2021.11.18

主要内容：GEFCom2014 第一名的 GAM + 分位数回归方法进行概率预测，以及在 intermittent demand 上的实现（M5数据）；

已经做了 WSS（基于历史销量数据模拟需求与需求发生概率）与 VZ （基于历史数据模拟需求与需求间隔时间）两种基于历史数据的方法，重复模拟1000次后去产生0.01-0.99分位数；并尝试使用GAM+分位数回归的方法进行概率预测。

分位数回归概率预测

来自于 ***Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting*** (Gaillard et al., 2016) (International Journal of Forecasting)，介绍了在该比赛的 load 和 price 赛道获得第一名的 Tololo 队的结果。这里仅关注将 GAM 与分位数回归联合使用的方法（也是他们的主方法）。这个方法他们称作 **quantGAM。**

方法介绍

广义加性模型（GAM）可以写成下式，g 表示均值 μ 的连接函数，右侧诸 f 为协变量的平滑函数。在比赛中，连接函数使用的是**恒等变换**（Identity），协变量选取的函数为三次样条。具体操作他们称在 mgcv 包完成。

$$g\left(\mu(\boldsymbol{X}_t)\right) = f_1(X_{t,1}) + f_2(X_{t,2}) + f_3(X_{t,3}, X_{t,4}) + \cdots,$$

（鉴于电力负荷与电价或可看成连续变量，使用恒等变换是简单合理的。）

比赛中 **quantGAM** 的算法步骤：

1. **拟合GAM**：又可分为两部分：

   a. 第一步是拟合均值，估计均值的各个可加成分影响 f_i，则诸成分之和为均值的估计。

   b. 第二步是拟合方差（可选项），将残差项的平方与各个协变量再做一次 GAM，可以获得与残差项有关的诸可加成分 g_i。

2. **分位数回归**：做因变量与各个计算好的成分 f_i , g_i 的线性分位数回归，0.01-0.99。因为 GAM 可以保证因变量与成分之间是线性关系，这样最优化也比较方便。（相当于用协变量构造的成分来代替自身去进行估计）

论文中，对于比赛的实际场景，这里以中期、负荷预测为例展示预测过程。其使用协变量为：该日在一年中的比例，时间序号 t，该日温度，该日类型（周一，周二-周四，周五，周六，周日，公共假日当天，公共假日前后几天）。没有使用负荷的滞后变量，时间带来的周期性通过协变量来引入。两式表明需要的均值和方差成分，之后用这些成分进行分位数估计。

$$\begin{aligned} Y_t &= f_1(Toy_t) + f_2(t) + f_3(T_t) \\ &\quad + h(DayType_t) + \varepsilon_t, \end{aligned}$$

$$\left(Y_t - \widehat{Y}_t\right)^2 = g_1(Toy_t) + g_2(T_t) + \varepsilon_t,$$

而对于气温，操作中不能提前知道，故还需要提前预测气温——气温的0.01-0.99分位数。气温的估计式如下：

$$T_t = f_1(Toy_t) + \varepsilon_t,$$

$$\left(T_t - \widehat{f_1}\left(Toy_t\right)\right)^2 = g_1(Toy_t) + \varepsilon_t,$$

在负荷的分位数估计中，每次估计使用气温的一个分位数，最终对不同气温值的分位数估计求平均值以获得最终负荷的分位数估计。

在M5数据的尝试

这里仅是一个尝试，目的是找到可行方法并代码实现，故变量选用有待斟酌。

还是仿造如上方式，对每条时间序列各自生成GAM模型，提取成分，进行分位数回归，估计概率。取最后28天作为测试数据。

考虑的协变量：该日在全年的比例、该日在星期的比例（周六-周五，按照M5的表格来）、事件（event_name1，event_type1）、SNAP（每个州一个，共3个）、价格（原始价格、该日价格相对于同类商品最高价的比例、该日价格相对于历史最高价的比例（这个变量最终放弃了））

变量的问题：（线性分位数回归要求不能出现完全共线性，然而M5的协变量却经常会产生这种问题；事实上，M5提供的协变量几乎没有连续且取值丰富的变量）

- 关于事件变量的说明：（其实有name2与type2，但是出现事件次数5年仅4次，且预测过程中由此引发了矩阵奇异的问题，故这次实现暂时放弃）（更好的方式应该是每个事件一个哑变量，不过一共30+事件，感觉代码有点麻烦。。。就没用；但估计时应该和使用event_name1引入了相同的哑变量）；此外，使用 event_name1 变量，在数据长度较短时，会出现测试数据出现的事件训练数据没有的问题。。。
- 关于价格变量的说明：M5商品的价格变化次数很少，有的甚至不变价格；对于价格单调变化的商品，相对于历史最高价的比例会带来完全共线性，而这种现象过于常见，就弃用；而对于价格不变的商品，也会带来完全共线性，故对于这种时间序列需要放弃该变量。
- 没有引入滞后的销量：一个是确实有些参赛者没有引入滞后项，另外引入滞后的销量对于预测阶段也会产生麻烦（短的滞后期需要用预测去生成预测，需要进行多次蒙特卡罗实验；之后将测试的参数方法也会面临这个问题）

预测过程：

首先，对销量与协变量做GAM，但是鉴于销量是计数数据，故先尝试连接函数为泊松回归的连接（mgcv好像不能直接支持负二项甚至tweedie的连接函数？），可以得到各协变量的成分（只考虑均值，不考虑残差）。事实上，只有"该日在全年比例"可以用三次样条，其它变量由于重复值过少只能直接使用。

接下来是大问题：**所提取的成分与销量均值的连接函数线性关联，故不能直接做成分与销量的线性分位数回归**！

（尝试非线性分位数回归，需求不对，也没搞懂）；计数数据分位数回归?

最终借鉴 ***Quantiles for Counts*** （Machado & Silva, 2005) (Journal of the American Statistical Association) 的做法得以解决：

1. 对因变量 Y 加一个 U(0,1) 的随机数，记作 Z，使其连续；
2. 对 Z 做如下变换，使其可以与协变量有一定线性关系：

$$T(Z;\alpha) = \begin{cases} \log(Z-\alpha) & \text{for } Z > \alpha \\ \log(\varsigma) & \text{for } Z \le \alpha \end{cases}$$

3. 对 T 与协变量做线性分位数回归；
4. 将所得分位数做逆变换得到 Z 的分位数，按照如下关系得到 Y 的分位数：

$$Q_Y(\alpha|\mathbf{x}) = \lceil Q_Z(\alpha|\mathbf{x}) - 1 \rceil$$

这个估计被证明是一致的，但为了更稳定的结果，可以重复多次取平均；但个人感觉对大多数分位数影响不大，对间断销量数据而言，可能会在靠近1的分位数有些不同；由于目前还在调试，故这部分只做了一次。

由于变量、数据的原因，程序在有些时间序列上没完全跑通（占约5%），而且变量、具体的操作过程还需完善，暂不展示预测结果。

个人想法

- 无论是基于历史数据的概率预测，还是分位数回归，在间断数据上的结果基本上接近于分布列（基于历史数据的方法，在界点上的分位数可能出现小数，但大多数都是0、1、2……这样持续一段概率值的分布列）。（一开始认为分位数回归会出现连续的结果，如0.01分位数为0.01，0.05分位数为0.08这样的结果）
- 对于今后要做的组合问题，在间断数据上就是离散分布的组合。
- 之前提到的基于给定参数分布的概率预测也会去做，包括GAM、基于一定随机过程的模拟（如Snyder et al. （2012）的方法）以及DeepAR的尝试
- 对概率预测结果的观察（PIT 观察calibration、DRPS 等 sharpness 指标，以及不同预测方法的相关性）
- 果然写代码就会出现各种各样问题。。。