Assignment 1
Dengue Case Prediction

Yi-Ju Chen

Po-Chih Kuo

Introduction

- Millions of people suffer from the mosquito-borne disease dengue, whose primary habitat and breeding place is natural and artificial still-water bodies, which are directly related to environmental variables like temperature, precipitation, and so on.
- Early prediction, detection, and subsequent action may reduce the risk of mass infection and save lives.
- So in this assignment, we need to analyze the given dataset and predictive models to predict weekly dengue cases based in different cities in South America.

Dataset

- True Data
  - From the cities in South America
- Parts of data have been manipulated
  - Outliers and missing values were added
  - Dates were shifted
- Time-dependent data
  - Case number
  - Temperature
  - Precipitation
- Time-independent data
  - Elevation
  - Socio-economic variables

Goal

- Predict weekly dengue cases based in different cities.
- Combine temperature, precipitation, elevation, or other socio-economic variables to predict the weekly dengue cases
- Implement the regression model to achieve the prediction
- Preprocess/Split the data for model training/testing

Grading Policy

Item	Score
Basic Implementation	60%
Advanced Implementation	35%
Report	5%

Basic Implementation (60%)

- Given the average temperature and dengue fever cases in three cities in the past 94 weeks
- Build a regression model using temperature as an input variable to predict the number of cases in the next 10 weeks
- Number of cases could be another input variable
- Please use the file we provide as your input
- Print the coefficients of your model (5%)

Basic Grading Policy

- Baseline (including printing coefficients 5%) – 40%
  - Get all if average MAPE across all cities <= 35%
- Ranking – 20%
  - Average MAPE across all cities to compete with the whole class (only those above baseline)

Advanced Implementation (35%)

- Combined with other conditions or in a different way than the basic part to help your predictions for dengue cases in the next 10 weeks
- You can finish this part in any other way you like
- Using only temperature (same approach as basic model) will not get points.

Advanced Grading Policy

- Baseline – 25%
  - Average MAPE across all cities <= 30%
- Ranking – 10%
  - Average MAPE across all cities to compete with the whole class (only those above baseline)

Template

- You must use the given file “hw1_template.ipynb” to build the model
- Except for the imported packages in the template, you cannot use any other packages in the basic part
- There is no restriction on the format of the advanced part

HW1: Regression

In *assignment 1*, you need to:

1. Basic Part: Implement the regression model to predict dengue cases

   - Step 1: Split Data
   - Step 2: Preprocess Data
   - Step 3: Implement Regression
   - Step 4: Make Prediction
   - Step 5: Call the function of Step 1 to Step 4

2. Advanced Part: Implement the regression model with addtional conditions to predict dengue cases

1. Basic Part (60%)

In the first part, you need to implement the regression to predict dengue cases

Please put the prediction result in a csv file **hw1_basic.csv**

Import Packages

Note: You **cannot** import any other packages in the first part (implementation)!

```
[ ]  import numpy as np
     import matplotlib.pyplot as plt
     import csv
     import math
     import random
```

Basic Input File Format

- Named "**hw1_basic_input.csv**" and containing a (n+1) * 7 matrix, n means the number of weeks
- Each row represents "epiweek , TemperatureA, TemperatureB, TemperatureC, CityA, CityB, and CityC
- CityA is the number of dengue case, and so are CityB, and CityC
- TemperatureA is the average temperature of CityA, and so are TemperatureB,and TemperatureC
- The part to be predicted (202143 ~ 202152 of CityA, CityB, and CityC) is filled with 0
- We will use this format of csv file to test your model with n = 10 (202143 ~ 202152)
- Please make sure your model can be correctly input into this format of csv file

epiweek	Temperatu	Temperatu	Temperatu	CityA	CityB	CityC
202001	21.48	22.24	9.16	147	89	9
202002				146	99	7
202003	24.66	22.32	24.84	198	78	13
202004	23.89	24.9	29.66	180	69	14
202005	22.85	23.74	29.78	162	57	8
202006	27.49	25.41	30.38	127	52	14
202007	12.74	24.23	29.73	108	47	11
202008	26.2	21.51	27.98	99	51	15
202009	23.51	23.77	26.54	94	50	7
202143	29.48	19.61	28.86	0	0	0
202144	26.97	24.44	26.72	0	0	0
202145	27.01	22.53	28.7	0	0	0
202146	27.41	22.5	27.36	0	0	0
202147	24.39	23.14	27.18	0	0	0
202148	22.13	21.44	24.41	0	0	0
202149	24.51	22.16	30.4	0	0	0
202150	28.39	17.27	26.63	0	0	0
202151	25.93	19.07	29.29	0	0	0
202152	26.73	23.72	28.96	0	0	0

Advanced Input File Format

- Named “**hw1_advanced_input1.csv**” and “**hw1_advanced_input2.csv**”
- “hw1_advanced_input1.csv” is the precipitation data, which contains a (n+1) * 4 matrix, n means the number of weeks
- “hw1_advanced_input2.csv” is the sociodemographic data, which contains a (n+1) * 26 matrix, n means the number of cities

epiweek	Precipitati	Precipitati	PrecipitationC
202001	3.17	1.7	4.81
202002	7.8	8.21	0.16
202003	3.5	0.05	0.59
202004	0.35	0	0.89
202005	2.21	3.21	0.97
202006	0.32	0.66	0
202007	2.46	0.13	0
202008	0.81	4.58	8.46
202009	4.72	1.59	2.97
202010	15.1	9.36	5.51

	Populatior	Age0-4(%	Age5-14(%	Age15-29(	Age>30(%	Peoplewitl
CityA	2206804	5.37	12.38	24.73	57.52	9.67
CityB	2414616	5.16	11.96	26.71	56.16	6.27
CityC	521409	6.82	15.8	26.75	50.63	4.33

Output File Format

- The prediction of both basic and advanced you turned in must follow this format
- Named as “**hw1_basic.csv**” and “**hw1_advanced.csv**”, both contain a 10 * 4 matrix
- Each row represents “epiweek, CityA, CityB, and CityC” without header
- Please make sure your model can correctly output this format of csv file

202143	1	2	3
202144	1	2	3
202145	1	2	3
202146	1	2	3
202147	1	2	3
202148	1	2	3
202149	1	2	3
202150	1	2	3
202151	1	2	3
202152	1	2	3

Report

- Named as “**hw1_report.pdf**”
- Write down your **regression equation** in basic part (1%)
- Briefly describe the **variables** you used in the advanced part (1%)
  - No point would be given for the advanced part if you do not clearly point out the difference between the basic part and the advanced part
- Briefly describe the difficulty you encountered (1%)
- Summarize how you solve the difficulty and your reflections (2%)
- No more than one page

Assignment 1 Requirement

- Do it individually! Not as a team! (The team is for final project)
- Announce date: 2022/9/29
- Deadline: 2022/10/12 23:59 (Late submission is not allowed!)
- Hand in your files in the following format (Do not compressed!)
    - hw1_basic.csv
    - hw1_advanced.csv
    - hw1.ipynb
    - hw1_report.pdf
- Assignment 1 would be covered on the exam next time

The Evaluation Metric

- MAPE (Mean absolute percentage error):

$$\text{MAPE} = \frac{100\%}{N} \sum_{i=1}^{N} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

- For example:
  - The value you predicted:
  - $\hat{y}$ = [592, 486, 538, 689, 752, 841, 491]
  - Ground Truth :
  - y = [491, 584, 541, 599, 615, 741, 512]
  - MAPE = 1/7 *0.928 = 0.1326 = 13.26%
- We would evaluate your assignment by the average MAPE across all cities

HMI

Penalty

- 0 points if any of the following conditions happened
  - Plagiarism
  - Late submission
  - Not using a template or importing any other packages in the basic part
  - Incorrect prediction format
  - Incorrect submission format

Questions?

- TA: Yi-Ju Chen (ss111062511@gapp.nthu.edu.tw)
- Do not ask for debugging.