We release parts of the codes in paper [1]. All the codes are **MATLAB** scripts. Codes are organized in two folders, Check-out Proportion and Check-in Inference. Please cite the following two papers when using the codes.

[1] Yexin Li, Yu Zheng, Huichu Zhang and Lei Chen. Traffic prediction in a bike-sharing system. In *Proceedings of* the 23rd ACM SIGSPATIAL GIS, 2015.

[2] Yu Zheng, Licia Vapra, Ouri Wolfson, and Hai Yang. Urban Computing: concepts, methodologies, and applications. ACM Transactions on Intelligent Systems and Technology, vol. 5, no. 3, pp. 38:1-38:55, 2014.

In all the examples given in this document, we assume that there are 5 historical hours as training data and 3 future hours as testing data. We assume that there are 100 bike stations, which are clustered into 4 clusters.

Check-out Prediction

The codes in "Check-out Prediction" folder aim to predict the check-out across clusters.

Run "p_O.m" in "Codes\Check-out Proportion".

➢ **Input**

- *Historical entire traffic*
  The historical entire traffic is a vector, e.g. (200, 170, 210, 300, 100), each entry of which is the entire traffic of a special hour, e.g. the entire traffic in the first historical hour is 200.
- *Historical check-out across clusters*
  The historical check-out across clusters is a matrix, e.g.

$$\begin{pmatrix} 50 & 100 & 20 & 30 \\ 45 & 55 & 10 & 60 \\ 100 & 10 & 70 & 30 \\ 70 & 35 & 115 & 90 \\ 10 & 10 & 55 & 25 \end{pmatrix}.$$

  Each row corresponds to a historical hour and each column corresponds to a cluster, e.g. in the first historical hour, the check-out across clusters is 50, 100, 20, 30 respectively.
- *Historical check-out proportion*
  The historical check-out proportion is a matrix, e.g.

$$\begin{pmatrix} 0.25 & 0.5 & 0.1 & 0.15 \\ 0.265 & 0.324 & 0.059 & 0.352 \\ 0.476 & 0.048 & 0.333 & 0.143 \\ 0.233 & 0.117 & 0.35 & 0.3 \\ 0.1 & 0.1 & 0.55 & 0.25 \end{pmatrix}.$$

  Each row corresponds to a historical hour and each column corresponds to a cluster, e.g. in the first historical hour, the check-out proportion of each cluster is 0.25, 0.5, 0.1, 0.15 respectively.
- *Historical feature*
  The historical feature is a matrix, e.g.

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 41 & 10.4 \\ 1 & 0 & 0 & 0 & 1 & 41 & 5.8 \\ 1 & 0 & 0 & 1 & 0 & 39.9 & 9.2 \\ 0 & 0 & 1 & 0 & 0 & 42.1 & 4.6 \\ 1 & 1 & 0 & 0 & 0 & 45 & 6.9 \end{pmatrix}.$$

Each row corresponds to a historical hour and each column corresponds to a feature. In each row, the first entry corresponds to a working day (when it is “1”) or a weekend/holiday (when it is “0”); the second to the fifth entries describe the weather, corresponding to snowy, rainy, foggy and sunny respectively, e.g. the weather of the first historical hour in the matrix is (0, 0, 0, 1), which means a sunny hour; the sixth and the seventh entry stands for the temperature (^{o}F) and wind speed (mph) respectively, e.g. to the first historical hour in the matrix, the temperature is 41^{o}F and the wind speed is 10.4 mph.

- *Predicted entire traffic*

  Predicted entire traffic is a vector, e.g. (210, 300, 100), each entry of which is the predicted entire traffic of a future hour, e.g. the entire traffic in next hour is 210.

➢ **Output**

Predicted check-out across clusters

It is a matrix. Each row corresponds to a future hour and each column corresponds to a cluster. The format is the same with that of historical check-out across clusters, e.g.

$\begin{pmatrix} 120 & 40 & 10 & 30 \\ 80 & 30 & 50 & 40 \\ 68 & 54 & 38 & 40 \end{pmatrix}$, showing the predicted check-out across clusters for the future 3 hours.

➢ **Test Case**

We use the data from 1st, Apr. to 30th, Sep. in New York City as the test case. Check-out across clusters is stored in variable “m_o”, entire traffic is stored in “m_O”, check-out proportion across clusters is stored in variable “rho” and features are stored in variable “fea”. All these are statistics results, which are stored in “fea_rho.mat”.

Check-in Inference

The codes in “Check-in Inference” folder aim to infer the check-in across clusters.

❖ Trip Duration

These codes aim to learn the trip duration between each pair of clusters.

Run “tD_M.m” in “Codes\Check-in Inference”.

➢ **Input**

- *Trip Duration Set*

  It is a matrix of sets, e.g.

$$\begin{pmatrix} set_{11} & set_{12} & set_{13} & set_{14} & set_{15} \\ set_{21} & set_{22} & set_{23} & set_{24} & set_{25} \\ set_{31} & set_{32} & set_{33} & set_{34} & set_{35} \\ set_{41} & set_{42} & set_{43} & set_{44} & set_{45} \\ set_{51} & set_{52} & set_{53} & set_{54} & set_{55} \end{pmatrix}$$

Each entry corresponds to a set of trip duration between a pair of clusters, e.g. $set_{12} = \{266, 1321, 673, 1063, 951, 749, 297\}$ means that the trip duration from cluster 1 to cluster 2 has the following historical values, 266 seconds, 1321 seconds, 673 seconds, 1063 seconds, 951 seconds, 749 seconds and 297 seconds.

➢ **Output**

Trip Duration matrix

It is a matrix, each entry of which corresponds to a pair of parameters, e.g.

$$\begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} & p_{15} \\ p_{21} & p_{22} & p_{23} & p_{24} & p_{25} \\ p_{31} & p_{32} & p_{33} & p_{34} & p_{35} \\ p_{41} & p_{42} & p_{43} & p_{44} & p_{45} \\ p_{51} & p_{52} & p_{53} & p_{54} & p_{55} \end{pmatrix}.$$

Each entry, $p_{ij} = (u_{ij}, sigma_{ij})$ corresponds to the two parameters of a lognormal distribution, which describes the trip duration distribution from cluster i to cluster j.

➢ **Test Case**

We use the data from 1st, Jul. to 31th, Aug. in New York City as the test case. Trip Duration Set is stored as variable “tD” in “tD.mat”.

❖ Check-in

For bikes borrowed before

Run “runAll_b.m” in “Codes\Check-in Inference”

➢ **Input**

- *Bike usage information*

It is a matrix storing when and where a bike is checked out and when and where the bike is checked in. Each row stands for a bike usage record and the four columns mean the start time, stop time, start station and stop station respectively, e.g.

$$\begin{pmatrix} 10000 & 10047 & 29 & 25 \\ 10018 & 10025 & 14 & 25 \\ 40831 & 40846 & 85 & 26 \\ 161954 & 162012 & 25 & 50 \\ 261005 & 261049 & 35 & 57 \end{pmatrix}$$

for Sep. There are 5 records. The first record tells that a bike was checked out from station 29 at 00:00am on 1st, Sep and was checked in to station 25 at 00:47am on 1st, Sep; the second record tells that a bike was checked out from station 14 at

00:18am on 1st, Sep. and was checked in to station 25 at 00:25am, on 1st, Sep.; the forth record tells that a bike was checked out from station 25 at 19:54pm on 16th, Sep. and was checked in to station 50 at 20:12pm on 16th, Sep., etc.

- *Predicted inter-cluster transition matrices*
  It is a set of matrix. Every hour that need to be predicted has an inter-cluster transition matrix.
- *Trip duration matrix*

➢ **Output**

The number of bikes, which were borrowed before, will be returned to each cluster in the predicted period.

➢ **Test Case**

We predict the check-in of each cluster, which comes from the "not returned bikes", from 11th, Sep. to 30th, Sep. Bike Usage Information is stored as variable "R" in "R.mat". Predicted inter-cluster transition matrices are stored as variable "T_C"in "T_C.mat". Trip Duration Matrix is stored as variable "dM" in "dM.mat".

<u>For bikes will be borrowed in the future</u>

Run "runAll_f.m" in "Codes\Check-in Inference"

➢ **Input**

- *Predicted check-out*
  It is a matrix, e.g.
  $\begin{pmatrix} 7 & 6 & 5 & 7 & 9 \\ 4 & 3 & 3 & 4 & 4 \\ 5 & 10 & 6 & 3 & 4 \end{pmatrix}$. It is the predicted check-out of each cluster for future 3 hours. Each row stands for a predicted hour and each column stands for a cluster. For the first predicted hour, the predicted check-out of each cluster is 7, 6, 5, 7, 9 respectively; for the second hour, the predicted check-out of each cluster is 4, 3, 3, 4, 4 respectively, etc.
- *Predicted inter-cluster transition matrices*
- *Trip duration matrix*

➢ **Output**

The number of bikes, which will be borrowed in the predicted period, will be returned to each cluster in the predicted period.

➢ **Test Case**

We predict the check-in of each cluster, which come from the "will be borrowed bikes", from 11th, Sep. to 30th, Sep. Predicted check-out is stored as varibal "p_o" in "p_o.mat". Predicted inter-cluster transition matrices are stored as variable "T_C"in "T_C.mat". Trip Duration Matrix is stored as variable "dM" in "dM.mat".

Contact

Dr. Yu Zheng, yuzheng@microsoft.com

Lead Researcher at Microsoft Research

http://research.microsoft.com/en-us/people/yuzheng/