3月19日讨论会纪要

内容摘要

1. 探索一下基于1小时、2小时的统计情况（申）。
2. 补充了几个变量（花开）： a. 每条link,有前面的link后面的link,如果有intop,outtop有两个link,说明是交叉路口,所以补充特征,每条路线遇到的交叉路口数 b. 路途的路径宽度: 1车道2车道3车道4车道的link数，1车道2车道3车道4车道的长度 c. 是否过节,国庆节:'2016-10-01-2016-10-07',中秋节''2016-09-15-2016-09-17''
3. 是否将十一期间的数据清洗掉？（孔）
4. 查看以前的赛事，看看大概要做到什么程度，分享一些论文和主题（孔）

5. 时间特征刻画的方法（孔）：
   a. 按星期把训练集和测试集分为七个，周一到周日，每天共有3×20个特征，每天要预测的Y有12个。这12个Y都用上周这60个特征来预测，用前一天和上一周该星期的数据。
   周一 特征
   Y x1 x2 x3 x4 ...

   Y1(y1,y2,...)

   Y2(y1,y2,...)

   Y1是一个向量它共有十二个元素，每个元素代表周一要预测的一个时间窗口，y1,y2,...都用同样的特征值
   Y1是9月26号，Y2是10月3号...
   x1是9月19号的第一个时间窗口也就是0点到0点20分，x2是后一个，以此类推，共3×20个特征

用前一天和上一周该星期的数据。

周一	特征						
Y	x1	x2	x3	x4	…		
Y1(y1, y2, …)							
Y2(y1, y2, …)							

6. 模型（孔）：

   a. 1.时间序列的处理方法有ARIMA，指数平滑，季节性指数平滑，缺点是只用到车流信息，用不上天气等其他信息

   b. 把每一天的数据抽出来，做一个训练集，得出参数，然后把测试集中绿色那段数据和我们的模型进行匹配，找到最匹配的那一天的模型，然后用该模型预测红色那段数据

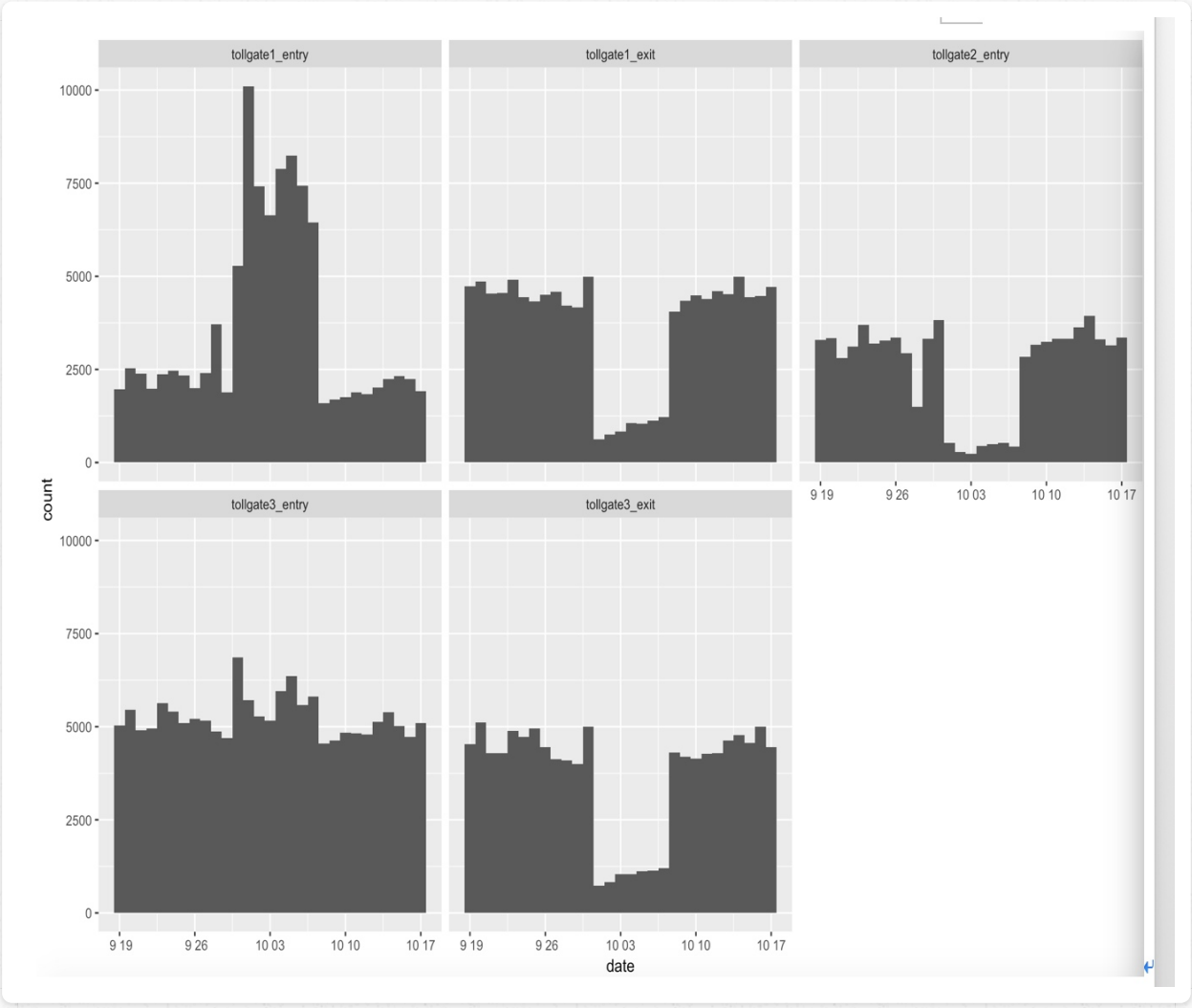
7. 模型（申）：

   a. 我尽量讲的通俗一点。我能想到的一个最合适的模型，是首先用深度构造一个网络，最后一步用一个回归模型，可以用简单的线性回归

   b. 模型的最后一步，可以用general linear regression，也可以用孔提到的集成学习中的回归方法，因为他之前看历届的大赛，很多问题用集成学习效果不错

   c. 输入变量我考虑分三块，一块是不随时间，天气影响的固定因素，比如道路的长短，道路的结构等等，一块是要求时间段前时间段的特性，另一块是前一天，前一周同时间段；我的意思是，比如要计算2号的10点，我取2号九点的天气，风速等具体特征，但1号的10点，我只取算好了的交通量，至于1号天气咋样就不管了

8. 明显的假期模式（free）；



9. 关于流量预测的看法（free)

a. 对五个”收费口-方向“分别建立训练集进行训练。

会议共识

1. 需要提取出每条记录是周几；
2. 需要提取出每条记录的时间段（几点几分到几点几分之间），窗口20分钟？
3. 需要提取出每个route是否有岔路
4. 需要提取出道路的宽度、长度；
5. 需要提取出每条记录是否假期；
6. 可采用集成学习；
7. 变量暂时选取：星期、时间段、是否假期、湿度、降雨、风速、收费口、方向、路径长度、路径宽度、vehicle_model、岔路口。
8. 对于时间刻画，分这样几个层次：天，每小时，每20分钟，同时连接到周几。
9. 变量不需要归一化
10. 先搞任务2

下阶段任务

1. 提取记录是周几
2. 提取记录的时间段
3. 提取记录是否为假期
4. 提取每个route是否有岔路
5. 提取道路长度、宽度；
6. 确定每个变量的特征化格式（比如，周几用1-7数字表示，是否假期用0-1表示）