Traffic Prediction

Abstract

Accurate traffic forecast is crucial for traffic control. In this competition, participants will be asked to give traffic predictions for 228 sensor stations in a region based on traffic speeds of the previous hour.

1 Description

Consider the traffic speed prediction problem for 228 sensor stations in a region. We split one day into 288 time periods (5 minutes per time period), and count an average speed in each time period for every sensor station.

The problem you will solve is: Given the speed of the pervious hour (12 time periods) of these 228 stations $v_{1,i}, ..., v_{12,i}$, predict the speed after 15, 30, 45 minutes $\hat{v}_{15,i}, \hat{v}_{18,i}, \hat{v}_{21,i}$, where $1 \leq i \leq 228$.

Note that we also give the distance relationship of these 288 sensor stations, you may be able to consider the interaction of their traffic conditions through these distance information.

2 Dataset

The dataset was collected from Caltrans Performance Measurement System (PeMS) in real-time by over $39,000$ sensor stations, deployed across the major metropolitan areas of California state highway system [1]. The dataset is aggregated into 5-minute interval, so there is 288 time periods per day. We randomly select among the District 7 of California containing 228 stations as data sources. The time range of the dataset is in the weekdays of May and June of 2012. We select the first 34 weekdays of historical speed records as training set, and the rest serves as the test set.

- distance.csv: a matrix $D_{228\times228}$, where $d_{i,j}$ denotes the distance between sensor station i and sensor station j (unit: meter).

- train.tar.gz: including 34 csv files, each file is a matrix $V_{288\times228}$, contains the traffic speed data of 228 sensor stations in one weekday, where $v_{t,i}$ is the traffic speed of sensor station i in the t-th time period (5 min) that day.
- test.tar.gz: including 80 test files, each file is a matrix $X_{12\times228}$, contains traffic speed of 228 sensor stations in an hour. Participants should predict the traffic speed after 15, 30, and 45 minutes, i.e. given traffic speed of time period $1, 2, ..., 12$, predict the traffic speed of time period $15, 18, 21$.

3 Submission

As shown in figure 1, the submission file should be a csv file containing 2 columns: Id and Expected.

Id is a string "*d_t_i*", denotes the d-th test file, the i-th sensor station, t minutes prediction.

Expected column: you need to fill in your predictions in the corresponding position.

	A	B
1	Id	Expected
2	0_15_0	70.3
3	0_15_1	64.9
4	0_15_2	67.9
5	0_15_3	56.8
6	0_15_4	72.3
7	0_15_5	67.4
8	0_15_6	63.4
9	0_15_7	61.8
10	0_15_8	62.5
11	0_15_9	16.2
12	0_15_10	62
13	0_15_11	66.1
14	0_15_12	64.8
15	0_15_13	70.5
16	0_15_14	61.2
17	0_15_15	63.6
18	0_15_16	65.1

Figure 1: A sample submission file

See "sample.csv" for the correct format.

4 Evaluation

We adopt RMSE metric, for prediction $\hat{v}_{d\_t\_i}$ and actual traffic speed $v_{d\_t\_i}$, the loss function is:

$$L = \sqrt{\underset{\forall d,t,i}{\text{mean}} ||\hat{v}_{d\_t\_i} - v_{d\_t\_i}||_2^2} \tag{1}$$

References

[1] Chen C, Petty K, Skabardonis A. Freeway Performance Measurement System: Mining Loop Detector Data[J]. *Transportation Research Record Journal of the Transportation Research Board*, 2001, 1748(1).