COM S 573: Home work 1 Spring 2014

Write your name on each page. Maximum score is 35 points, due date is **Friday, February 7, 2014** . Please upload a CLEAN version of the solutions (**one** PDF file) on BlackBoard (before 10am) or hand in the solutions (CLEAN version) on the due date in class (hard copy). If you turn in the home work in class make sure it is stapled!

1. The table below provides a training data set containing 6 observations, 3 variables (or predictors) and 1 qualitative response variable. Suppose we wish to use this data set to make a prediction for

Observation	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

   Y when $X_1 = X_2 = X_3 = 0$ using k-nearest neighbors.

   (a) [3 points] Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

   (b) [2 points] What's your prediction with $k = 1$? Explain.

   (c) [2 points] What's your prediction with $k = 3$? Explain.

   (d) [2 points] If the Bayes decision boundary in this problem is highly nonlinear, then we would expect the best value for k to be large or small? Explain.

2. (a) [5 points] Suppose we would like to fit a straight line through the origin i.e., $Y_i = \beta_1 x_i + e_i$ with $i = 1, \ldots, n$, $\mathbf{E}[e_i] = 0$, $\mathbf{Var}[e_i] = \sigma_e^2$ and $\mathbf{Cov}[e_i, e_j] = 0, \forall i \neq j$. Find the least squares estimator $\hat{\beta}_1$ for the slope β_1.

   (b) [5 points] Calculate the bias and the variance for the estimated slope $\hat{\beta}_1$.

3. [10 points] Solve Exercise 8 in Chapter 2 on page 54 of the textbook (*An Introduction to Statistical Learning with Applications in R*).

   Guidelines regarding Question 3

   1. This exercise has multiple parts. Please, answer each part separately and in a brief way. Be direct to the point!

2. Type each question before you answer it, and provide a clear separation between each part.
3. All relevant computer output should be provided unless noted otherwise.
4. Attach your R code as an Appendix. Make sure to provide comments on what your code is doing. Keep it clean and clear!

Hints

1. In part (a), when you read the data into R, make sure to check if the data has a header or not.
2. In part (b), you dont need to use the fix() function to view the loaded dataset. Instead, and since we are using Rstudio, you can view the data by clicking on the data name (college) in the Workspace window in Rstudio. I am saying this, because sometimes the fix() function might crash Rstudio especially if you are using Macs. Another option would be to use the command View(college).
3. Parts (a) and (b) are for data manipulation (i.e. there is no need to include any output in the report for submission). You will be mainly answering questions from part (c).
4. For part (c-iii), make sure to annotate the plot (title, x-axis, and y-axis).
5. If you want to learn more about a certain R function, you can use the command ?. For example, if you want to learn more about the plot() function, type the command ?plot, and a help document will pop up.
6. In part (c-v), when you use the command par(mforw(2,2)), the plotting screen should split into 2x2 = 4 parts. To go back to the original setting, run the command par(mforw(1,1)).

4. Consider the following equation of a straight line $Y_i = \beta_0 + \beta_1 x_i + e_i$ with $i = 1, \ldots, n$, $\mathbf{E}[e_i] = 0$, $\mathbf{Var}[e_i] = \sigma_e^2$ and $\mathbf{Cov}[e_i, e_j] = 0, \forall i \neq j$.
   (a) [2 points] Calculate the bias for the estimator of the intercept $\hat{\beta}_0$.
   (b) [4 points] Calculate the variance for the estimator of the intercept $\hat{\beta}_0$.