**Project 1. Implementing a Regression Project**

In Project 1, you will implement a regression project by applying a real-world bike sharing dataset!

The dataset collected hourly bike volume from 113 neighborhoods in Shanghai, China during 8~9 AM and 14~15 PM on August 1, 2016. You are now predicting hourly shared bike volume in each neighborhood using a bunch of features shown in Table 1. The features fall into three categories, built environment factors (including land use, road design, and facilities), social demographics, and temporal factors.

The project shall consist of the following components:

* (15) Read the data into Google Colab and describe the data in the following process including (not limited to): the number of rows and columns, statistic summary (e.g. mean, median, min, and max values of each variable). Do not forget to visualize the data by presenting histogram or boxplot of variables, and scatter plots between the outcome variable and features.
* (30) Linear regression. First, split the input data into training and test dataset. Fit the multiple regression model using the training dataset. You need to print the coefficients, intercept, and R square. Then, you need to apply the fitted model to the test dataset. Be sure to evaluate the model performance. This goal can be achieved by comparing actual and predicted values in one table and in bar plot, and printing training error and test error. You should plot the residual plot for the training and test dataset and analyze when the linear model fits the dataset well and when it does not. Do you think the Linear model is a good fit to the dataset? If not, you are encouraged (but not required) to provide some approaches to improve its performance.
* (20) Ridge regression. Similar to steps in Linear regression, you should fit the multiple regression using the training dataset and apply the fitted model to the test dataset. Besides, you shall determine the best by cross validation.
* (20) Lasso regression. Similar to steps in Linear regression, you should fit the multiple regression using the training dataset and apply the fitted model to the test dataset. Besides, you shall determine the best by cross validation.
* (15) Model comparison. Present coefficients of Linear regression, Ridge regression, and Lasso regression in one table and their performances (R square, training error, test error, and sum of absolute weights) in another. Plot the coefficients magnitude for three models. Analyze their differences. Which of the three models is better for this bike sharing dataset?

Have fun! :)

Table 1. Description of Features

|  |  |  |
| --- | --- | --- |
| Category | Variable | Description |
| Land use | Pct. Residential | Percentage of residential land, in % |
| Pct. Parks | Percentage of park land use, in % |
| LUM | The entropy of four types of land use, ranges from 0 to 1 |
| Urban density | FAR | Floor area ratio, the zonal-average number of floors, in counts |
| Road design | Bus stop density | The density of bus stops in the corresponding neighborhood, in counts/km2 |
| Metro-station density | The density of metro stations in the corresponding neighborhood, in counts/km2 |
| Primary road density | The length of primary roads in the corresponding neighborhood, in counts/km2 |
| Secondary road density | The length of second roads in the corresponding neighborhood, in counts/km2 |
| Tertiary road density | The length of tertiary roads in the corresponding neighborhood, in counts/km2 |
| Intersection density | The density of intersections in the corresponding neighborhood, in counts/km2 |
| Facilities | School density | The density of schools in the corresponding neighborhood, in counts/km2 |
| College density | The density of colleges in the corresponding neighborhood, in counts/km2 |
| Demographics | Pct. Females | Percentage of females in the corresponding neighborhood, in % |
| Pct. Seniors | Percentage of people older than 65 in the corresponding neighborhood, in % |
| Pct. Children | Percentage of people younger than 16 in the corresponding neighborhood, in % |
| Temporal | Peak hour | If bikes are counted during peak hours, 1, else, 0 |