CSL 7020 Assignment 1

Topic: Linear regression

Due: 15.09.2019

1. Dataset

   Air pollutant data is collected from two air quality monitoring sites
   (a) Alsip Village (AV) in Southern Cook country, Illinios (AQS ID: 17-031-0001; latitude/longitude:41.670992/−87.732457
   (b) Lemont Village (LV) in southwestern Cook County, Illinois (AQS ID: 17-031-1601; latitude/longitude: 41.66812/−87.99057.

   Air pollutant data has concentration of O_3, $PM_{2.5}$ and SO_2. We have used Sample measurement as concentration.
   Air pollutant data is downloaded from the U.S. EPA's Air Quality System (AQS) database (https://www.epa.gov/outdoor-air-quality-data).

   Meteorological data is collected from two weather stations
   (a) The weather station situated in Lansing Municipal Airport (LMA) is the closest meteorological site (MesoWest ID: KIGQ; latitude/longitude: 41.54125/−87.52822) to the AV air quality monitoring site.
   (b) The weather station positioned at Lewis University (LU) is the closest meteorological site (MesoWest ID: KLOT; latitude/longitude: 41.60307/−88.10164) to the LV air quality monitoring site.

   We have used Temperature, relative humidity, wind speed and direction, wind gust, precipitation accumulation, visibility, dew point, wind cardinal direction, pressure, and weather conditions.
   Meteorological data is downloaded from MesoWest (http://mesowest.utah.edu/).

2. Approach

   2.1 Pre-processing

   We have paired meteorological data and air pollutant targets on the basis of time to apply linear regression. We need value of each variable for every hour but meteorological data has multiple values in interval of one hour. Therefore we have taken hourly mean of each numerical values if there were multiple entries within an hour. For categorial data we have chosen one having highest frequency. There were missing values for some variables. We have interpolated those missing values. We deleted days having missing values even after interpolation.
   We have made one hot encoding for two categorial variables, the cardinal wind direction (16 values, e.g., N, S, E, W, etc.) and weather conditions (26 values, e.g., sunny, rainy, windy, etc.).

We have also used weekday and weekend as two boolean features.
Finally, we obtained total 55 features (9 numerical meteorological features, 16 hot vectors for wind direction, 26 hot vectors for weather conditions, 2 boolean features for weekday/weekend, 1 numerical feature for previous day pollutants, and 1 bias term).
We have done normalization for all the features and pollutant targets so that values will lies in range [0,1].

2.2 Model Description

we have used former day's pollutant data and metrological data to predict the next day's hourly pollutants.
Let (x_i; y_i) denote the ith training data, where $y_i \in R^{24*1}$denotes the concentration of a certain air pollutant on a day, and $x_i = (u_i; v_i)$ denotes the observed data on the previous day where component $u_i \in R^{24*54}$ includes all meteorological data over 24 h for the previous day and $v_i \in R^{24*1}$ includes the hourly concentration of the same air pollutant on the previous day. W^{55*1} denotes the parameters of the model.

We have minimised $\frac{1}{n}\sum_{i=1}^{n} || < x_i , \mathrm{W} > -\mathrm{yi} ||^2$ w.r.t to W using gradient descent.

3. Experiments

3.1 Setup

We have concatenated Meteorological data for two MesoWest ID: KLOT and KIGQ column wise. Then we have lexicographically sorted the data and then interpolated it in both direction. Now we are left with no empty values. Then we have taken hourly mean for all the numerical variables and taken highest frequency term for categorical data. Then we have formed one hot vector with size 16 for wind direction and one hot vector with size 26 for weather condition.

In case of Air pollutant data, data is repeated for whole year, So we have taken mean of it. We have made Date-time format of air pollutant data same as in case of metrological data.
Now we have taken inner join of air pollutant data and metrological data.
We have given 55 features and pollutant target as input to linear regression model.
Regression model returns weight matrix which will be used to obtain output.

3.2 Results

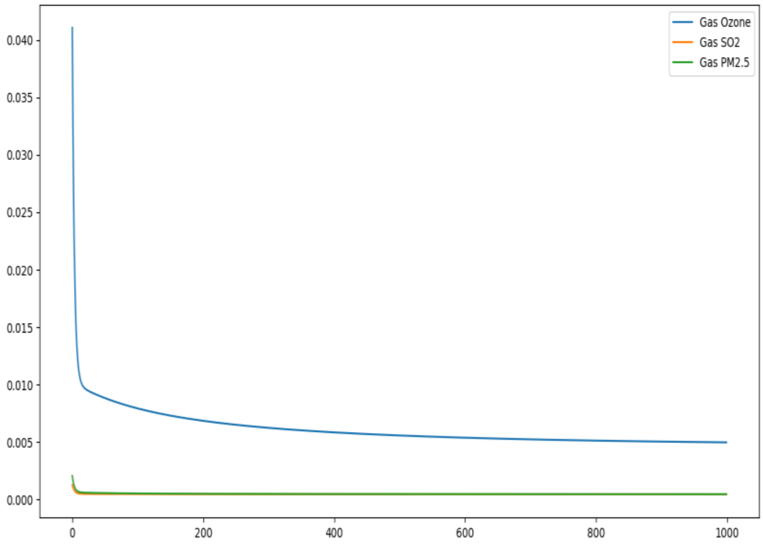

Figure 1: Empirical loss vs iteration curve (Target normalised)

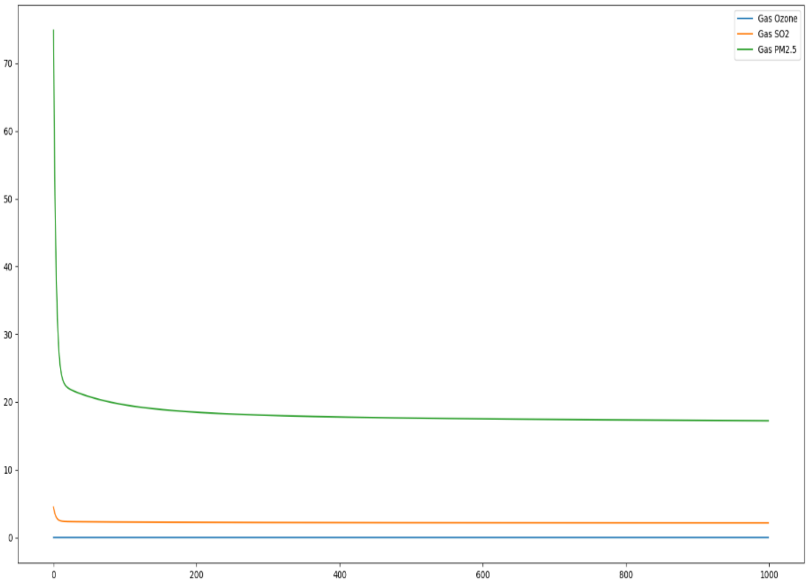

Figure 2: Empirical loss vs iteration curve (Target not normalised)

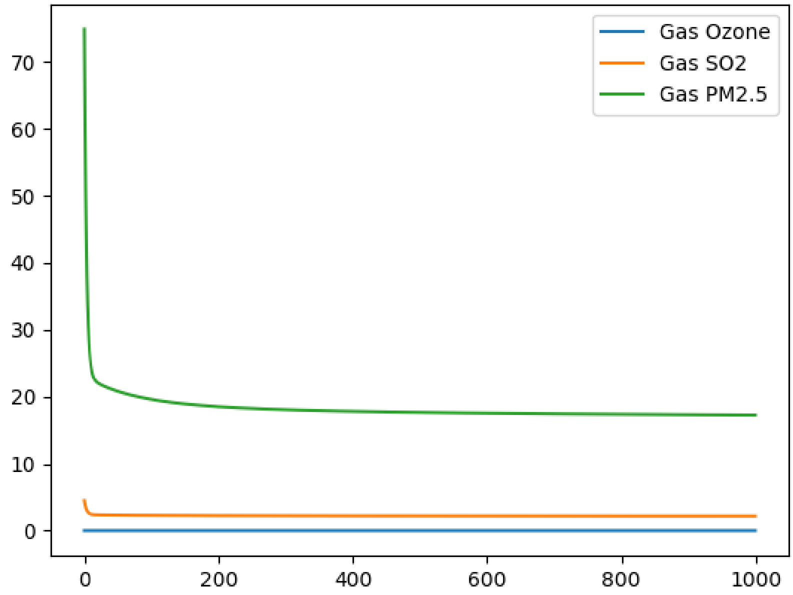

Figure 3: Empirical loss vs iteration curve with regularization (Target normalised)

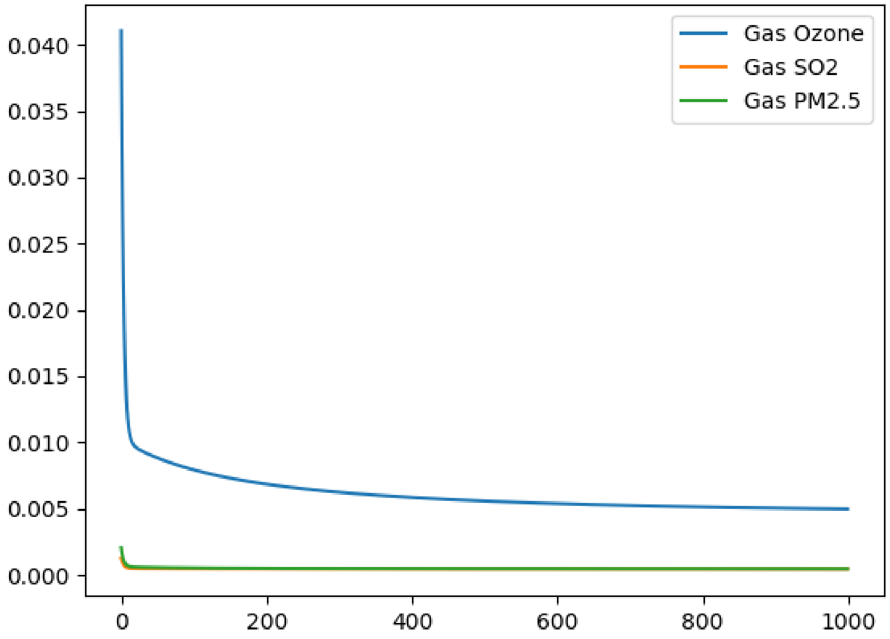

Figure 4: Empirical loss vs iteration curve with regularization (Target not normalised)

4. Analysis
    a) Without normalisation, it was taking more time for training and result was also not optimum.
    b) When we have normalised target, loss decreased.
    c) Loss decreases with regularization.