CEE 690: Health and Environmental Data Science, Spring 2019
Homework 4: 70 Total Points
Due: April 3rd by 11:59 PM
Submissions required electronically through Sakai

Problem 1: Assessing Performance on Blind Predictions From HW3 (10 points)
In Homework 3, we asked you to make predictions on a blind data set. Here, we want to assess how well you did on that problem.

For HW 3, problem 5, please assess the AUC between the true outcomes in "hw3_problem_5_y_blind.csv" versus your predictions.

(a) What was your hold-out AUC?
(b) A "good" AUC here is about .79 using **4 components**. How did your results compare to this? Were you able to reduce the number of components by incorporating the PCA into the validation procedure?
(c) Discovering that 4 components was ideal would require a much larger parameter search space than you may of used. How does the larger search space affect multiple comparisons?

Problem 2: Permutation Testing (25 points)

We want to examine the impact of dataset size on permutation testing. To evaluate this, try running a cross-validation grid search procedure (you can choose the procedure) on permutated labels with *N* datapoints, where *N* is [20,35,50,100,200,400]. Use 10 repeats for each setting.

(a) Plot the mean and standard deviation of the results as a function of the number of data.
(b) Looking at the trends, what can we say about the bias in the predictive performance results and the effect of more data?
(c) How many distinct models did your cross-validation procedure try? How would increasing or decreasing the search space affect the results?

Problem 3: Forcasting Time Series (30 points)

The famous CO2 dataset from the Mauna Loa observatory is in the class repository. In the homework template, I have placed code to preprocess the dataset to address missing data. Then, there are two versions of the data: the first is the raw values, and the second is "detrended" data, where a linear trend has been removed To start building an understanding of time series, please use the time series validation methods and evaluate the following:

(a) A linear regression model to forecast one step ahead on the raw data. Determine the optimal number of lags.
(b) A k-Nearest Neighbor model to forecast one step ahead on the raw data. Determine the optimal number of neighbors.

(c) On the raw data, which method worked better? Why?
(d) A linear regression model to forecast one step ahead on the detrended data. Determine the optimal number of lags.
(e) A linear regression model to forecast one step ahead on the detrended data. Determine the optimal number of lags.
(f) A k-Nearest Neighbor model to forecast one step ahead on the detrended data. Determine the optimal number of neighbors.
(g) On the detrended data, which method worked better? Why? Was performance much closer than before?
(h) The "detrended" data simply removes a linear fit line. Did a linear fit line remove the background "trend?"

Problem 6: (5 points)
How many hours did it take you to complete this assignment?
Affirm that you adhered to the Duke Honor Code.