REPORT

Swati Sisodia
Shreyashi Sharma

Goal:

To predict the correct flow measurement, given a collection of erroneous measurement data (e.g. flow, speed, occupancy) where most of the measurement is correct.

Intuition:

The flow measurement of a particular lane is affected by the speed and occupancy measurement of the same lane as well as the hour of the day and the flow measurement of the nearby lanes. Hence we thought of considering all these factors in order to accurately predict the flow measurement.

Approach:

- **Prediction model:**

We decided to use SGDRegressor in order to train and test the data and finally predict our flow values. SGDRegressor is a linear model fitted by minimizing a regularized empirical loss with SGD. SGD stands for Stochastic Gradient Descent: the gradient of the loss is estimated each sample at a time and the model is updated along the way with a decreasing strength schedule (aka learning rate). It trains the dataset in mini batches making it much faster than a linear or logistic regressor.

$$w := w - \eta \nabla Q(w) = w - \eta \sum_{i=1}^{n} \nabla Q_i(w),$$

- **Implementation steps:**

1. We split our data in two types of samples – training sample and corrupt sample. Training sample consisted of all the valid data. Corrupt sample contained the improbable data (probability = 0) and the missing data.
2. We grouped the training data hour wise in order to find out the hourly average measurements.
3. All the invalid values in the corrupt sample were replaced by these average measurements.

4. The training sample was used to train our prediction model based on SGDRegressor as mentioned above. We used a learning rate of 0.00005 which was found out experimentally.
5. We then combined the training and corrupt sample and made predictions for the complete dataset using the model.

Challenges:

The biggest and most important challenge was the size of the dataset. Since most of the prediction models take a considerable amount of time to train as well as predict hence we overcame this challenge by using the mini batch training approach of SGDRegressor.

Also it took us a while to find out the correct value of the learning rate which gave us close predictions.

Results and Runtime:

Mean absolute error for the results came out to be 1.76 approx. Also, the runtime for the largest dataset came out to be 357462 milliseconds.