2-Day Project

You are given 2 questions in this area. Answer 1 question of your choice.

1. Seoul Air Prediction

Seoul Air Quality dataset[1] was used in [1, 2, 3][2], including Seoul air quality data from 2008 to 2018. Air quality is impacted by many factors such as traffic volume, neighboring area situations, weather, seasonal information, and other economic activities. Many works have addressed the relationship between the air quality level and other factors via numerous modeling approaches. For instance, during the Chuseok holidays, the air quality index tends to get better, while it is serious during weekdays, especially with foggy weather conditions or in the yellow dust season. You may refer to [1, 2, 3] for more information on how researchers used this dataset in their works.

Columns	Meaning
Datetime	Date and time
District	District code 0-25 (Code 0 represents the average value of all 25 districts in Seoul). Other districts are identified from 1 to 25.
PM10_CONC	PM10 concentration ($\mu g/m^3$)
PM2_5_CONC	PM2.5 concentration ($\mu g/m^3$)
O3	Ozone concentration ($\mu g/m^3$)
NO2	NO2 concentration ($\mu g/m^3$)
CO	CO concentration ($\mu g/m^3$)
SO2	SO2 concentration ($\mu g/m^3$)
PM10_AQI	PM10 AQI Index according to US Standard AQI Index
PM2_5_AQI	PM2.5 AQI Index according to US Standard AQI Index

Table 3: Seoul Air Quality Dataset

(a) **Additional Data Collection.** Of course, it is okay to treat the problem as a pure time-series problem. After that, you can apply traditional methods like ARIMA or SVR to solve the problem. However, these methods are only appropriate for data that have repeated patterns. In many problems, especially Air Quality forecasting, more relevant data help improve the accuracy of prediction models. Therefore, you may collect additional weather and holiday information. In [1], the author clearly explains how different data sources are collected. If you can collect more than two data sources and use them in your model, you get full credits. Please specify what additional datasets you used.

(b) **Data Cleansing & Data Loading.** Describe how you preprocess the data into a format you use for training.

(c) **Learning Algorithm.** Describe your learning algorithm and how it is implemented.

(d) **Evaluation Metric.** Define your evaluation metric and explain why it is appropriate for the problem.

(e) **Experimental Settings.** Clearly specify experimental settings and how you split the dataset into train/val/test sets.

(f) **Result and Discussion.** Present your work at the oral exam, including (but not limited to) problem setting, hypotheses, how you collect and preprocess data, description and justification of your learning algorithm, prediction results, interesting observations or insights learned, and potential future work.

[1] https://cleanair.seoul.go.kr/airquality/localAvg

[2] https://github.com/alexbui91/Air-Quality-Prediction-Datasets

2. New COVID Infection Prediction

Predicting future COVID19 infection is of a major importance in public health. The rate of infection can be affected by many factors including weather/temperature, government containment policy, economic policy, etc. Recently vaccination becomes an important factor. In this project, students need to use many different sources of data to build a model to predict short term (next 10 day) and more longer term (next 1 month) prediction of the number of new COVID cases. The following data can be used.

- COVID new case statistics
  `https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-data.csv`
- Government response data
  `https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker#data`
  `https://github.com/OxCGRT/covid-policy-tracker`
- Vaccination data
  `https://github.com/owid/covid-19-data/tree/master/public/data/vaccinations/country_data`

(a) **Country to build the model.** To build a model to predict COVID infection for two countries: Korea and US

(b) **Additional Data Collection.** In addition to the data listed, you can also used additional data to fit the model. When you use additional data, please specify the datasets.

(c) **Data Cleansing & Data Loading.** Describe how you preprocess the data into a format you use for training.

(d) **Learning Algorithm.** Describe your learning algorithm and how it is implemented.

(e) **Evaluation Metric.** Define your evaluation metric and explain why it is appropriate for the problem.

(f) **Experimental Settings.** Clearly specify experimental settings. To find the model, you may need to split the dataset into train/val/test sets. Please specify how this can be done.

(g) **Result and Discussion.** Present your work at the oral exam, including (but not limited to) problem setting, hypotheses, how you collect and preprocess data, description and justification of your learning algorithm, prediction results, interesting observations or insights learned, and potential future work.

References

[1] Tien-Cuong Bui, Joonyoung Kim, Taewoo Kang, Donghyeon Lee, Junyoung Choi, Insoon Yang, Kyomin Jung, and Sang Kyun Cha. STAR: Spatio-temporal prediction of air quality using a multimodal approach. In *Proc. of SAI Intelligent Systems Conference*, pages 389–406, 2020.

[2] Sookyung Kim, Jungmin M Lee, Jiwoo Lee, and Jihoon Seo. Deep-dust: Predicting concentrations of fine dust in seoul using LSTM. *Climate Informatics*, 2019.

[3] Van-Duc Le, Tien-Cuong Bui, and Sang-Kyun Cha. Spatiotemporal deep learning model for citywide air pollution interpolation and prediction. In *IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 55–62, 2020.