $\hat{\mu}$. Interpret this result.
Hint: We can compute the standard error of the sample mean by dividing the sample standard deviation by the square root of the number of observations.

(c) (5 points) Now estimate the standard error of $\hat{\mu}$ using the bootstrap. How does this compare to your answer from (b)?

(d) (5 points) Based on your bootstrap estimate from (c), provide a 95% confidence interval for the mean of `medv`. Compare it to the results obtained using `t.test(Boston$medv)`. Hint: You can approximate a 95% confidence interval using the formula $[\hat{\mu} - 2SE(\hat{\mu}), \hat{\mu} + 2SE(\hat{\mu})]$.

(e) (5 points) Based on this data set, provide an estimate, $\hat{\mu}_{med}$, for the median value of `medv` in the population.

(f) (5 points) We now would like to estimate the standard error of $\hat{\mu}_{med}$ Unfortunately, there is no simple formula for computing the standard error of the median. Instead, estimate the standard error of the median using the bootstrap. Comment on your findings.

(g) Based on this data set, provide an estimate for the tenth percentile of `medv` in Boston suburbs. Call this quantity $\hat{\mu}_{0.1}$ (You can use the `quantile()` function.)

(h) (5 points) Use the bootstrap to estimate the standard error of $\hat{\mu}_{med}$. Comment on your findings.