12-741: Data Management
Assignment#2

Instructor: Mario Bergés
TA: Rami Ariss
Due on: Friday November 5, 11:59pm ET

October 29, 2021

Some notes before you begin:

When answering the following questions, please provide all of your calculations to arrive at the answer (in addition to the answer itself). Your calculations should be very clear and easy to understand. They should include your assumptions, and a step-by-step explanation of how you arrived at the solution. Also, make sure you type your name and AndrewID on the top of each page.

Some final recommendations:

- Before finding the answer to each question or looking at the next step in the solution, take some time to think about how you can come up with this on your own.
- Again, make sure you document everything you do, and not just write down the answer to the question. This will both help during grading as well as improving your learning process.
- Do not write down any solution or process that you do not understand. If you feel that you do not understand how to do something, seek some help. The preferred method for this is to post your questions on the discussion board for the course, in Canvas and/or Piazza.

1 Warming up to Time-Series Data Again (15%)

After having so much rain a couple of weeks ago, I thought it would be appropriate to analyze some rainfall time-series data.
Visit the 3 Rivers Wet Weather website to learn about historical rainfall data that you can obtain (either programmatically, or via their web interface) for the city of Pittsburgh. Pick a specific rainfall gauge that you are intersted in, and download data for the last 2 years on an hourly interval.

1. Plot the full time-series, ensuring that the axes of your plot have meaningful labels and units.

2. Using the techniques described in Lectures 2 and 3, describe how you would proces that dataset (i.e., find and remove outliers, model trends in the data, etc.) in a manner that you find fitting. *(There is no single correct answer here but, clearly, there are wrong answers)*

2 Water in the foundation: yikes! (60%)

The File "Water_Temp_Strain.txt" reports the recordings or 3 sensors, during a one-year campaign, with a sampling period Δt of 144 minutes (corresponding to 10 measurements per day). The file has 3651 rows. Column 1 reports the water level in the soil under a building (in meters), Column 2 the temperature (in Celsius), and Column 3 the strain on a column of the building (in micro-strains).

You can process the measurements by using the following model:

$$y_i = q(t_i) + \gamma T_i + \delta W_i + n_i$$

Here $q(t_i) = \alpha t_i + \beta$, y indicates the strain measurementments, T the temperature measurements, W the water level, q the actual strain component not depending on W or T; t is the time, n the noise, and subscript i refers to the i-th recording, at time t_i. Finally, α, β, γ and δ are the model parameters.

(a) Using linear regression, find the best fitting parameter vector $\hat{w} = [\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\delta}]^T$, specifying the corresponding physical units.

(b) Using the model identified in (a), infer q_{3000}, i.e. the strain at t_{3000}

(c) Again, using the model identified in (a), predic q_{5y} (the actual strain component, not depending on water level or temperature, 4 years after the end of the monitoring campaign recorded in the dataset.

(d) The strain measurements may contain some outliers. Plot the residuals obtained from the model fitted in part (a), and report their mean value and standard deviation.

(e) By using "Chauvenet's criterion", mark as outliers and remove from dataset all points for which the residual is more than 3 standard deviations away from the mean. Repeat the regression analysis and the outlier removal until convergence. During these iterations, report the value of the removed outliers and the position in the original dataset (i.e. the row number). Also report the value of $\hat{w}$ after having removed all outliers.

(f) Using the model identified in (e), re-compute q_{3000} and q_{5y}.

(g) Suppose now that you do not have access to the recording of the temperature data. Re-calibrate your model (i.e., find $\hat{w}$) and re-estimate q_{3000} and q_{5y}. What do you find?

3 Wet Databases (15%)

Now that you have a model that can predict strain on a column given water level in the soil and temperature, you consider starting a company that will do predictive maintenance of structural elements by predicting possible failure modes during heavy rain events. Your company will install sensor kits in critical structural members, measuring just temperature and water level in the soil at the site, and leverage existing rainfall data from the city's sensor network (like the 3 Rivers Wet Weather one) to better understand rainfall and water content in the soil across the region and improve prediction. Of course, this is purely hypothetical, but it allows us to get to the following:

Create an entity-relationship diagram for the design of the company's main database system, which will contain information about: the sensor kits that are installed, the structural members that are being monitored, the rainfall gauges that the city has installed, the actual data collected by botht he rainfall gauges and the company's sensors, the clients that are being served, and the maintenance records performed on the sensor kits.

4 Set Theory (10%)

One tenth of the baseball players are Dominicans, and one sixth of Dominicans are baseball players. Which group is larger? What is the ratio of the sizes of these groups?