竞赛题目

本次大赛要求选手根据广州市内及广佛同城公交线路的历史公交刷卡数据，挖掘固定人群在公共交通中的行为模式。建立公交线路乘车人次预测模型，并用模型预测未来一周（20150101-20150107）每日06时至21时每小时段各个线路的乘车人次。Part2将更换一批新数据。

大赛开放20140801至20141231五个月广东部分公交线路岭南通用户刷卡数据，共涉及近200万用户2条线路约800多万条数据记录。同时大赛提供20140801至20150131期间广州市的天气状况信息。

数据说明

乘车刷卡交易数据表（gd_train_data）

列名	类型	说明	示例
Use_city	String	使用地	广州
Line_name	String	线路名称	线路1
Terminal_id	String	刷卡终端ID	4589bb610f9be53a43a7bc26bb40e44d
Card_id	String	卡片ID	8ce79e0b647053f191d20c5552eb49f0
Create_city	String	发卡地	佛山
Deal_time	String	交易时间(yyyymmddhh)	2014091008
Card_type	String	卡类型	学生卡

数据库导入（表名：gd_train_data）样例：

use_city	line_name	terminal_id	card_id	create_city	deal_time	card_type
广州	线路6	d1b5feda077818e286aefe5b1cbc75c6	7e017a9d22141cc1b536d	广州	2014122908	普通卡
广州	线路6	4fe4beb30f21df1b16b0769a0ceac473	f983f75ad1146366dea7b	广州	2014122909	普通卡
广州	线路6	4fe4beb30f21df1b16b0769a0ceac473	7615005a78528a5651437	广州	2014122909	普通卡
广州	线路11	78da304535276e90f941f644028848c2	e7921ba6f9fba313c2326c	广州	2014122813	普通卡
广州	线路11	78da304535276e90f941f644028848c2	040e0386b1ee6f8b7f03f1	广州	2014122813	普通卡
广州	线路6	d6d5806e8ea3db1b7ce2d61d995bb8da	393591054157b0a725e28	广州	2014122906	老人卡
广州	线路11	562689c04e6feda9314312569ec1ba05	e3c4ec853a3d968e73984	广州	2014122920	普通卡
广州	线路11	a54c4fef8e8af04c188abadb618fc230	83d1402668c9768284f08	广州	2014122910	老人卡
广州	线路6	98852b28c4d7eebcb078186f815e97ff	daefee665536be4aeac37	广州	2014122907	普通卡
广州	线路6	98852b28c4d7eebcb078186f815e97ff	f4af1b89e920ca7996b6d	广州	2014122907	普通卡
广州	线路6	98852b28c4d7eebcb078186f815e97ff	3cb9600393757d1b7f075	广州	2014122907	普通卡
广州	线路6	86c04ce314b74bcb8dfdc937c1ee8b83	f1e2bd15bec45182529a4	广州	2014122813	普通卡
广州	线路11	f3819ff946bd5657ca8e1a0f9a1161a7	4eb8a77f5b6a998e6a979	佛山	2014122912	普通卡
广州	线路6	f3819ff946bd5657ca8e1a0f9a1161a7	b27bedd005b292021efc2	广州	2014122912	普通卡
广州	线路11	f3819ff946bd5657ca8e1a0f9a1161a7	840b0c32c78a6957d2958	广州	2014122912	老人卡
广州	线路11	df99e8a997ca95f714c24e1350e269a4	240ad795e7103272a937b	广州	2014122917	普通卡
广州	线路11	fcd8d53f9a5994fc18a759eb2c5d485f	faf77aed4903226308a53e	广州	2014122907	普通卡

公交线路信息表（gd_line_desc）

列名	类型	说明	示例
Line_name	String	线路名称	线路1
Stop_cnt	String	线路站点数量	24
Line_type	String	线路类型	广州市内/广州佛山跨区域

数据库导入（表名：gd_line_desc）样例：

line_name	stop_cnt	line_type
线路1	18	广州市内
线路2	26	广佛跨区域
线路3	22	广州市内
线路4	31	广佛跨区域
线路5	21	广州市内
线路6	14	广州市内
线路7	27	广佛跨区域
线路8	34	广佛跨区域
线路9	30	广州市内
线路10	30	广佛跨区域
线路11	31	广州市内
线路12	17	广佛跨区域
线路13	12	广州市内
线路14	23	广州市内
线路15	37	广佛跨区域
线路16	24	广州市内
线路17	13	广州市内
线路18	10	广州市内
线路19	34	广州市内
线路20	50	广州市内
线路21	22	广州市内

广州市天气状况信息（gd_weather_report）

列名	类型	说明	示例
Date_time	String	日期	2014/8/1
Weather	String	天气状况（白天/夜间）	小雨

Temperature	String	气温（最高/最低）	36℃/26℃
Wind_direction_force	String	风向风力（白天/夜间）	无持续风向≤3级/无持续风向≤3级

数据库导入（表名：gd_weather_report）样例：

date_time	weather	temperature	wind_direction_force
2014/8/1	晴/雷阵雨	36℃/26℃	无持续风向≤3级/无持续风向≤3级
2014/8/2	雷阵雨/雷阵雨	35℃/26℃	无持续风向≤3级/无持续风向≤3级
2014/8/3	雷阵雨/雷阵雨	35℃/25℃	无持续风向≤3级/无持续风向≤3级
2014/8/4	多云/多云	34℃/26℃	无持续风向≤3级/无持续风向≤3级
2014/8/5	雷阵雨/多云	34℃/26℃	无持续风向≤3级/无持续风向≤3级
2014/8/6	雷阵雨/雷阵雨	34℃/26℃	无持续风向≤3级/无持续风向≤3级
2014/8/7	雷阵雨/雷阵雨	32℃/26℃	无持续风向≤3级/无持续风向≤3级
2014/8/8	雷阵雨/雷阵雨	34℃/25℃	无持续风向≤3级/无持续风向≤3级
2014/8/9	雷阵雨/雷阵雨	34℃/25℃	无持续风向≤3级/无持续风向≤3级
2014/8/10	雷阵雨/雷阵雨	33℃/26℃	无持续风向≤3级/无持续风向≤3级
2014/8/11	雷阵雨/雷阵雨	33℃/26℃	无持续风向≤3级/无持续风向≤3级
2014/8/12	雷阵雨/雷阵雨	33℃/25℃	无持续风向≤3级/无持续风向≤3级
2014/8/13	大雨/中到大雨	30℃/24℃	无持续风向≤3级/无持续风向≤3级
2014/8/14	中到大雨/雷阵雨	31℃/25℃	无持续风向≤3级/无持续风向≤3级
2014/8/15	多云/多云	33℃/25℃	无持续风向≤3级/无持续风向≤3级
2014/8/16	多云/多云	34℃/25℃	无持续风向≤3级/无持续风向≤3级
2014/8/17	多云/多云	34℃/25℃	无持续风向≤3级/无持续风向≤3级
2014/8/18	多云/雷阵雨	34℃/25℃	无持续风向≤3级/无持续风向≤3级
2014/8/19	大雨/大到暴雨	32℃/24℃	无持续风向≤3级/无持续风向≤3级
2014/8/20	中到大雨/雷阵雨	30℃/25℃	无持续风向≤3级/无持续风向≤3级
2014/8/21	雷阵雨/雷阵雨	31℃/25℃	无持续风向≤3级/无持续风向≤3级
2014/8/22	多云/多云	33℃/24℃	无持续风向≤3级/无持续风向≤3级

预测数据集为这些公交线路在20150101-20150107每个线路每日06时至21时各个小时段的乘车人次总和。（注：21时指的是21:00-21:59这个时间段）

选手需要提交结果表（gd_predict.txt）

列名	类型	说明	示例
Line_name	string	线路名称	线路1
Deal _date	string	日期	20150101
Deal_hour	string	小时段	08
Passenger_count	bigint	乘车人次	1234

提交文件示例

文件需用UTF-8字符编码；提交的文件内容格式如下，或参见文件sample_for_offline.txt。

```
线路10,20150104,07,0
线路10,20150104,10,799
线路10,20150104,13,834
线路10,20150104,16,106
```

评估指标

评估指标的设计主要期望选手对未来一周（20150101-20150107）每天06时至21时每个小时段各个线路乘车人次的总量数据预测的越准越好，积分公式的计算方法：计算每天每个小时段各个线路预测值的相对误差，然后根据用户预测乘车人次的相对误差，通过得分函数映射得到每个预测记录的得分，最后将所有预测记录得分求和除以理想状况的满分，得到最终评分。

具体的评分步骤如下：

1）计算各个线路每天每个小时段在测试集中乘车人次的预测偏差。

预测偏差 $\text{deviation}_i = \frac{|\text{count}_t - \text{count}_p|}{\text{count}_t}$

注：count_t为真实乘车人次，count_p为预测乘车人次

2）每个预测记录得分由预测偏差决定，偏差与得分之间的计算公式F（deviation_i）不公布，但保证该计算公式为单调递减的，即偏差越小，得分越高，偏差越大，得分越低。当deviation_i为0时，得分为 10 分；当deviation_i> 0.3 ，其得分为 0 。

3）最终得分precision = $\sum_i^N F(\text{deviation}_i)/(10 * N)$

明确问题：回归预测

1.数据下载和导入

2.预处理数据

（1）

INSERT INTO gd_train_data_count_number

SELECT line_name,create_city,deal_time,card_type,COUNT(*) FROM gd_train_data GROUP BY

deal_time,line_name,create_city,card_type

line_name	create_city	deal_time	card_type	count_number
线路11	佛山	2014080100	普通卡	3
线路11	广州	2014080100	员工卡	1
线路11	广州	2014080100	学生卡	5
线路11	广州	2014080100	普通卡	98
线路11	广州	2014080100	老人卡	1
线路11	汕尾	2014080100	普通卡	2
线路6	佛山	2014080100	普通卡	4
线路6	广州	2014080100	员工卡	1
线路6	广州	2014080100	学生卡	6
线路6	广州	2014080100	普通卡	71
线路6	广州	2014080100	老人卡	1
线路6	汕尾	2014080100	普通卡	1
线路11	佛山	2014080101	普通卡	1
线路11	广州	2014080101	普通卡	23
线路6	佛山	2014080101	普通卡	1
线路6	广州	2014080101	普通卡	6
线路11	佛山	2014080104	普通卡	2
线路11	广州	2014080104	员工卡	1
线路11	广州	2014080104	学生卡	1
线路11	广州	2014080104	普通卡	32
线路11	广州	2014080104	老人卡	3
线路6	佛山	2014080104	普通卡	2

（2）

INSERT INTO **gd_train_data_all_count_number**

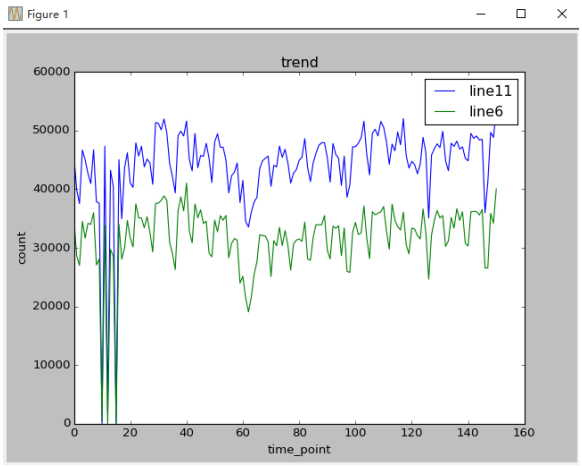
SELECT line_name,deal_time,SUM(count_number) FROM gd_train_data_count_number

GROUP BY line_name,deal_time

line_name	deal_time	all_count
线路11	2014080100	110
线路11	2014080101	24
线路11	2014080104	39
线路11	2014080105	161
线路11	2014080106	1639
线路11	2014080107	4108
线路11	2014080108	5189
线路11	2014080109	3799
线路11	2014080110	2872
线路11	2014080111	2216
线路11	2014080112	1722
线路11	2014080113	1899
线路11	2014080114	2070
线路11	2014080115	2272
线路11	2014080116	2434
线路11	2014080117	3842
线路11	2014080118	3954
线路11	2014080119	2430
线路11	2014080120	2156
线路11	2014080121	1956
线路11	2014080122	876
线路11	2014080123	316

3.数据观察

观察异常值，去掉异常点，分析可能影响结果的因素

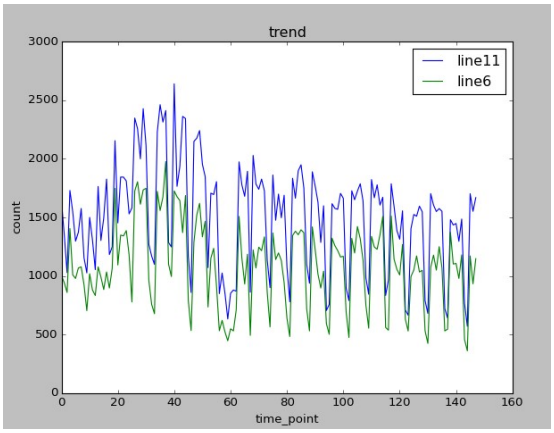


2015-09-26:35540
2015-09-27:28282
2015-09-28:30750
2015-09-29:31579
2015-09-30:31297
2015-10-01:23974
2015-10-02:25143
2015-10-03:21458
2015-10-04:19080
2015-10-05:21563
2015-10-06:25448
2015-10-07:27621
2015-10-08:32222
2015-10-09:32074
2015-10-10:32019
2015-10-11:30992
2015-10-12:25099
2015-10-13:31223
2015-10-14:30332
2015-10-15:33480
2015-10-16:30383
2015-10-17:32929
2015-10-18:30423
2015-10-19:26182
2015-10-20:30652
2015-10-21:31318
2015-10-22:31507

去噪音：去掉8月13、15、18（8月11日和12日数据不存在），经观察，分析十一假期客流量，得出可能会影响最后的预测结果（拉低结果值），所以去掉10月1、3、4、5号的数据

验证每天的数据中6点21点（程序验证）：没有空缺的数据

我们也可以从其他角度比如某个时间点来观察这些数据，比如，观察8月1日到12月31日6点的数据走向



2015-09-24:35482
2015-09-25:34692
2015-09-26:35540
2015-09-27:28282
2015-09-28:30750
2015-09-29:31579
2015-09-30:31297
2015-10-01:23974
2015-10-02:25143
2015-10-03:21458
2015-10-04:19080
2015-10-05:21563
2015-10-06:25448
2015-10-07:27621
2015-10-08:32222
2015-10-09:32074
2015-10-10:32019
2015-10-11:30992
2015-10-12:25099
2015-10-13:31223
2015-10-14:30332

4.特征选取

较为简单的方式：

day_1	day_2	day_3	day_4	day_5	day_6	day_7	day_8	day_9	day_10	day_11	day_12	day_13	day_14

day_15	day_16	day_17	day_18	day_19	day_20	day_21	day_22	day_23	day_24	day_25	day_26	day_27	day_28

day_29	day_30	day_31	hour_6	hour_7	hour_8	hour_9	hour_10	hour_11	hour_12	hour_13	hour_14	hour_15

hour_16	hour_17	hour_18	hour_19	hour_20	hour_21	线路6	线路11	stop_cnt	line_type_0	line_type_1	星期一	星期二	星期三	星期四	星期五	星期六	星期日

雨_daytime	雨_night	晴_daytime	晴_night	多云_daytime	多云_night	阴_daytime	阴_night	微寒_max	微寒_min	凉_max	凉_min	温凉_max	温凉_min	温和_max	温和_min

微温和_max	微温和_min	温暖_max	温暖_min	暖_max	暖_min	热_max	热_min	炎热_max	炎热_min	暑热_max	暑热_min	酷热_max	酷热_min	奇热_max

奇热	极热	极热	wind_direction_force_daytime	wind_direction_force_night	count

_min	_max	_min	wind_direction_force_daytime	wind_direction_force_night	count

实际上效果比较好的：

day_1	day_2	day_3	day_4	day_5	day_6	day_7	day_8	day_9	day_10	day_11	day_12	day_13	day_14

day_15	day_16	day_17	day_18	day_19	day_20	day_21	day_22	day_23	day_24	day_25	day_26	day_27	day_28

day_29	day_30	day_31	hour_6	hour_7	hour_8	hour_9	hour_10	hour_11	hour_12	hour_13	hour_14	hour_15	hour_16

hour_17	hour_18	hour_19	hour_20	hour_21	is_holiday	not_holiday	线路2	线路14	线路9	线路4	线路11	线路19	线路6	线路15

线路12	线路17	线路18	线路20	线路8	线路1	线路5	线路13	线路7	线路21	线路16	线路3	线路10	stop_cnt

line_type_0	line_type_1	星期一	星期二	星期三	星期四	星期五	星期六	星期日	holiday_start	holiday_mid	holiday_end	none

大雨_daytime	大雨_night	中雨_daytime	中雨_night	雨_daytime	雨_night	晴_daytime	晴_night	多云_daytime	多云_night	阴_daytime	阴_night	temperature_max	temperature_min	wind_direction_force_daytime	wind_direction_force_night	count

5.特征处理与映射

离散化：尽量将所有的特征离散，这样达到的效果是最好的，因为离散化后的特征更能够表现数据的特性，现有模型对于连续值的预测效果不如离散化后的特征（实践证明）

从gd_train_data_count_number表中筛选数据，构建特征矩阵，比如

line_name	deal_time	all_count
线路11	2014082012	1145

，

线路6	14	广州市内
线路7	27	广佛跨区域
线路8	34	广佛跨区域
线路9	30	广州市内
线路10	30	广佛跨区域
线路11	31	广州市内

，

2014/8/20	中到大雨/雷阵雨	30℃/25℃	无持续风向≤3级/无持续风向≤3级

，它的特征为：

day_1	day_2	day_3	day_4	day_5	day_6	day_7	day_8	day_9	day_10	day_11	day_12	day_13	day_14
0	0	0	0	0	0	0	0	0	0	0	0	0	0

day_15	day_16	day_17	day_18	day_19	day_20	day_21	day_22	day_23	day_24	day_25	day_26	day_27	day_28
0	0	0	0	0	1	0	0	0	0	0	0	0	0

day_29	day_30	day_31	hour_6	hour_7	hour_8	hour_9	hour_10	hour_11	hour_12	hour_13	hour_14	hour_15	hour_16
0	0	0	0	0	0	0	0	0	1	0	0	0	0

hour_17	hour_18	hour_19	hour_20	hour_21	is_holiday	not_holiday	线路2	线路14	线路9	线路4	线路11	线路19	线路6
0	0	0	0	0	0	1	0	0	0	0	1	0	0

线路15	线路12	线路17	线路18	线路20	线路8	线路1	线路5	线路13	线路7	线路21	线路16	线路3	线路10
0	0	0	0	0	0	0	0	0	0	0	0	0	0

stop_cnt	line_type_0	line_type_1	星期一	星期二	星期三	星期四	星期五	星期六	星期日	holiday_start	holiday_mid	holiday_end	none
31	0	1	0	0	1	0	0	0	0	0	0	0	1

大雨_daytime	大雨_night	中雨_daytime	中雨_night	雨_daytime	雨_night	晴_daytime	晴_night	多云_daytime	多云_night	阴_daytime	阴_night	temperature_max	temperature_min	wind_direction_force_daytime	wind_direction_force_night	count
1	0	0	0	0	1	0	0	0	0	0	0	30	25	0	0	1145

特殊处理：

```
def is_holiday(date_time):
    dt=datetime.strptime(date_time, "%Y/%m/%d")
    if(dt.month==10 and (dt.day in range(1,8))):
        return "is_holiday"
    if(dt.month==9 and (dt.day in range(6,9))):
        return "is_holiday"
    if(dt.month==1 and (dt.day in range(1,4))):
        return "is_holiday"
    if(dt.month==10 and dt.day==11):
        return "not_holiday"
    if(dt.month==1 and dt.day==4):
        return "not_holiday"
    if(0<=dt.weekday() and dt.weekday()<=4):
        return "not_holiday"
    if(5<=dt.weekday() and dt.weekday()<=6):
        return "is_holiday"
    return "not_holiday"
```

```
def is_holiday_start(date_time):
    dt=datetime.strptime(date_time, "%Y/%m/%d")
    if((dt.month==9 and dt.day ==6) or (dt.month==1 and dt.day ==1) or (dt.month==10 and dt.day ==2)):
        return "holiday_start"
    if((dt.month==9 and dt.day==7) or (dt.month==1 and dt.day ==2) or (dt.month==10 and dt.day ==6)):
        return "holiday_mid"
    if((dt.month==9 and dt.day==8) or (dt.month==1 and dt.day ==3) or (dt.month==10 and dt.day ==7)):
        return "holiday_end"
    return "none"
```

6.编程语言以及库使用

python和scikit-learn

7.模型选择

试验了多种模型，发现只有随机森林回归是最好的，后来，又发现了Bagging组合方法（2%-3%的提升），可以将若干个随机森林结合，提高预测准确率

8.程序编写

初始的数据量为11596035条记录，使用SQL脚本处理后，只剩下84287条记录，因此，可以使用基于内存的方式做，速度很快

（1）记录所有特征列名对应的index，并统计特征的个数

（2）取得所有天的天气信息为dict

（3）取得要预测的线路的信息（part1阶段是线路10和15，part2阶段是线路6和11）

（4）遍历**gd_train_data_all_count_number**的信息，将相应位置的特征列置1（默认为0），得到训练矩阵train_X和y

（5）train_X和y，训练模型得到一个model

（6）验证对原数据的拟合程度：将train_X再次带入model中，得到score_y，计算score_y和y的偏差（评分参考官方的评分：偏差小于0.3得1分，否则为0分）

（7）使用交叉评测观察model的得分，在0.93左右比较好

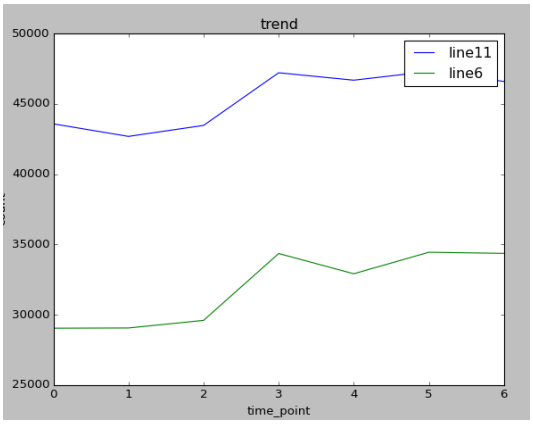
（7）生成训练数据，带入model进行预测，得出结果

（8）控制台打印输出的结果信息

```
自身评测 满分4608.得分4588
RandomForestRegressor：交叉评测得分0.940283690181
[  654.48  1705.93  2897.29 ...,  1860.76  1394.1   1273.73]
线路11,20150101,06,709
线路11,20150101,07,1937
线路11,20150101,08,3280
线路11,20150101,09,3787
线路11,20150101,10,3261
线路11,20150101,11,3118
线路11,20150101,12,2833
线路11,20150101,13,3086
线路11,20150101,14,3351
线路11,20150101,15,2891
线路11,20150101,16,3034
线路11,20150101,17,3141
线路11,20150101,18,2616
线路11,20150101,19,1994
```

9.预测结果分析

```
线路11信息：
2015-01-01:43581
2015-01-02:42689
2015-01-03:43470
2015-01-04:47217
2015-01-05:46688
2015-01-06:47309
2015-01-07:46598
线路6信息：
2015-01-01:29042
2015-01-02:29058
2015-01-03:29598
2015-01-04:34355
2015-01-05:32915
2015-01-06:34447
2015-01-07:34366
```



预测结果和中秋节、周六、周日、国庆进行比较，发现线路6在1月1日到1月3日的数量每天大约为29000左右合适