一、特征优化处理

1. 将天气情况按是否下雨结构化加入运算
2. 在特征中加入急诊人次作为基准
3. 加入特征的变化因素

第三点的意思是例如气温的变化，空气质量的变化。因为回归模型中每条训练数据之间并无关系，但是我们的回归模型中，每条数据之间可能是有相关关系的，目前的回归模型可能出现同样的的气温，同样的环境质量却有不同的急诊人次分布。出现这样的原因和气温变化可能有关。

举个例子，假设1号气温是1度，2号气温是16度。若我们预测窗口为0，则是通过2号16度气温，预测2号的急诊人次，即16度预测一个急诊人次。1号的气温1度到2号气温16度是升温，即气温变化为15度。

又假设1号气温为31度，2号气温为16度。同样的预测窗口为0，2号气温同样是16度，真实数据中却可能得出不一样的急诊人次，是因为这是降温。气温变化为-15度。因此需要将各个特征因素的变化也作为特征加入运算，目前首先尝试当天和前一天的因素变化（例如气温变化）作为特征加入

如下表所示：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | X1 | X2 | X3 | X4 | y |
| date |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | X1 | ΔX1 | X2 | ΔX2 | X3 | ΔX3 | X4 | ΔX4 | Y |
| date |  |  |  |  |  |  |  |  |  |

二、结局变量处理

通过诊断史（气候会影响的病种），进一步优化结局变量，得到一个有周期且有规律的解决变量。

1. 数据源优化处理
2. 区间切片：例如每年每个季度的模型提取出来，然后将每年每个季度的数据拼接在一起，假如按春夏秋冬四季切片，则每个季度应该有相应的预测模型。如何切片需要寻找规律。
3. 离群点去除，目前箱线图可标注出离群点但是没有去除代码。对于离群点去除的代码网上大部分都是通过聚类（K—means）去进行，我认为不太靠谱。