SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Intro to Big Data Science: Project 2

Due Date: May 21, 2019

✏ **Problem**(Bike sharing demand: regression)

Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return back has become automatic. Through these systems, user is able to easily rent a bike from a particular position and return back at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousands bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

Apart from interesting real world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of important events in the city could be detected via monitoring these data.

This dataset contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system in Washington, DC with the corresponding weather and seasonal information.

Data description:

Both hour.csv and day.csv have the following fields, except hr which is not available in day.csv

- instant: record index
- dteday: date

- season: season (1:springer, 2:summer, 3:fall, 4:winter)
- yr: year (0: 2011, 1:2012)
- mnth: month (1 to 12)
- hr: hour (0 to 23)
- holiday: weather day is holiday or not (extracted from [Web Link])
- weekday: day of the week
- workingday: if day is neither weekend nor holiday is 1, otherwise is 0.
- weathersit:
  * 1: Clear, Few clouds, Partly cloudy, Partly cloudy
  * 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  * 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  * 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp: Normalized temperature in Celsius. The values are derived via $(t-t_{min})/(t_{max}-t_{min})$, $t_{min}=-8$, $t_{max}=+39$ (only in hourly scale)
- atemp: Normalized feeling temperature in Celsius. The values are derived via $(t-t_{min})/(t_{max}-t_{min})$, $t_{min}=-16$, $t_{max}=+50$ (only in hourly scale)
- hum: Normalized humidity. The values are divided to 100 (max)
- windspeed: Normalized wind speed. The values are divided to 67 (max)
- casual: count of casual users
- registered: count of registered users
- cnt: count of total rental bikes including both casual and registered

Your goal is to predict the count of total rental bikes including both casual and registered. This can be played in several tasks:

1. Split the data to training and test sets in this way: extract the items (rows) in the first 19 days of each month and treat them as the training data; the remaining items form the test dataset. E.g., the data within in the period "2011-01-01 to 2011-1-19" belong to training set, while the data within in the period "2011-01-20 to 2011-1-31" belong to test set. Treat the attributes except for "causal", "registered" and "cnt" as covariate vector $\mathbf{X}$, and treat "cnt" as the response Y. Construct your own models to do this prediction, and validate your models on the test set. You can do this on both data sets "day.csv" and "hour.csv".
2. "casual" means some users did not register and just rent the bike occasionally. Treat "casual" as Y and do the same as in step 1. What can you say about the prediction of "casual"?

3. Examine the variation tendency of the "cnt" value, 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, National Day of 2011, Hurricane Sandy on 2012-10-30 (abnormal weather or event, the exact date and period for these events can be found on the web). Use the new features you just produce to predict "cnt" again, and validate your models.
4. Select the most important several features for the "cnt" forecast. These features could be the original attributes in the dataset, or some new features you created in step 3; 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) Event and Anomaly Detection: Count of rented bikes are also correlated to some events in the town which easily are traceable via search engines. For instance, query like "2012-10-30 washington d.c." in Google returns related results to Hurricane Sandy. Some of the important events are identified in [1]. Therefore the data can be used for validation of anomaly or event detection algorithms as well. Can you find a way to do this anomaly detection?

Your report should include the following several aspects:

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

You are required to write a report, just like the format of the paper provided. **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. You 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**. Then the zip file shall be uploaded to cookdata.

Your project will be graded based on several factors, including the accuracy (e.g., mean square error, mean absolute error, R^2 score, or adjusted R^2, etc.), comparison of different methods, whether your methods are innovative, the quality of your report, the

analysis on your methods and you results (e.g., computational efficiency, model interpretability, etc.), and the quality of your codes.

Students must finish and submit their work **individually**.

For references, see the website:

http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset#

The paper is also packed with this file and data.

REFERENCES

[1] FANAEE-T, HADI, AND GAMA, JOAO, *Event labeling combining ensemble detectors and background knowledge*, Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg.