Assignment

Description

Traffic conditions on a section of a road can be characterized using the average speed of vehicles passed during a fixed interval of time (e.g. 15 minutes) as measured by the sensors installed on the road. For example, during a traffic congestion, the average speed will be lower than the usual observed value for that time of the day. Traffic conditions depend on many factors: number of vehicles, hour of the day, day of week, type of day (holiday, before holiday, etc…), weather conditions and events occurred on the road (e.g. accidents and roadworks).

Your task is to predict the average speeds, for each sensor present in the dataset, for the next hour when traffic is influenced by these events on the monitored kilometers. You are required to produce 4 predictions, one for each quarter of the next hour. The average speed is computed as the mean speed of all the vehicles over a 15-minutes interval. The evaluation metric is the Mean Absolute Error (MAE).

Data

To complete this assignment, you are given five datasets as gzipped csv files with the semicolon as separator (apart for *distances.csv.gz* which will be treated separately). They consist of records from 01-08-2018 to 01-12-2018.

- *speeds.csv.gz*. The main dataset, it contains the average speeds observed during a 15-minutes periods. Periods start at DATETIME_UTC and end 15 minutes later. Speeds are observed by sensors located on the road KEY at kilometer KM. In addition to average speed (SPEED_AVG), also minimum, maximum and standard deviations of the speeds of the passed vehicles are provided. N_VEHICLES is the number of vehicles passed. Speeds have been rescaled to anonymize data. You can assume that they are expressed in kilometers per hour. If a value is not provided, it is missing. This means either that there were no vehicles or that the sensor was not working properly; there is no way to distinguish these possibilities.
- *events.csv.gz*. The dataset with the events occurred on the roads. The start and end of the events are provided as they are inserted by operators. The event can be located on the roads using the identifier of the road (KEY) and the initial and final kilometers. Every event has a type EVENT_TYPE (“accident”, “alarm”, etc…) and a subtype EVENT_DETAIL (the values have been remapped into integers).
- *weather.csv.gz*. The dataset with the weather conditions observed in a limited set of fixed weather stations. Weather conditions are recorded hourly. The stations can be very distant from the roads. Moreover, there are missing values. In order to assign weather conditions to kilometers you should find for each kilometer and datetime the nearest station which transmitted weather conditions. ID represents the identifier of the weather station. Minimum and maximum temperatures should be intended as daily extrema.

- *distances.csv.gz*. This file is a csv file but should be interpreted as a key-value file. Each row should be first split on "|" and two strings should be obtained. The first string is constituted by two values (separated by a comma), the road KEY and the kilometer KM. The second string is a list of values (the separator is the comma) that should be interpreted as tuples; the odd values are the ID of the weather stations and the even ones are the distances between the stations and the kilometer inferred from the first string.
- *sensors.csv.gz*. Some KPIs about the road characteristics near the traffic sensors.

Considerations

1. The speed dataset contains the average speeds for both normal traffic conditions and traffic affected by the presence of events. Remember that the model objective is to predict well the average speed in presence of events. For the assignment, we do not care about prediction of speeds in normal traffic conditions.
2. Road names and kilometers have been anonymized. However, the distances between sensors on the same road are correct and also those between sensors and events since we simply added a value dependent on the road.
3. When you merge the different datasets, it is possible that some keys are not present in all datasets. These rows can be ignored (after having imputed missing values).
4. All the provided datasets are real ones. As such, they are noisy and with missing values. Average speeds and weather conditions do have missing values. You are supposed to treat them in both cases.
5. Events are noisy in the sense that often they have little to no effect to typical traffic conditions; considering this fact can be useful when defining the strategy to solve the task.
6. Only two datasets are necessary to solve the task (speeds and events); however, you are encouraged to use all of them (also because it is quite reasonable that traffic depends on road characteristics and weather conditions).
7. The test set will contain the data from 01-12-2018 to 31-12-2018. It will be disclosed right before the expositions of the results. It will consist of three files containing speeds, weather conditions and events.

Evaluation

The evaluation metric is the mean absolute error between the predicted speed and the real one. Performances should be evaluated on the test set built as previously described. The following errors should be computed:

- Total MAE: MAE on the predictions of the test set, filtering only the time intervals when events occurred.
- MAE per event type: MAE on all the events type regardless of the quarter.

MAE is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|pred_i - real_i|$$

You can use the sklearn function mean_absolute_error (from sklearn.metrics import mean_absolute_error).