MLPractice-01 神舟优车订单预测（CSV）

讲义

1. 任务描述

在网约车业务中，由于系统中的专车资源是有限的，而乘客资源也是有限的，但是这两类有限的资源是各自散布在一个城市的若干区域中，并且由于城市的不同区域的功能不同，乘客资源散布在各个区域的密度有所差别，如何使专车的散布区域密度尽量符合乘客散布密度是其中很重要的一个需求。因此我们提出这样的需求，预测某个时间段内从区域A出发到区域B的订单数量。

现在给定2017年7月份和8月份的部分订单数据作为训练数据，预测8月份规定时间段内的数据。

2. 数据集

2.1. 训练集

我们以七月份的订单数据作为训练集，其中包含的各个字段的含义如下：

字段名称	字段含义	数据示例
id	订单id的hash值	583411b46a31bcc5d12d4402c928a146
driver_id	司机id的hash值（未出行成功则为一个特殊司机编号）	3e69e17a6e5a726fe44d71896bee4f32
menber_id	乘客id的hash值	6b4d6e4992191fe96b9f27921520d551
create_date	订单创建日期	2017-07-01

字段名称	字段含义	数据示例
create_hour	订单创建时间（0-23小时）	00
status	订单状态：0是未预约成功，1是预约取消，2是出行成功	2
estimate_money	预估行程金额（元）	140.00
estimate_distance	预估行程距离（米）	20099.00
estimate_term	预估行程时间（分钟）	18.00
start_geo_id	起点区域id的hash值	6d7827e8dcfa09497954a31e6f7e6ee6
end_geo_id	终点区域id的hash值	85e49ded1fa70a7bfa01ab0212a6e538

见附件：train_July.csv

2.2. 测试集

我们以八月份第一周内随机抽取的若干条数据作为测试集。测试集字段含义如下所示：

字段名称	字段含义	数据示例
test_id	该条测试用例的id	1
start_geo_id	起点区域id的hash值	6d7827e8dcfa09497954a31e6f7e6ee6
end_geo_id	终点区域id的hash值	85e49ded1fa70a7bfa01ab0212a6e538
create_date	订单创建日期	2017-08-01
create_hour	订单创建时间（0-23小时）	01

见附件：test_Aug.csv

2.3. 其他数据

2.3.1. 区域内设施数据

包含某一区域内公共设施的数量描述，其中第一列为区域id的hash值（与训练及测试数据中的区域对应），后面分别为各类设施的类型及数量，例如：

```
3d99665144344fc090b5b7450ffe72f5,加油站,4,超市,43,住宅区,151,地铁站,4,公交站,36,咖啡厅,22,中餐厅,597,ATM,59,写字楼,47,酒店,45
```

表示该区域内有加油站4个，超市43个，等等。

见附件：poi.csv

2.3.2. 天气数据

训练集与测试集所涉及的时间段内的天气情况（有极少部分缺失）。各个字段含义如下：

字段名称	字段含义	数据示例
date	日期（精确至半小时）	2017-7-10:30
text	天气现象文字	晴
code	天气现象代码	1
temperature	温度	29
feels_like	体感温度	28
pressure	气压	998
humidity	相对湿度	62
visibility	能见度	9.3
wind_direction	风向文字	南
wind_direction_degree	风向角度，范围0-360，0为正北，90为正东，180为正南，270为正西	200
wind_speed	风速	8
wind_scale	风力等级	2

见附件：weather.csv