STAT 542/CS 598 Fall 2021

# Coding Assignment 3

**Due Monday, October 11**

## Part I: Select optimal span for `loess` (2pt)

Write your own functions to use LOO-CV and GCV to select the optimal span for `loess`.

- Test your code on data set `Coding3_Data.csv`, which can be downloaded from Campuswire.
- Report your CV and GCV for 15 span values: $0.20, 0.25, \ldots, 0.90$.
- Report the optimal value(s) for span based on CV and GCV.
- Plot the fitted curve(s) using the optimal value(s) for span.

## Part II: Clustering time series (3pt)

Download the `Sales_Transactions_Dataset_Weekly` dataset from UCI Machine Learning Repository Link.

This dataset contains weekly purchased quantities of 811 products over 52 weeks, i.e., each product has a time series with 52 measurements. For this assignment, we want to cluster time series with similar fluctuation patterns even if their means are different. So remove the mean from each time series and store the data as an 811-by-52 matrix $\mathbf{X}$.

1. Fit each time series with a NCS with df = 10, which is equivalent to a NCS with 8 interior knots. That is, treat each row of $\mathbf{X}$ as the response and the one-dimensional feature is just the index from 1 to 52. Save the corresponding coefficients (without the intercept) as an 811-by-9 matrix $\mathbf{B}$.

   The matrix $\mathbf{B}$ can be obtained as follows. Let $\mathbf{F}$ denote the 52-by-9 design matrix without the intercept, which, for example, can be obtained by calling the `ns` command in `R`. Remove the column mean from $\mathbf{F}$ as we do not care about the intercept. Then $\mathbf{B}^t = (\mathbf{F}^t\mathbf{F})^t\mathbf{X}^t$.

2. Run k-means algorithm on $\mathbf{B}$ to cluster the 811 products into 6 clusters. Display time series for products in the same cluster in one figure along with the corresponding cluster center; arrange the 6 figures in 2-by-3 format.

3. Run k-means algorithm on $\mathbf{X}$ to cluster the 811 products into 6 clusters. Display time series for products in the same cluster in one figure along with the corresponding cluster center; arrange the 6 figures in 2-by-3 format.

**What you need to submit?**

- A Markdown (or Notebook) file in HTML format, which should contain all necessary code and the corresponding output/results.

- For Part II, set the seed at the beginning of your code to be the last 4-dig of your University ID. So once we run your code, we can get the same result.

  You do not need to set seed for Part I as nothing is random.

- Name your file starting with `Assignment_2_xxxx_netID..`, where "xxxx" is the last 4-dig of your University ID and make sure the same 4-dig is used as the seed in your code.