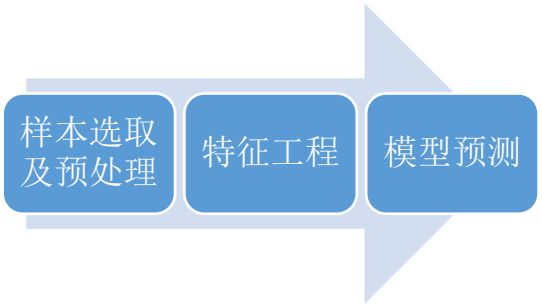
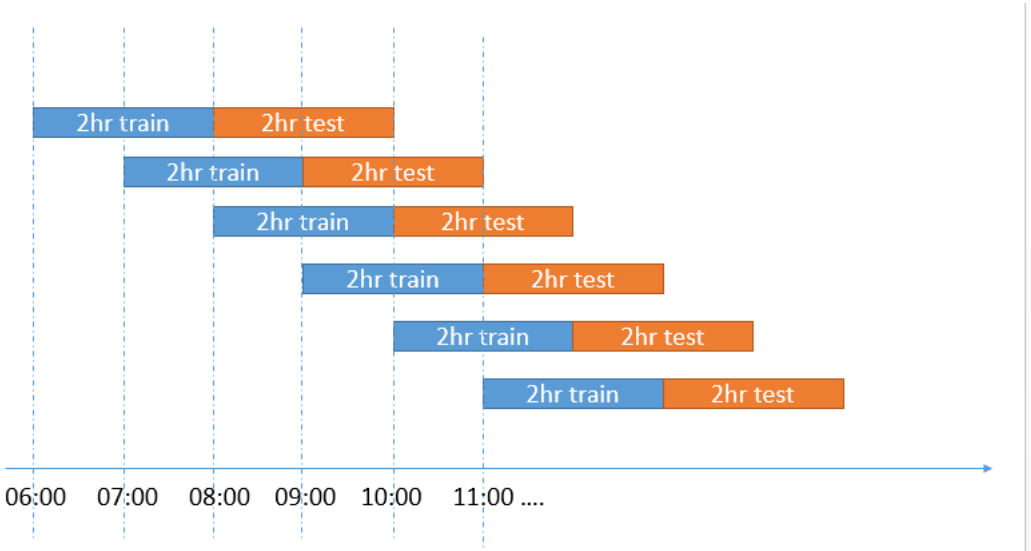
一.题意理解：
根据 2016 年 7~10 月的高速车流数据，预测 10 月 25 到 10 月 31 日指定时间的平均通行时间

二.解决框架




三.样本选取及预处理：
1.样本构造
a. 采用前 2 小时预测后 2 小时的方式，如 6-7 预测 8-9，7-8 预测 9-10

b .抽取白天的样本，即每天 6-20 点的数据.
则训练样本的时间从 6-7 预测 8-9，一直到 17-18 预测 19-20 为止




如上图，用 6-7 点的蓝色部分作为特征区间，预测 8-9 点时间段的平均时间。

2.样本选取及预处理：

a.全局按 travel_time 与均值之差大于 5 倍标准差的方式进行样本过滤

b.每 20min 再按 travel_time 与均值之差大于 5 倍标准差的方式进行第二层过滤

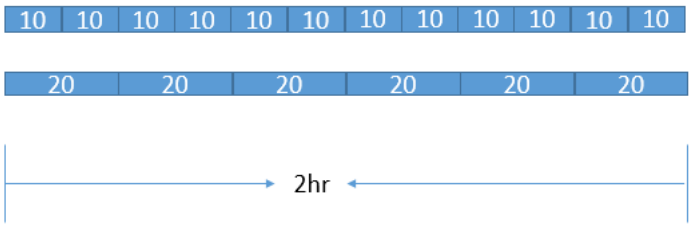
c.计算 label,计算每 20 分钟的 travel_time 均值，则每 2 小时有 6 个 label。选取 2 小时内含

有完整 6 个 label 的样本作为最终训练样本

四．特征工程：
主要分 5 部分：统计特征、路线特征、天气特征、各路段特征、以及其他　特征

1. 统计特征：
a. 统计每个样本特征空间内，每 10/15/20/30 分钟的 travel_time 均值（缺失值用中位数填充）

如




b. 由 a 所得，分别求不同窗口所得均值的统计特征，如均值、中位数、最大、最小、标准差、范围、25%分位数、75%分位数
c. 每 20/40/60/120 分钟经过的记录数作为车流量特征

2. 路线特征：
主要是每条路线的客观特征。
a. 所有路线总共有 25 个路段（link）,对每条路线，对其路段做 0-1 编码，共 25 维。即 1 代表该路线包含该路段，0 代表不包含。
b. 统计每条路线的路段总长、总面积（路段长*路段宽度之和）

3. 天气特征：
a. 原始天气值直接 merge
b. 对风速、温度、湿度等按实际天气标准进行离散化分级

4. 各路段特征：
由于轨迹表含有 travel_seq 这样的轨迹字段
115#2016-10-18　06:00:28#9.35;102#2016-10-18　06:00:38#13.54;109#2016-10-18 06:00:51#4.35;104#2016-10-18 06:00:55#24.96;112#2016-10-18 06:01:20#22.73;111#2016-10-18 06:01:43#16.16;103#2016-10-18 06:01:59#18.43;122#2016-10-18 06:02:20#27.53

于是将每条记录的 travel_seq 全部拆开，形成下图的表格：

| Link | time | Travel_time |
|---|---|---|
| 115 | 2016-10-18 06:00:28 | 9.35 |
| 102 | 2016-10-18 06:00:38 | 13.54 |
| 109 | 2016-10-18 06:00:51 | 4.35 |
| 104 | 2016-10-18 06:00:55 | 24.96 |
| … | … | … |

对组成的新表，可以看做是路段信息表,并构造如下特征：

a. 每小时 路段平均通行时间
b. 每小时 路段车流量/单位长度路段车流量/单位面积路段车流量
c. 每小时 路段车速

5. 其他特征：

a. 训练时间标识，共 6 维，如 100000 表示训练第一个 20 分钟，010000 表示训练第二个 20 分钟，依次类推
b. 时间相关标识：是否周末、是否节假日（国庆、中秋）、按小时分段、是否周末与时间分段的交叉，如是否周末早上、工作日中午等
c. 路段标识，主要是 intersction_id 与 tollgate_id 的交叉

五．模型训练：

线下：采用 train-test 的方式调参，选取线上测试前一周（18~24）的 8 点和 17 点数据作为线下验证集，18 号前的作为训练集

线上：全部数据

模型：Xgboost，模型参数如下

```
params = {
    'booster': 'gbtree',
    'objective': 'reg:linear',
    'stratified': True,
    'max_depth': 8,
    'min_child_weight': 1,
    'gamma': 1,
    'subsample': 0.7,
    'colsample_bytree': 0.7,
    'lambda': 1,
    'eta': 0.01,
    'seed': 20,
    'silent': 1
}
```

成绩提升较大的几个点：

1. 异常值过滤
2. 添加各路段特征
3. 根据 xgb 输出的 feature importance 进行特征选取，剔除 importance 大但对模型有反效果的特征