YZV202E Optimization for Data Science Project Proposal Traffic Congestion Prediction Optimization

Anıl Dervişoğlu
Artificial Intelligence and Data Engineering
Istanbul Technical University
dervisoglua22@itu.edu.tr
150220344

Ömer Faruk Zeybek
Artificial Intelligence and Data Engineering
Istanbul Technical University
zeybeko22@itu.edu.tr
150220743

I. Project Description

With the increasing population and crowded traffic in Istanbul, the aim of this project is to understand the historical data and predict a self made traffic congestion level showing a value between $[0, 1]$ range indicating how dense the traffic actually is.

Following our motivation, we also decided to understand and explore the harder path, optimization in neural networks. To do so we have analyzed several models as to understand which would fit the limited data we have in hand the best. And with input parameters of the specific day of the week, hour, month (for seasonal patterns), the geospatial location, whether it is a special day, grid population in the location, the weather (will be optional), average speed (optional - as the information requires observation data unlike other inputs), number of railway and bus stations in 1 km radius; as the data we had were in time series format and we wanted also use the information from the past inputs, we decided to go on with LSTM model.

For the sustainability area, our project is directly related with public health with the increased air pollution, stress (people might get angry), less physical activity and so on depending on the traffic congestion. It also affects the country economically as it reduces immediate access to many services. With our project, by predicting congestion accurately, we expect to support the ”sustainable” decision making in transportation.

II. Problem Definition

As this is only project proposal, we will define our problem state. Formally, the problem we try to solve in this project is to understand and predict the traffic congestion level given specific time, hour, day, month, location and many other parameters described below; with a value between $[0, 1]$.

Below, each input parameter and their specific range can be seen also with their abbreviations to use later in the proposal, $t \geq 0$ stands for the time step while x_n stands for the input parameter order:

$$\begin{aligned}
\text{latitude} &= la \in [40.8120, 41.4000] = x_1^t \\
\text{longitude} &= lo \in [28.9760, 29.4960] = x_2^t \\
\text{is\_special\_day} &= s \in \{0, 1\} = x_3^t \\
\text{grid\_population} &= p \geq 0 = x_4^t \\
\text{\#\_of\_bus\_stops} &= bs \geq 0 = x_5^t \\
\text{\#\_of\_railway\_stops} &= rs \geq 0 = x_6^t \\
\text{hour} &= h \in \{0, 1, \dots 23\} = x_7^t \\
\text{day\_of\_week} &= d \in \{1, 2, \dots, 7\} = x_8^t \\
\text{month} &= m \in \{1, 2, \dots 12\} = x_9^t
\end{aligned}$$

We can formally represent the input vector and output as follows:

$$\begin{aligned}
&X_t = [x_1^t, x_2^t, \dots x_9^t] \\
&y_t \in [0, 1]
\end{aligned}$$

The objective function (LSTM) can be then represented as (where k is the number of time steps used):

$$\hat{y}_t = f(X_{t-k}, X_{t-k+1} \dots X_t)$$

and the loss function to use our optimization functions would be (we will start by applying mean squared error and experiment different ones as well):

$$L(\hat{y}, y)$$

We choose to first apply the Bayesian directly as the optimization to see the affects (the general internal parameter θ optimization can shown as, with θ^* as the optimal):

$$\theta^* = \arg\min_{\theta} L(\hat{y}, y)$$

We will also try a combination of Particle Swarm Optimization with gradient as another addition alongside the direction, personal best and global best. Nonetheless, we will compare our results with gradient based methods like SGD and Adam to have a better comparison and understanding.

III. Dataset

We chose seven different datasets to complete our project. Descriptions of the datasets are given below:

- **Hourly Traffic Density Dataset**
  - This dataset contains hourly Istanbul location density and traffic information.
  - **Purpose of the dataset:** This dataset contains both location and hourly data. It also shows the average, maximum and minimum speed and the number of vehicles. This allows us to better predict the peak hours of traffic congestion. Also we will use this dataset to help our model understand monthly patterns. After converting time into a numerical format with a sine cycle, we will divide by 12 to represent each month of each year. This will allow the model to better learn the annual and monthly cycle and accurately predict the impact of changes.
  - **Data Source:** Ulusal Akıllı Şehir Platformu
- **Road Maintenance Works Web Service**
  - This dataset provides access to daily road maintenance works carried out by the Road Maintenance and Infrastructure Coordination Department.
  - **Purpose of the dataset:** This dataset shows areas where road works are taking place. We can predict that traffic density will be higher in areas close to these areas, or we can predict that roads connecting to this road may also affect other roads.
  - **Data Source:** Ulusal Akıllı Şehir Platformu
- **IETT Bus Stops Data**
  - This dataset contains vector data of IETT bus stops.
  - **Purpose of the dataset:** Since buses are also traffic generators, it is reasonable to expect higher traffic density around bus stops, especially within a 1 km radius. This assumption is based on our estimates so it might not be correct. We aim to find out by testing.
  - **Data Source:** Ulusal Akıllı Şehir Platformu
- **Rail System Station Points Data**
  - This dataset contains vector data of rail system station points.
  - **Purpose of the dataset:** It has a similar logic to the logic that there will be more traffic density near bus stops, but since the subways are not in traffic, we think that it can reduce traffic density around the area where it is located. Again, we will find out by testing.
  - **Data Source:** Ulusal Akıllı Şehir Platformu
- **Population Density Dataset**
  - This dataset shows the density of population by region.
  - **Purpose of the dataset:** Using this dataset, which includes the population density in the districts of Istanbul, we can predict that traffic will be higher on roads with higher population density and lower on roads with lower population density.
  - **Data Source:** TÜİK
- **Weather Condition Dataset (External Data - Optional)**
  - This dataset contains weather information for Istanbul from 2020 to 2025.
  - **Purpose of the dataset:** Weather conditions are one of the most important factors in traffic density, this dataset contains weather information on a daily basis. We know that traffic congestion is higher when it is rainy, snowy and foggy. With this dataset, we can get better results.
  - **Data Source:** Visual Crossing

IV. Methodology

- **Data Preparation:** We will select some set of features that are important and necessary for our model. Temporal features such as days, hours and months will be processed using a sine function to capture periodic patterns. Additional features include weather conditions, regional population data and indicators for public holidays or special events, as well as traffic congestion levels from previous periods, and especially metro and bus stations near roads within a 1 km radius. After feature selection, we will normalize the continuous variables in our data to achieve efficient training and convergence.
- **Model Architecture:** As discussed, we will be using LSTM in our project as we are dealing with time series prediction. Given an input of k time steps in the form of X_t feature vectors, we will be providing the several input features we have to the model. Then, our base model will contain number of LSTM layers that learns the relationships within data. Also, to have an output $\hat{y}_t$ within the given $[0, 1]$ range we will be using a sigmoid activation function as to output. We will also try to improve our model with dropout layers to prevent overfitting and generalize the model.
- **Optimization Techniques:** At first, we will try Bayesian Optimization as it is cost effective compared to many other methods and try to see the effects of gradient free methods. On another hand we will also try using a hybrid approach using Particle Swarm Optimization with additional parameter of gradient. We expect this approach to have quite the cost and might change it along the way but our main reason is to see if we can find global optimums rather than local approximations and have a better model in the end. finally we will compare our results with gradient based methods of SGD and Adam.
- **Evaluation Metrics:** As explained in the problem definition, we will evaluate our metrics with Mean Squared Error initially but will experiment alternatives as well.