SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Intro to Big Data Science: Project

Deadline: May 27, 2022

✏ **Problem**(Beijing PM2.5 Diagnosis and Forecast)

Beijing and a substantial part of China are experiencing chronic air pollution. The main pollutants are fine particulate matter, and PM2.5 in particular. PM2.5 consists of airborne particles with aerodynamic diameters of less than 2.5 μm. They are known to influence visibility, human health and even climate. Epidemiological evidence shows that exposure to PM2.5 can cause lung morbidity, serious respiratory and cardiovascular diseases, and even death. Therefore, the assessment and monitoring of PM2.5 in Beijing is of great importance and of particular interests.

It seems to be a common sense that the PM2.5 could be influenced by meteorological effects, at least according to the experiences of Beijing residents. For instance, strong northwestern wind could alleviate the PM2.5 pollution. As a result, examining the relationship between the PM2.5 and meteorological effects is a quite useful and important task for PM2.5 diagnosis and forecast.

You are provided with historical data of PM2.5 concentrations and some meteorological indices in Beijing between Jan 1st, 2010 to Dec 31st, 2014. Note that there is some missing data in the data file "PRSA_data.csv".

This is a **regression** problem. The corresponding measures for the model quality are Mean Square Error, (adjusted) R^2 score, etc.

Data description: PRSA_data.csv

This is the historical training data, which covers to 2010-01-01 to 2014-12-31. Within this file you will find the following fields:

- **No** - row number
- **year** - year of data in this row

- **month** - month of data in this row
- **day** - day of data in this row
- **hour** - hour of data in this row
- **No** - row number
- **pm2.5** - PM2.5 concentration ($\mu g/m^3$)
- **DEWP** - Dew Point
- **TEMP** - Temperature
- **PRES** - Pressure (hPa)
- **cbwd** - Combined wind direction
- **lws** - Cumulated wind speed (m/s)
- **ls** - Cumulated hours of snow
- **lr** - Cumulated hours of rain

Your goal is to predict the PM2.5 concentration. This can be played in several tasks:

1. Just delete all items with missing PM2.5 ("NA"). Then split the remaining data to training and test sets by following this rule: extract the items (rows) for every 7-day from 2010-01-07 to 2014-12-25; that is, choose the items on 2010-01-07, 2010-01-14, 2010-01-21, 2010-01-28, 2010-02-04, 2010-02-11, ..., the last item is on 2014-12-25. Collect these items as the test set, and keep the remaining items as training set. Use the meteorological indices DEWP, TEMP, PRES, cbwd (need to be processed by one-hot coding), lws, ls, and lr, as covariate vector $\mathbf{X}$, and treat pm2.5 value as the response Y. Construct your own models to do this prediction, and validate your models on the test set.
2. Examine the tendency of PM2.5 variation, and its dependency on time period of the day (e.g, morning, afternoon, evening, night), the week (weekdays, weekends, or holidays), the season (spring, summer, fall, winter), and the year. You shall produce more features using the time information, e.g, Monday morning, evening of the weekend in December, spring festival of 2012, or State Council of China in October, 2013 (policy effect, the exact date and period for these events can be found on the web). Use the new features you just produce to predict PM2.5 again, and validate your models.
3. Now keep all items without deleting any missing data. Try to impute the missing values by some approaches, e.g, interpolating, averaging, model-based kNN filling, etc. Then combine all the information (full data) to do the prediction (the same thing as in part 1 and part 2).
4. Select the most important several features for the PM2.5 forecast. These features could be the original attributes in the dataset (e.g., DEWP, PRES, lws, etc); or some new features you constructed just now; or combinations of them (e.g., linear combination by PCA). Explain what they mean, and how important they

are (you should compute this based on some indices, e.g., correlation coefficients in linear regression, maximum information coefficients, cumulative eigenvalue proportion in PCA, feature importance in random forest, etc).

5. (Optional) You can even do something more. For example, as suggested in the reference given at the end, you may partition the PM2.5 time series into three states: low PM state when PM2.5 $\leqslant 35\mu g\, m^{-3}$; polluting episode when PM2.5 $> 35\mu g\, m^{-3}$; and very high PM when $PM2.5 > 150\mu g\, m^{-3}$. Then smooth the time series over 3-hour moving windows (e.g., the smoothed TEMP value at 2010/01/01 1h, is equal to the average TEMP values at 2010/01/01 0h, 2010/01/01 1h, and 2010/01/01 2h). Then treat this as a classification problem and make the prediction.

Your report should include the following several aspects:

1. Data exploration: data statistics or data visualization;
2. Data preprocessing: including detecting missing values (if any) and outlier samples (if any), data conversion and normalization (if necessary);
3. Model construction: you could use any model you prefer, even the model we did not cover in class;
4. Feature selection and model selection;
5. Model evaluation;
6. Conclusion.

This project should be finished in groups. Each group should consist of one or two students. You can find your partners by yourselves. Each group should contribute a final report in the submission.

DO NOT just submit the code file. Necessary statements, analysis, formula, figures, and tables should be included in your report. You should also have a complete set of codes. Your report (typically in pdf format) and codes should be compressed in a zip file. Please use your student ID and name to rename your zip file, e.g., "11800000_张三". Then the zip file shall be uploaded to BlackBoard system.

Your project will be graded based on several factors, including the accuracy (e.g., F_1 score, R^2 score, etc.), comparison of different methods, whether you have innovative ideas, the quality of your report, the analysis you made based on you results (e.g., computational efficiency, model interpretability, etc.), and the quality of your codes, but not limited to these.

Last but not the least, we hope you really get familiar with whole data science procedure and discover the new world of your own.

For references, we provide one paper concerning the PM2.5 problem:

http://rspa.royalsocietypublishing.org/content/471/2182/20150257

The paper is also packed with this file and data.