VFPUGEN numerical column regression tree algorithm

jiuluan lv

July 2024

Algorithm 1 数值列采用回归树处理算法

输入: 训练数据集 D

输出: 回归树 $f(x)$

在训练数据集所在的输入空间中，递归地将每个区域划分为两个子区域并决定每个子区域上的输出值，构建二叉决策树:

(1) 选择最优切分变量 j 与切分点 s，求解

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2\right] \tag{1.1}$$

遍历变量 j，对固定的切分变量 j 扫描切分点 s，选择使式 (1.1) 达到最小值的对 (j,s)

(2) 用选定的对 (j,s) 划分区域并决定相应的输出值：

$$R_1(j,s)=\{x|x^{(j)}\le s\},R_2(j,s)=\{x|x^{(j)}>s\} \tag{1.2}$$

$$\hat{c}_m=\frac{1}{N_m}\sum_{x_i\in R_m(j,s)}y_i,x\in R_m,m=1,2 \tag{1.3}$$

(3) 继续对两个子区域调用步骤 (1)，(2)，直至满足停止条件

(4) 将输入空间划分为 M 个区域 $R_1,R_2,...,R_M$，生成决策树：

$$f(x)=\sum_{m=1}^{M}\hat{c}_mI(x\in R_m) \tag{1.4}$$
