STT3851 Homework 2

Dr. Lasanthi Watagoda

Due – Sep 5

Note: *Submit a printed pdf or an html file. ggplots must be used for all graphics. Graphics created using base commands will not receive credits.*

1) 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.

2) Describe three real-life applications in which *classification* might be useful.

   a) Describe the response, as well as the predictors.

   b) Is the goal of each application inference or prediction? Explain your answer.

3) Describe three real-life applications in which *regression* might be useful.

   a) Describe the response, as well as the predictors.

   b) Is the goal of each application inference or prediction? Explain your answer.

4) Describe the differences between a parametric and a non-parametric statistical learning approach.

   a) What are the advantages of a parametric approach to regression or classification (as opposed to a nonparametric approach)?

   b) What are its disadvantages?

5) Extra credit: Start with $E(Y - \hat{Y})^2 = E[f(X) + \epsilon - \hat{f}(X)]^2$ and show that the $E(Y - \hat{Y})^2 = [f(X) - \hat{f}(X)]^2 + Var(\epsilon)$