KDD Cup 2017

M10515031 黃佳郁
M10515036 謝奇元
M10515104 羅煜賢

指導教授:李漢銘 教授

Problem Description

TASK 1:

Travel Time Prediction

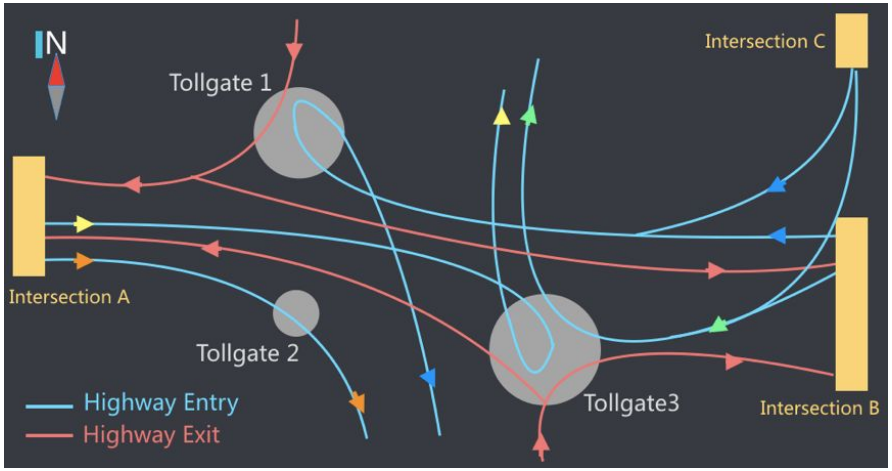
For every 20-minute time window, estimate the average travel time of each route.

A. Intersection A - Tollgates 2 & 3
B. Intersection B - Tollgates 1 & 3
C. Intersection C - Tollgates 1 & 3

TASK 2:

Traffic Volume Prediction

For every 20-minute time window, predict the entry and exit traffic volumes at tollgates 1, 2 and 3.

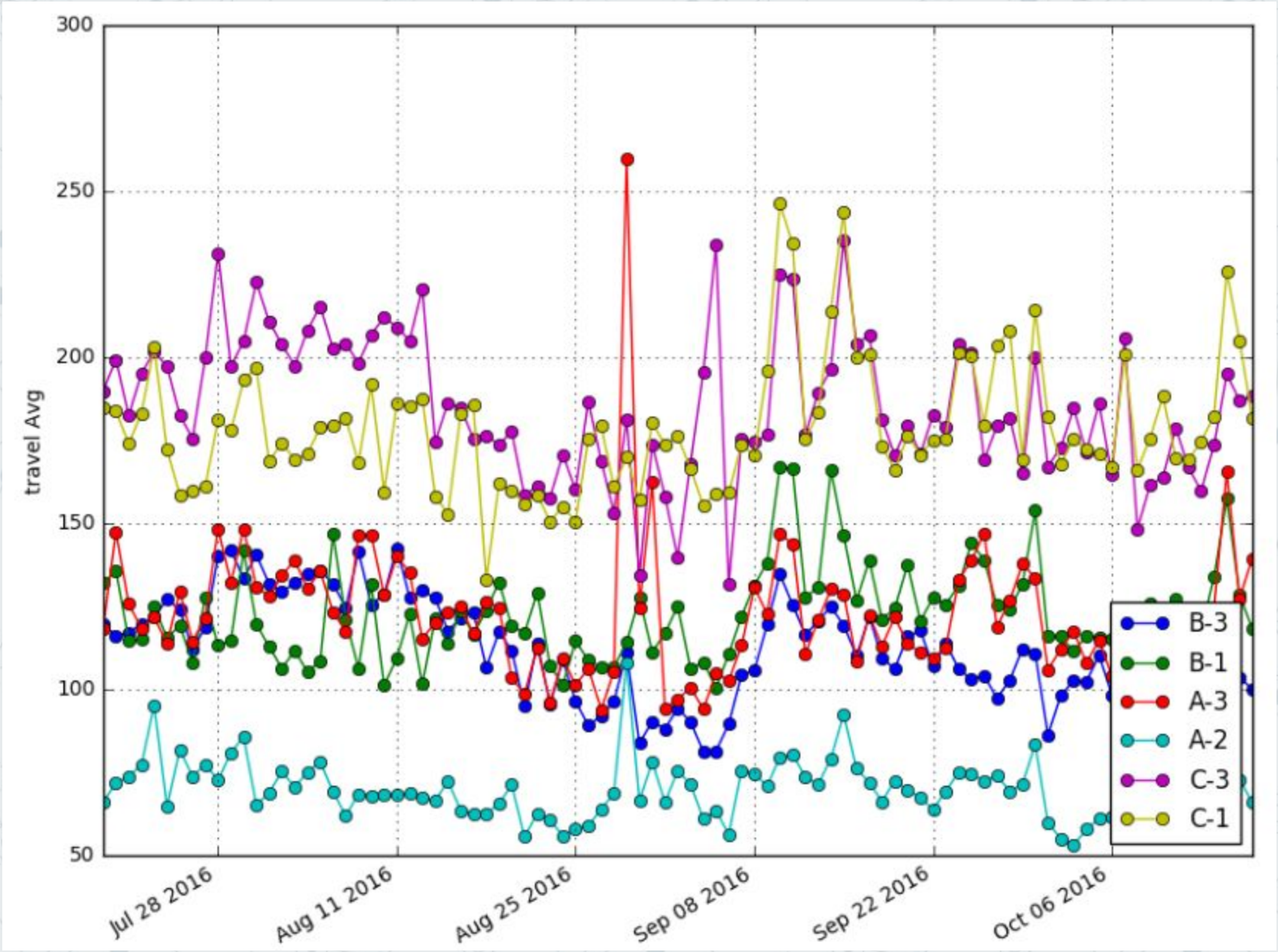


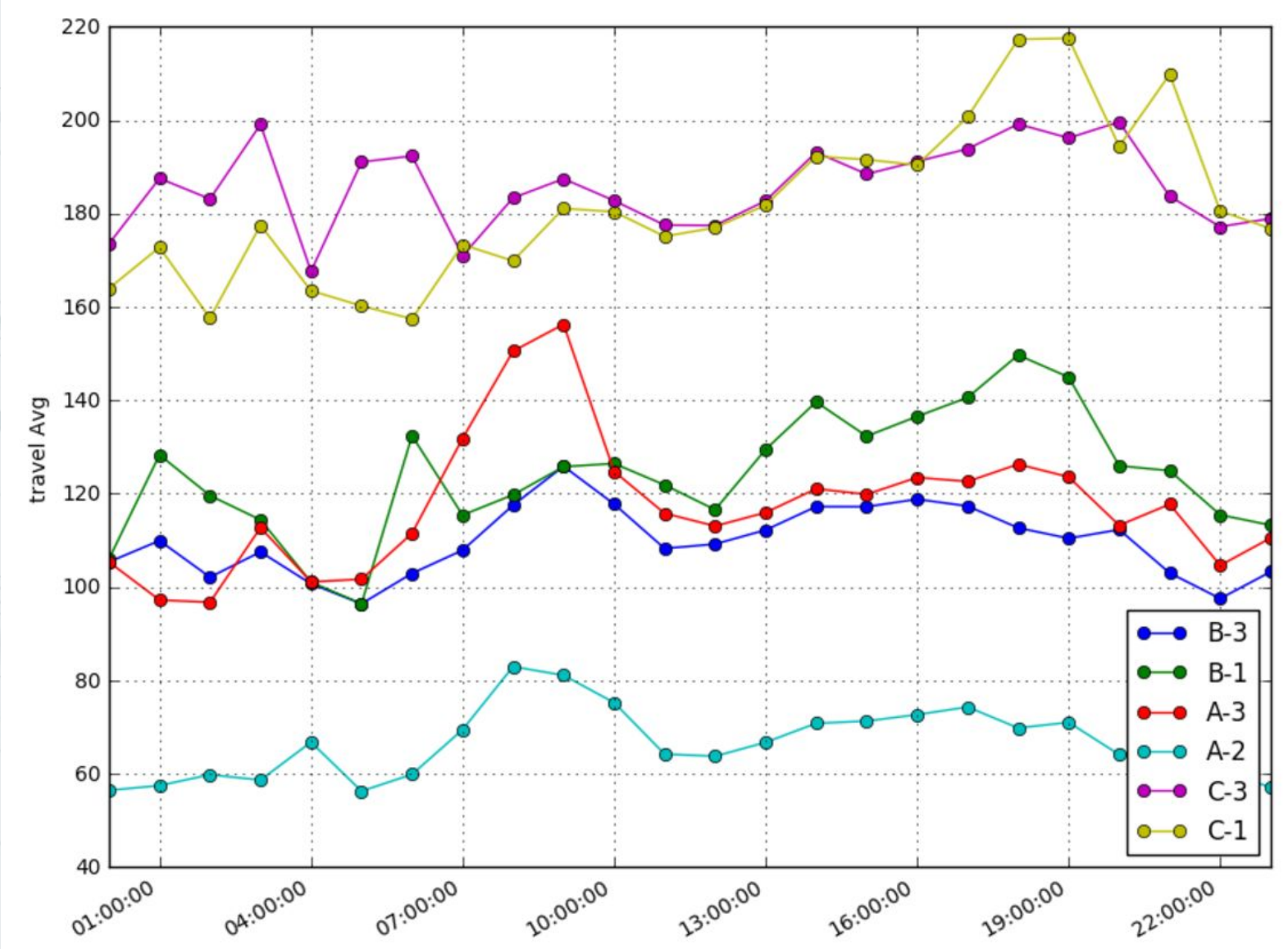
Competition website : [KDD Cup 2017]

Grade


Volume Prediction 261 / 0.2539
Travel Time Prediction
Volume Prediction
时间
MAPE
当天排名
2017-04-11 01:10:29
0.2539
73
2017-04-09 17:30:04
0.3923
86
2017-04-07 14:15:05
5.5468
84
2017-03-12 23:44:07
无
无

How To Do?


300
250
200
150
100
50
travel Avg
Jul 28 2016
Aug 11 2016
Aug 25 2016
Sep 08 2016
Sep 22 2016
Oct 06 2016
B-3
B-1
A-3
A-2
C-3
C-1


travel Avg
220
200
180
160
140
120
100
80
60
40
01:00:00
04:00:00
07:00:00
10:00:00
13:00:00
16:00:00
19:00:00
22:00:00
B-3
B-1
A-3
A-2
C-3
C-1

Linear Regression

Task 1

- What day?
- Start point time
- Spending time of each link of this route
- Total travel time

Field	*Type*	*Description*
intersection_id	string	intersection ID
tollgate_id	string	tollgate ID
vehicle_id	string	vehicle ID
starting_time	datetime	time point when the vehicle enters the route
travel_seq	string	trajectory in the form of a sequence of link traces separated by ";", each trace consists of link id, enter time, and travel time in seconds, separated by "#"
travel_time	float	the total time (in seconds) that the vehicle takes to travel from the intersection to the tollgate

Table 5

starting_time	travel_seq	travel_time
2016/7/19 00:14	105#2016-	70.85
2016/7/19 00:35	105#2016-	148.79
2016/7/19 00:37	105#2016-	79.76
2016/7/19 00:37	110#2016-	58.05
2016/7/19 00:56	105#2016-	137.98
2016/7/19 00:56	115#2016-	113.54
2016/7/19 01:26	105#2016-	176.7
2016/7/19 01:36	110#2016-	74.47
2016/7/19 01:36	110#2016-	94.57
2016/7/19 01:36	115#2016-	214.87

TASK 1

```
Get six files which belongs to six routes(A-2, A-3, B-1, B-3, C-1, C-3)
Address data type(travel_seq, starting_time)
For i in six files:
  Ignore the files without enough info
  Label starting_time
  Regress with Linear Regression
  Fit the model
  Predict travel time
```

Problem - Travel Time

A-2

mean	70.123898
std	45.561928
min	9.260000
25%	44.980000
50%	58.660000
75%	82.715000
max	1569.640000

A-3

mean	123.824527
std	83.335008
min	19.790000
25%	88.860000
50%	107.710000
75%	137.210000
max	6711.110000

B-1

mean	128.078528
std	57.578811
min	19.460000
25%	96.290000
50%	117.850000
75%	144.510000
max	1627.380000

B-3

mean	113.412535
std	53.858812
min	11.740000
25%	78.700000
50%	106.315000
75%	137.942500
max	1498.970000

C-1

mean	184.307117
std	73.699985
min	38.500000
25%	142.140000
50%	171.455000
75%	210.382500
max	2489.570000

C-3

mean	187.242564
std	72.014020
min	32.040000
25%	142.830000
50%	176.200000
75%	217.170000
max	1260.760000

Linear Regression

Task 2

- What day?
- Time range
  - One unit / 20 min
  - 8-10, 17-19
- Volume of two previous time range
  - if 8:00-8:20 => 7:20-7:40, 7:40-8:00
- Average volume of the same time range(8-10 or 17-19) of that day

Field	*Type*	*Description*
time	datetime	the time when a vehicle passes the tollgate
tollgate_id	string	ID of the tollgate
direction	string	0: entry, 1: exit
vehicle_model	int	this number ranges from 0 to 7, which indicates the capacity of the vehicle (bigger the higher)
has_etc	string	does the vehicle use ETC (Electronic Toll Collection) device? 0: No, 1: Yes
vehicle_type	string	vehicle type: 0-passenger vehicle, 1-cargo vehicle

Table 6

time
2016/9/19 23:09
2016/9/19 23:11
2016/9/19 23:13
2016/9/19 23:17
2016/9/19 23:16
2016/9/19 23:18
2016/9/19 23:18
2016/9/19 23:19
2016/9/19 23:19

Day	startTime	x1	x2	h	y
1	1	71	103	127	118
1	7	90	115	90	86
1	2	103	118	127	168
1	8	115	86	90	85
1	3	118	168	127	161
1	9	86	85	90	91
1	4	168	161	127	145

TASK 2

```
1  Label date with 1-7 which means Mon. to Sun.
2  Label time with 1-12(1: 8:00-8:20; 7: 17:00-17:20)
3  Use sqlQuery to count volume of each time range(every 20 mins)
4  Get average volume of 8-10 and 17-19 of each day
5  Create training data with day label, time, volumes of two previous time
•  range, average volume of the same time range it belongs to(8-10 or 17-19)
6  Regress with Linear Regression
7  Fit the model
8  Predict volume
```

Tools

VM

OS: Ubuntu

RAM: 3GB

Anaconda

iPython Notebook

MySQL

SQL Query

sklearn

Future Work

- Address noise data
- Try to give weight to each feature
- Predict with more features, e.g. temperature, humidity, width of each link
- Try to train and predict with xgboost
  - In some competition, this was used by first place participant
- Test deep learning module, e.g. Keras
- To improve the efficiency, we expect to use cloud computing technique

Thanks!