Intro to Data Science, Fall 2019
Homework 3

Due Nov 17 at 11:59 PM via Canvas

---

1. Consider the data set given in `https://archive.ics.uci.edu/ml/datasets/Air+quality`. Read the description of the data on the webpage. As the website says, the dataset contains 9358 instances of hourly averaged responses from an array of chemical sensors located in a significantly polluted area in Italy. Data were recorded from March 2004 to February 2005 (one year). For every time point t, your job is to write code to predict the values at time $t + k$, for some fixed value of k. I recommend you try out values of k equal to 1 hour, 12 hours, 1 day, 2 days, and 7 days. You can use the ARIMA tool in Python or R.

   For each prediction, you will need to compute the following measures to compute the accuracy:

   (a) Mean Absolute Percentage Error (MAPE)

   (b) Mean Error (ME)

   (c) Mean Absolute Error (MAE)

   (d) Mean Percentage Error (MPE)

   (e) Root Mean Squared Error (RMSE)

   (f) Lag 1 Autocorrelation of Error (ACF1)

   (g) Correlation between the Actual and the Forecast (corr)

   (h) Min-Max Error (minmax)

   Make sure your results are visualized. Submit your Python notebook ot R code along with the visualizaed results. Deliverable are:

   (a) Outline of the method

   (b) Notebook files with executable source code with the instructions in the comments section. Make sure you do not submit SCREENSHOTS OF CODE OR CODE IN PDF OR DOC FILES

   (c) Line plots of real values and predicted values (in 2 different colors)

   (d) Table or plots of accuracy measures