Recitation 1

Big Data Science 18788
Spring 2023

Friday March 17th, 2023

Note on recitation slides

- Recitation slides are intended to be a guide on how to approach the assignment and not a prescription of exactly what to do.
- There could be many approaches to any problem
- Seek to understand the problem and solve it instead of just trying to reproduce the steps listed in the slides
- To avoid overdependence on the slides, start your assignment early!
- If you have to choose between a “creative” approach, or following assignment instructions in the PDF, choose to follow PDF instructions **always**

Assignment Objectives

- Studying the relationship between rainfall index and vegetation index
- Feature Engineering
- Fitting parametric and non-parametric models
- Evaluating performance of models

Question 1

1. Download the three datasets:
    - RwandaDistrictVegetation.csv
    - RwandaDistrictRainfall.csv
    - RwandaDistrictCentroidsLongitude-Latitude.csv

2. Load the two datasets in your working environment
    - RwandaDistrictVegetation.csv
    - RwandaDistrictRainfall.csv

Question 2

- Plot two time series graphs, one for **rainfall index data** and the other for the **vegetation index data** . Each should have a **6x5 subplots** for all districts.
    - Preprocess the dataframe. Transpose or other technique.
    - Drop the NaNs values where necessary.
    - Label the subplots appropriately
        - X label and y label
    - Title / header
    - Comment on any observed pattern.

Question 3

- Compute the statistical quantities for both rainfall and vegetation index for the **12 months.**
  - Mean, median, minimum and maximum.
- For each ith month where i (ranges from 1 to 12) find the four statistical summaries.
- Provide 2 graphs (one for rainfall and another one for vegetation index) with lines representing the mean, median, minimum and maximum values. Add figure legends
- Analyze your result.

Question 4

- Load the ***RwandaDistrictCentroidsLongitude-Latitude.csv*** dataset
- Calculate the correlation coefficient,C, for rainfall between each pair of districts.
- Compute the distance, d, measured in km between the pair of districts.

  *Hint [use **haversine** to compute the distance. Available in Python and in MATLAB]*

- Make an initial graph to show the correlation values against distance

Question 4 (cont'd)

- Fit a model of the form, **$C(d) = C_0 \exp(-ad)$**
- Estimate the params **C_0** and the decay constant **a**.
- Hint: curve fitting from scipy library for python users

  curvefit from the Optimization Toolbox for Matlab users

- Plot this curve on the graph to show how quickly the correlation declines with distance.

Question 5

- Synchronize the dates corresponding to both time series
- Provide a well labelled scatter plot of vegetation index against rainfall
- Use different colors and symbols to distinguish between the different districts

Question 6

- Transform the rainfall time series by delaying it by k[0:12].
- Calculate its correlation with vegetation index for every district
- Evaluate how many months it takes to see the effect of rain on vegetation in each district, i,e the k giving the highest correlation
- Is there a consensus?

Question 7

- Transform the rainfall time series by using simple moving averages(SMA)
- The idea of using SMA is to smoothen out short-term fluctuations to see the long-term pattern.
- You will be averaging over the last n months with a window of k ranging from 1 to 12.

Question 7 (cont'd)

- What value of K gives the highest correlation between the smoothened rainfall index and vegetation index for each district.
- Provide a graph of correlations against K
- Comment on the graph

Question 8

- Examine if a quadratic model explains the relationship between
    - Rainfall index and vegetation index.
    - Delayed rainfall and vegetation index
    - simple moving average rainfall and vegetation index

NB: For SMA and delayed, the window size will be equal to the optimum k

.

Question 8 (cont'd)

- Test the same relationship across these models:
  - Linear regression
  - Quadratic regression
  - Cubic regression
- Compute the performance metrics for each of the above
  - Adjusted R-squared, RMSE and R-squared
- Analyze your results

Question 9

- Delayed SMA
  - Use optimal window for simple moving average from Q7.
  - Delay the SMA by optimal k from Q6 .
  - Fit linear, quadratic and cubic models.
- Use cross validation (AKA train_test_split) to test your models on the out of sample data, while computing performance metrics.
- Repeat the same process for the other features(Rainfall, delayed rainfall and SMA Rainfall)

Question 10

- Consider linear, nonlinear and nonparametric models (2 non-parametric)
- Select best feature based on your conclusion from #9
- Use coefficient of determination (R2 or RMSE) to evaluate the models.
- Plot the graphs with the fitted models with vegetation index against the rainfall feature (5 graphs expected)

Submission Instructions

- Submissions should be made via canvas.
- **Single** Python/MATLAB code file(.ipynb or .m) **[Do not Submit checkpoints for .ipynb].** In addition, each line of code should be documented by text. This demonstrates that the code is unique and owned by the student
- Assignment report(.pdf) with full evidence that the assignment was completed by the student and demonstrate a full understanding of each step in the process including textual descriptions of each result (statistics, table, graph etc) represents and insights that can be gained
    - Indicate the libraries you have used in your code at the beginning of the report (After the title page)
- Data files (as given)

Submission Instructions cont’

Submission process:

1. Put source code **file and data files** in a single folder
2. Name of the folder should be the same as your andrew ID
3. **Zip this folder and attach the zipped file on assignment submission page (CANVAS)**
4. After attaching zipped file, click on "Add Another File" from assignment submission page and **attach your report**
5. Submit your assignment

N.B. This process will allow us to compile your reports in **Turnitin** to check for plagiarism.

Specific reasons for a submission being classified as incomplete include:

- Failure to correctly name your folder with your Andrew ID, report, and code file with andrewID_BDS_AssignmentNo. For example, mcsharry_BDS_Assignment1, mcsharry_BDS_Assignment2 and mcsharry_BDS_Assignment3.
- A missing report describing the steps, results, and insights
- A missing dataset required for running the code
- A missing code file such as .ipynb or .m file
- An error in the file path needed to run the code

Submission Instructions’

The student is responsible for checking that their submission is complete. Students will lose 10% as for usual late submission even if the submission is repaired during the 24 hours after the deadline has passed, and receive 0 for the assignment if it is not repaired.

The submission deadline is on:

Monday 27, March, 2023 17:59 Eastern Time (ET) /

Monday 27, March, 2023 23:59 Rwandan Time (CAT)

Q&A ?