COM S 573: Home work 2 Spring 2016

Write your name on each page. Maximum score is 40 points, due date is **Wednesday, March 28, 2016** . Please hand in the solutions (CLEAN version) on the due date in class (**hard copy**). Also paste the results of your R code and the code itself into your homework. Make sure your homework is stapled!

1. In this exercise you'll create some simulated data and fit a simple linear regression model to it.
   (a) [1 point] Perform the following commands in R

   ```
   > set.seed (1)
   > x1 <- runif (100)
   > x2 <- 0.5* x1+rnorm (100) /10
   > Y <- 2+2* x1 +0.3* x2+rnorm (100)
   ```

   Write out the form of the linear model. What are the regression coefficients?

   (b) [1 point] What is the correlation between $x1$ and $x2$? Create a scatterplot displaying the relationship between the variables.

   (c) [2 points] Using this data, fit a least squares regression to predict Y using $x1$ and $x2$. Describe the results obtained. What are $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$? How do these relate to the true β_0, β_1 and β_2? Can you reject the null hypothesis $H_0 : \beta_1 = 0$? How about $H_0 : \beta_2 = 0$?

   (d) [1 point] Now fit a least squares regression to predict Y using only $x1$. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

   (e) [1 point] Now fit a least squares regression to predict Y using only $x2$. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

   (f) [2 points] Do the results obtained in (c)-(e) contradict each other? Explain your answer.

   (g) [3 points] Now suppose we obtain one additional observation, which was unfortunately mismeasured.

   ```
   > x1 <- c(x1 , 0.1)
   > x2 <- c(x2 , 0.8)
   > y <- c(y,6)
   ```

   Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers and make suitable plots.

2. [6 points] This problem relates to the QDA model, in which the observations within each class are drawn from a normal distribution with a classspecific mean vector and a class specific covariance

matrix. We consider the simple case where $p = 1$; i.e. there is only one feature. Suppose that we have K classes, and that if an observation belongs to the kth class then X comes from a one-dimensional normal distribution, $X \sim N(\mu_k, \sigma_k^2)$. Recall that the density function for the one-dimensional normal distribution is given in Eq. 4.11 in the textbook. Prove that in this case, the Bayes classifier is not linear. Argue that it is in fact quadratic.

3. [6 points] Suppose that we wish to predict whether a given stock will issue a dividend this year ("Yes" or "No") based on X, last years percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was $\overline{\mathrm{X}} = 10$, while the mean for those that didn't was $\overline{\mathrm{X}} = 0$. In addition, the variance of X for these two sets of companies was $\sigma^2 = 36$. Finally, 80% of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was $X = 4$ last year.

4. This question should be answered using the Weekly data set, which is part of the ISLR 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) [2 points] Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?

   (b) [2 points] 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) [2 points] 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) [2 points] 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) [2 points] Repeat (d) using LDA.

   (f) [2 points] Repeat (d) using QDA.

   (g) [1 point] Is it justified to use QDA? Use appropriate hypothesis test(s) we've seen in class.

   (h) [2 points] Repeat (d) using KNN with $K = 1$.

   (i) [1 point] Which of these methods appears to provide the best results on this data?

   (j) [1 point] Could you create a better classifier? How would you do this?