Project 1: Data Analysis Competition Statistics 8330 (DAIII; Wikle) November 3, 2015

DUE: November 12 and November 17, 2015

Instructions: You may work in groups of 1-3 people on this project and you only need turn in one project paper per group. The project consists of two data analyses, a regression problem and a classification problem as outlined below. In both cases, you are to find the best model to answer the question at hand using any method we have discussed in Data Anlaysis III up to and including the lecture on November 3, 2015. You may also use methods that were covered in Data Analysis I and II if you like. You must use R to perform your analyses. I will provide you with training data sets on the class blackboard website. I will hold back a separate set of test data to be used as specified below.

Data Sets:

1. **Online News Popularity:** "This dataset summarizes a heterogeneous set of features about articles published by Mashable in a period of two years. The goal is to predict the number of shares in social networks (popularity)." You can read more about this at http://archive.ics.uci.edu/ml/datasets/Online+News+Popularity but note that I did not give you all of the data (and, I even after giving you the test data, you will not have received all of the data). The training data is given in `News_train.txt`. This consists of 30,000 records (rows) with 58 columns (columns 1 - 57 are features as described in the `names.txt` file and the last column is the response variable (number of shares). Your task is to predict the number of shares given the features.

2. **Hill-Valley Graph Prediction:** The file `Hill-Valley_train.txt` contains 700 records with thefirst 100 columns corresponding to the "y" values on a cartesian graph (the x-axis corresponds to 1,2,...,100). See the example plot in

   `Hill_Valley_visual_examples.jpg`.

   Your task is to classify each of these graphs as either a "hill" (upward bump) or "valley" (downward bump). The last column in the dataset contains the classifcation. These data are described http://archive.ics.uci.edu/ml/datasets/Hill-Valley, but note that I mixed the test and training data together from that site and didn't give you all of it. Please note that the overall mean values on these graphs do not matter for the classification (and, they are quite varied). It is just the overall pattern that matters!

Tasks:

- Each group will analyze both data sets to come up with their best solution based on the training data. How you do this is up to you (subject to the constraints given in the "Instructions" above).

- Each group will hand in to me a piece of paper with your group members and team name on Thursday, November 5.

- Each group will turn in one project report on November 12 (in class!!) that contains the following:

  - A **very brief** introduction to the problems (no more than 1 paragraph each).

- Any basic data analysis/description/plots that you feel are important; these must be **relevant** to your data analysis or don't include them!
- A **brief** description of how you decided on your final model and any pre-processing you do to the data.
- A **brief** description of your final model and your results in terms of misclassification error rate, false positive rate, AUC, for the classification data and the MSE error rate for the regression problem.
- The **exact** R commands necessary to obtain the exact misclassification error rate and false positive rate for your classification model, and the MSE for your regression problem. Include any preprocessing commands (I need to be able to duplicate your results exactly!) Put this in an appendix and **email just the R code to me, along with your group member names**!.
- A **very brief** conclusion.

- After class on November 12, I will make available the test data sets. You are then to run your **best model as indicated on the training data** on this test data set.

- You will make a 1 page report showing your results and hand this in on November 17 (in class) that gives the test misclassification error rate and false positive rate for the classification problem and the test MSE for the regression problem. In addition, you will include the **exact** R commands that gave you this output. **YOU MUST USE THE MODEL YOU SELECTED IN THE TRAINING PHASE!! FAILURE TO DO SO WILL DISQUALIFY YOU!**

- As this is a competition, the group with the best classification results and best regression results **wins**. HUGELY MASSIVE GINORMOUS PRIZES!!