Problem Set 2

Data Science 602: Data Analysis and Machine Learning

Spring 2022

1. **Data transformation** Using the `weather` dataset, construct a derived data frame having the following qualities:

   (a) Each row represents an observation day

   (b) Each column represents an hourly temperature observation. That is, the dataframe includes 24 columns of the form tmpmeasx where x ranges from 0 to 23 and represents the observation taken that hour. (For changes from daylight savings time to standard time, there are two observations at the 1:00 hour (local). You may discard one of the values arbitrarily.)

   Using this derived dataset, show the average difference in temperature, standard deviation, and max/min values between observations taken in hours 2 (about 2:56am) and 14 (about 14:56, or 2:56pm).

2. **Circular data and One-Hot Encoding** Open and clean the `citations` data frame using the script provided, and discard records where the violation timestamp is not between January 1, 2020 and July 1, 2021. Use this derived data frame to complete the following:

   (a) Identify the 5 most common violation types. Remove from the derivative data frame any violation types not in these 5 most common

   (b) Identify the average time of day for each violation in the derivative data frame. (You must use the circular mean to compute this average.)

   (c) The violation type column is a nominal field. Convert this field to one-hot encoded (OHE) variables, and add these variables to the data frame.

3. **Data analysis** Join the citations and weather datasets to produce a merged data frame. For each observation in the citations dataset, the merged data frame should provide the reported weather conditions for the closest observation in the weather dataset.

   From this joined data set, examine instances of fixed speed camera citations (violation type 32). Does weather affect citation volumes? Justify your conclusion.