分位数回归及其应用

Quantile Regression & its Applications

刘勤波

2022 年 4 月 9 日

分位数回归介绍

给定样本空间$\{(x_i, y_i)\}_{i=1}^N$, $\hat{y}_i$ 表示我们对y_i的估计值, 并定义样本点的损失函数为

$$\rho_{\tau,i} = \begin{cases} \tau(y_i - \hat{y}_i), & y_i \geq \hat{y}_i \\ (\tau - 1)(y_i - \hat{y}_i), & y_i < \hat{y}_i \end{cases}$$

其中 τ 记为分位数,则对应的损失函数可以表示为

$$Q = \sum_{i=1}^{N} \rho_{\tau,i} = \sum_{y_i \geq \hat{y}_i} \tau(y_i - \hat{y}_i) + \sum_{y_i < \hat{y}_i} (\tau - 1)(y_i - \hat{y}_i)$$

若使用连续随机变量表示，设y的分布函数为$F(y)$，估计值为$\hat{y}$，则对应的损失函数为:
$G(\hat{y}) = (\tau - 1)\int_{-\infty}^{\hat{y}}(y - \hat{y})dF(y) + \tau\int_{\hat{y}}^{\infty}(y - \hat{y})dF(y)$,
其中$F(\infty) = 1$，$F(-\infty) = 0$。

分位数回归介绍

求解损失函数的导数$G'(\hat{y})$。

- 引理1(微积分基本定理)：
  $F(b)-F(a)=\int_a^b F'(x)dx=\int_a^b dF(x)$。
- 引理2(含参数不定积分求导)：假设$g'(t)$是$[a,b]$上的连续函数，$f(t,x)$及其偏导数$f_x(t,x)$都是闭区域$[a,b]\times[c,d]$上的连续函数，且$\alpha(x)$，$\beta(x)$是$[c,d]$上的可微函数，并满足$a\leq\alpha(x),\beta(x)\leq b$，则函数

$$F(x)=\int_{\alpha(x)}^{\beta(x)} f(t,x)dg(t)$$

在$[c,d]$上可微，且有 $F'(x)=$

$$\int_{\alpha(x)}^{\beta(x)} f_x(t,x)dg(t)+f(\beta(x),x)g'(x)\beta'(x)-f(\alpha(x),x)g'(x)\alpha'(x)$$

分位数回归介绍

$G(\hat{y}) = (\tau - 1)\int_{-\infty}^{\hat{y}}(y - \hat{y})dF(y) + \tau\int_{\hat{y}}^{\infty}(y - \hat{y})dF(y)$

求解损失函数的导数$G'(\hat{y})$:

- 针对左半部分α取为常数， $\beta(\hat{y}) = \hat{y}$， $f(y, \hat{y}) = y - \hat{y}$，应用引理2，我们可以得到 $\frac{d}{d\hat{y}}\int_{-\infty}^{\hat{y}}(y - \hat{y})dF(y) =$

$$\int_{-\infty}^{\hat{y}} -1dF(y)+(\hat{y}-\hat{y})F'(\hat{y})(1)-0 = F(-\infty)-F(\hat{y}) = -F(\hat{y})$$

- 针对右半部分β取为常数， $\alpha(\hat{y}) = \hat{y}$， $f(y, \hat{y}) = y - \hat{y}$，应用引理2，我们可以得到 $\frac{d}{d\hat{y}}\int_{\hat{y}}^{\infty}(y - \hat{y})dF(y) =$

$$\int_{\hat{y}}^{\infty} -1dF(y)+0-(\hat{y}-\hat{y})F'(\hat{y})(1) = F(\hat{y})-F(\infty) = F(\hat{y})-1$$

因此，

$$G'(\hat{y}) = (\tau - 1)(-F(\hat{y})) + \tau(F(\hat{y}) - 1) = F(\hat{y}) - \tau$$

当$F(\hat{y}) = \tau$时， $G'(\hat{y}) = 0$，这也解释了为什么τ称为分位数。

项目实例:仓库推荐调拨服务

1. 仓库调拨现状分析
   - 1.1 订单满足率: 平均约70%
   - 1.2 调拨成本与配送成本:大区内RDC往FDC单位调拨成本普遍低于FDC单位配送成本
2. 算法优化目标
   - 2.1 提高订单满足率: $\geq$85%
   - 2.2 降低配送成本: 配送成本的降低大于调拨成本的提高
3. 算法方案
   - 3.1 SKU 需求量短期预测
   - 3.2 RDC-FDC调拨策略
4. 算法应用结果分析
   - 4.1 SKU 需求量预测准确率
   - 4.2 订单满足率提高以及对应的成本计算

分位数回归模型

- 为什么选取分位数回归模型
  - 1. 分位数回归适用于因变量具有异方差的情形。
  - 2. 分位数的引入可视为一种惩罚机制，用于平衡预测超出预期以及预测低于预期带来的影响，通常用于平衡持货成本与缺货成本。结合实际情况，我们采用分位数的目的主要是平衡基于预测调拨而产生的额外调拨成本与最优仓订单满足率。
  - 3. 分位数回归对于预测具有一定的可解释性，并且具有进一步提升的空间。

- 分位数回归模型的损失函数
  设y为真实值，其分布函数为$F(y)$，预测值为$\hat{y}$，分位数为τ，则对应的损失函数为：
  $G(\hat{y}) = (\tau - 1)\int_{-\infty}^{\hat{y}}(y - \hat{y})dF(y) + \tau\int_{\hat{y}}^{\infty}(y - \hat{y})dF(y)$。

令
(1) R表示区域RDC仓库;
(2) $F_i, i = 1, 2, 3, ...$表示区域内的FDC仓库;
(3) $j = 1, 2, 3, ...$表示商品SKU;
(4) $I_{R,j}, I_{F_i,j}$分别表示当前日期RDC=R和FDC= F_i对应SKU=j的库存量;
(5) $\hat{y}_{R,j}, \hat{y}_{F_i,j}$ 分别表示未来日期RDC=R和FDC= F_i覆盖区域内的SKU=j的需求量预测值;
(6) 目标订单满足率记为r (如$r = 85\%$);
(7) 决策变量为各个SKU从R到F_i的调拨量$Q_{F_i,j}$。

- 约束方程
  (a) $Q_{F_i,j} + I_{F_i,j} \geq r \cdot \hat{y}_{F_i,j},\ i = 1, 2, 3, ...,\ j = 1, 2, 3, ...;$
  (b) $I_{R,j} - \sum_{i=1,2,3,...} Q_{F_i,j} \geq r \cdot \hat{y}_{R,j},\ j = 1, 2, 3, ...;$
  (c) $Q_{F_i,j} \geq 0,\ i = 1, 2, 3, ...,\ j = 1, 2, 3, ...$。
- 目标函数约束方程有解时取调拨总数最小; 约束方程无解时取平均订单满足率最大值。

4. 算法应用结果分析

4.1 SKU 需求量预测准确率

适当选取分位数、预测时长、样本长度后模型大致可获得

- 3日预测平均准确率+80%。
- 预测模型拟合优度平均在0.18附近。

4. 算法应用结果分析

4.2 订单满足率提高以及对应的成本计算

以华中区为例，基于前述预测以及调拨模型:

- 最优仓订单满足率可提高至+90%。
- 调拨额外产生的每月费用约为2.5w。
- 配送成本每月可节省约10w。