Midterm Two - Statistics 153, Spring 2017

Due on April 18, 2017

April 4, 2017

On piazza, you will find five time series datasets: `q1_train.csv`,`q2_train.csv`,`q3_train.csv`,`q4_train.csv` and `q5_train.csv`. Each of these datasets is of length 525 and gives the google trends-like data (downloaded on April 04, 2017) for queries from the first week of January, 2004 to the second week of January, 2014. Your task is to predict the next 104 observations (2 years) of these time series.

You are encouraged but not required to work on all the five datasets. You are required to work on at least two of the datasets.

One member of your team must turn in the following. Please make sure the same team member turns in all parts:

1. Your predictions for the datasets that you have worked on. **These are due by 11:59PM on April 18, 2017**. You are required to turn in a txt-file for each data set. The text file should contain your predictions for the following 104 time points separated by "," or "newlines" and it should be named `Q[Number]_[Firstname]_[Lastname]_[SID].txt`. For example, `Q1_Alex_DAmour_123456.txt`. We assume that you will submit your values in an increasing order:

   $$\hat{X}_{526}, \hat{X}_{527}, \ldots, \hat{X}_{629}$$

   Please be aware that your submission must be of the right form in order to be valid.

   We are still working out the platform on which you will submit these files – we will update you with details.

2. A report describing your analysis. For one of the datasets that you have worked on, write a clean report describing your analysis attaching the relevant plots and R output. Include your R code as an Appendix to the report. Do not write a report for each of the datasets that you worked on. Just write it for one of those datasets and include your R work for the other datasets as an Appendix. The length of the report including the relevant plots and R output (and excluding the R code) cannot exceed 8 pages. **Make sure to include both team members' names on the report.**

You will be graded on the prediction accuracy as well as your report (**the report will be for 25 points and the prediction accuracy will be graded to a maximum of 10 points**). Here is a description of how your prediction accuracies will be evaluated. Suppose you decide to submit predictions for the dataset **q1**. Let your predictions be denoted by $\hat{X}_{526}, \ldots, \hat{X}_{629}$ and let the true values of **q1** (which we will have access to) are $X_{526}, \ldots, X_{629}$. We will first compute the sum of squares

$$\sum_{i=526}^{629} \left(\hat{X}_i - X_i\right)^2$$

This result will measure your discrepancy for **q1**. From here, we will compute:

$$5 * \frac{\text{best discrepancy for } \mathbf{q1} \text{ in the class}}{\text{your discrepancy for } \mathbf{q1}}$$

This will be your score for **q1**. Note that the maximum possible value for this score is 5. The minimum possible score is 0 (this will be the case if the best in-class discrepancy is zero). We will similarly compute your score for each of the datasets that you submit predictions for. To get your final points for the prediction part, we will take the sum of your highest two scores. (For example, if you submit predictions for four datasets and your scores are 4.1, 4.8, 3.7, 4.9; then you will get 4.8 + 4.9 = 9.7 points out of 10 for the prediction part).

You may work with a partner, but your group **may not** collaborate with other groups. You are allowed to use code from the lectures and the section without explicit citation. You are also allowed to consult books or online resources for your analysis but you must credit all such sources in your report. Anyone caught cheating (which includes copying code, reports etc.) risks failing the class and being referred to the Office of Student Conduct.