

Homework Assignment 1 (The chapter and exercise number are based on 2nd edition of the textbook)

(Please submit a softcopy of your assignment by Oct 15. TA will create a folder at Canvas for your submission.)

Chapter 2, Exercise 2 (page 52)

2. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p .
- (a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.
 - (b) We are considering launching a new product and wish to know whether it will be a *success* or a *failure*. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.
 - (c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

Chapter 3, Exercise 9 (page 123-124)

9. This question involves the use of multiple linear regression on the `Auto` data set.

- (a) Produce a scatterplot matrix which includes all of the variables in the data set.
- (b) Compute the matrix of correlations between the variables using the function `cor()`. You will need to exclude the `name` variable, which is qualitative.
- (c) Use the `lm()` function to perform a multiple linear regression with `mpg` as the response and all other variables except `name` as the predictors. Use the `summary()` function to print the results. Comment on the output. For instance:
 - i. Is there a relationship between the predictors and the response?
 - ii. Which predictors appear to have a statistically significant relationship to the response?
 - iii. What does the coefficient for the `year` variable suggest?
- (d) Use the `plot()` function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?
- (e) Use the `*` and `:` symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

(for this exercise, please find appropriate packages and functions in python to replace the R functions mentioned in the exercise)

13. This question should be answered using the `Weekly` data set, which is part of the `ISLR2` package. This data is similar in nature to the `Smarket` data from this chapter's lab, except that it contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.
- (a) Produce some numerical and graphical summaries of the `Weekly` data. Do there appear to be any patterns?
 - (b) Use the full data set to perform a logistic regression with `Direction` as the response and the five lag variables plus `Volume` as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?
 - (c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.
 - (d) Now fit the logistic regression model using a training data period from 1990 to 2008, with `Lag2` as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).
 - (e) Repeat (d) using LDA.
 - (f) Repeat (d) using QDA.
 - (g) Repeat (d) using KNN with $K = 1$.
 - (h) Repeat (d) using naive Bayes.
 - (i) Which of these methods appears to provide the best results on this data?
 - (j) Experiment with different combinations of predictors, including possible transformations and interactions, for each of the methods. Report the variables, method, and associated confusion matrix that appears to provide the best results on the held out data. Note that you should also experiment with values for K in the KNN classifier.

5. 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) Fit a logistic regression model that uses `income` and `balance` to predict `default`.
- (b) Using the validation set approach, estimate the test error of this model. In order to do this, you must perform the following steps:
 - i. Split the sample set into a training set and a validation set.
 - ii. Fit a multiple logistic regression model using only the training observations.
 - iii. 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. Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.
- (c) 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) 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.