Research Notes [Threshold regression]

zzz
Renmin University of China, Beijing
since August 7, 2014
Current Version: September 20, 2014

Contents

1 问题描述

对于回归树来说：给定训练集$D = (X_i, y_i)$,一个回归树对应着输入空间的一个划分以及在划分空间上的输出值。假设输入空间可以划分为M个类，$R_1, R_2, \cdots, R_M$.并且在每个类上有一个固定的输出值c_m。则模型可以写成$y = \sum_{m=1}^{M} c_m I(x \in R_m)$

对于threshold regression来说，假设真实的模型是：

$$y = \beta_0 + \beta_1 I(x_1 > t_1) + \beta_2 I(x_2 > t_2) + \epsilon$$

其可以写为

$$y = \beta_0 I(X_i \in R_1) + (\beta_0+\beta_1) I(X_i \in R_2) + (\beta_0+\beta_2) I(X_i \in R_3) + (\beta_0+\beta_1+\beta_2) I(X_i \in R_4)$$

其中R_i是x之前不同的排列组合组成的4类。当然这里还对系数之间有一些潜在的约束。模型可以视为4个类的输出，c_m相当于每一类的均值。

2 模拟结果

2.1 model1

$$y = 1 + 2I(x_1 > t_1) + 3I(x_2 > t_2) + c\epsilon$$

其中$t_1 = 0.5, t_2 = 0.2$

c=0.1时的决策树 模拟的$n = 1000$的数据的y的分布如下,相当于是从4类中生成出来的数据。

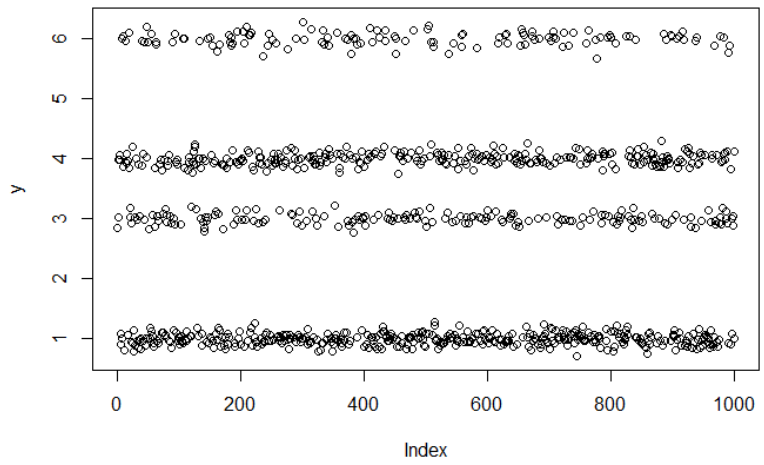

Figure 1: y的散点图,c=0.1

决策树的效果如下;

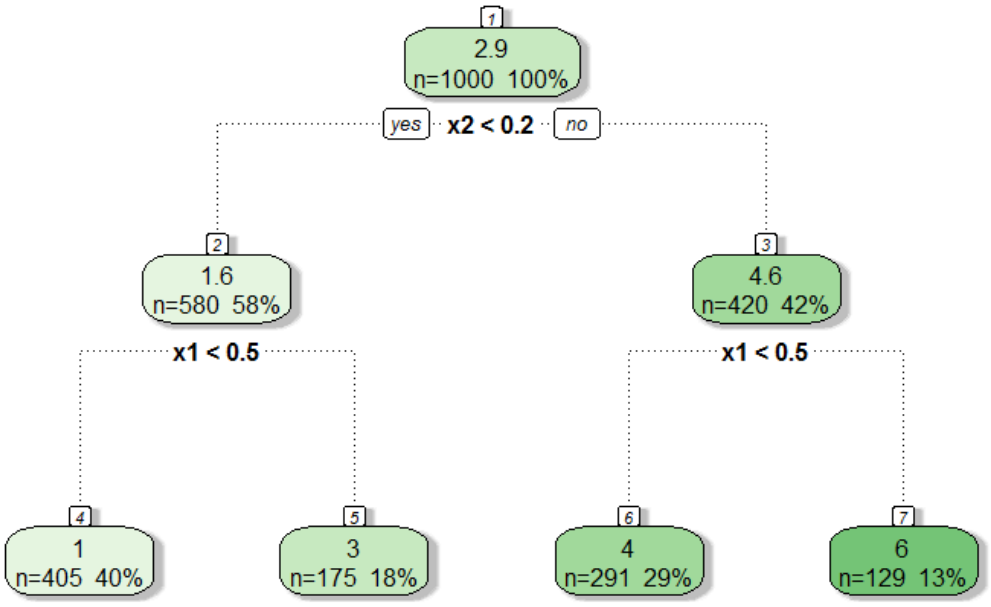

Figure 2: 噪声c=0.1

说明：

- 其中每个节点方框中的数代表当前单元下的估计的y的均值，即c_m
- 模型估计的叶子节点上的c_m精确的值是0.9998, 3.0003,3.9999 5.9996
- 与真实值1,3,4,6.即$\beta_0, \beta_0+\beta_1, \beta_0+\beta_2, \beta_0+\beta_1+\beta_2$ 几乎一致，估计的很准确。
- cut points的估计是$t_2 = 0.198 t_1 = 0.5, 0.5$

当逐渐增大噪声c,令c=1时，虽然从散点图上已经看不出数据的类，但是决策树效果仍然非常稳定。[在c较大时，出来的决策树变量有时会重复出现。这时可以给生成过程加入一些限制。比如maxdepth=p]

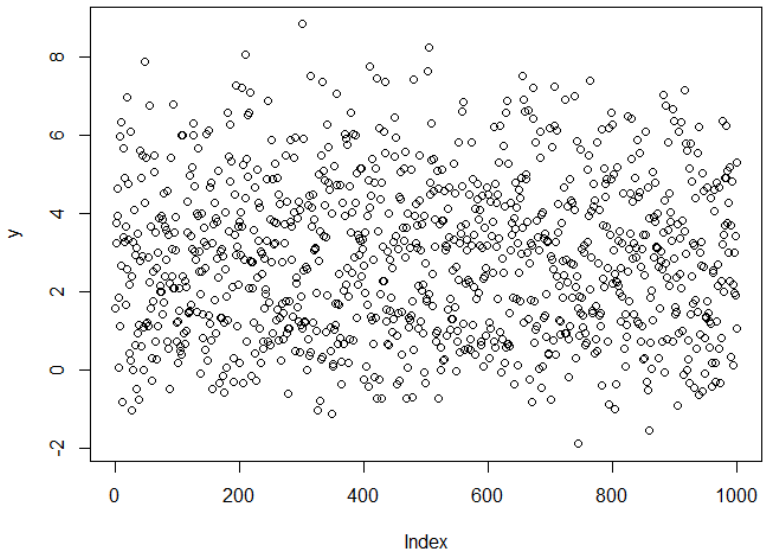

Figure 3: y的散点图,c=1

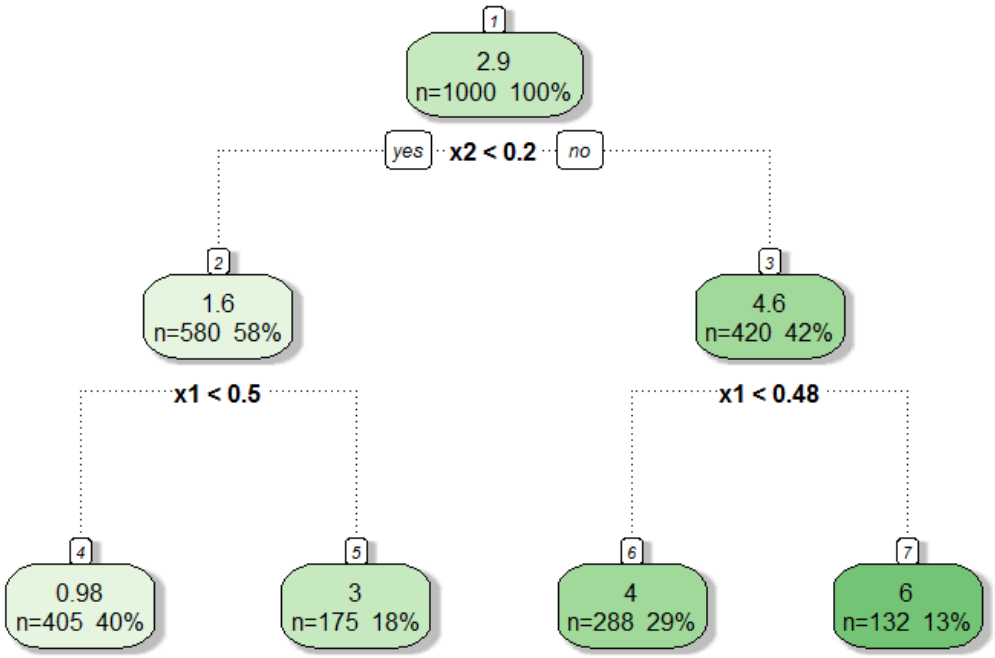

Figure 4: 噪声c=1

故意将噪声c加到5，决策树的结构如下

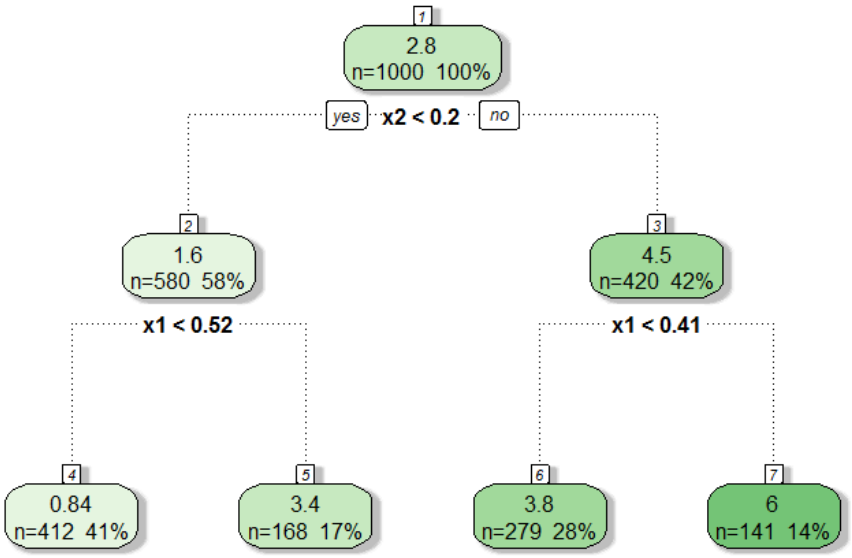

Figure 5: 噪声c=5

估计的cut points是$t_2 = 0.2$,$t_1 = 0.51$和0.42. c_m 是0.84，3.4，3.8，6. 与真实值1,3,4,6有一定误差，但是在噪音c=5的情况下，决策树仍然是很稳定的。

以上估计结果整理如下：

	估计的系数c_i	真实的系数c_i	估计的cut points	真实的cut points
p=2 噪音c=0.1	0.9998 3.0003 3.9999 5.9996	1 3 4 6	t2=0.1983821 t1=0.5001336 t1=0.5039988	t2=0.2 t1=0.5
p=2 噪音c=1	0.9763 3.0339 3.9809 5.9518	1 3 4 6	t2=0.1983821 t1=0.5001336 t1=0.4825968	t2=0.2 t1=0.5
p=2 噪音c=5	0.8436 3.3580 3.8321 5.9545	1 3 4 6	t2=0.1983821 t1=0.5232845 t1=0.412538	t2=0.2 t1=0.5

Figure 6: 估计结果

2.2 model2

假设有三个变量的：

$$y = 1 + 2I(x_1 > t1) + 3I(x_2 > t2) + 4I(x_3 > t3) + c\epsilon$$

其中$t_1 = 0.5, t_2 = 0.2, t_3 = 0$
效果仍然很好，这里只列出$c = 1$时候的一个例子

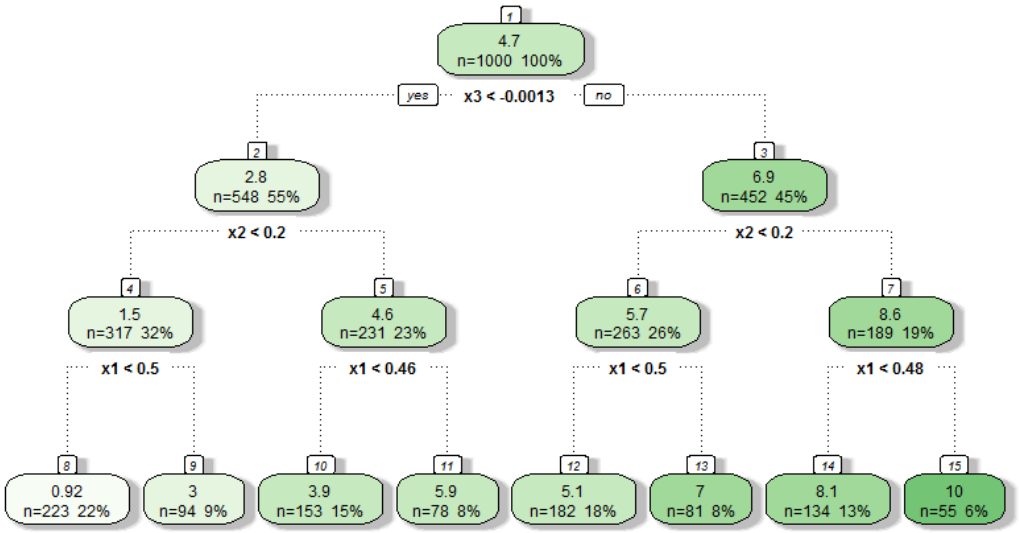

Figure 7: p=3,噪声c=1

从图中看，估计的cut point是$t_3 = -0.0013, t2 = 0.2, 0.2$,估计的$t_3$是0.5,0.46,0.5,0.48.估计的模型系数是0.92,3,3.9,5.9,5.1,7,8.1,10和真实的相差不大

3 比较

用cart回归树模型和原来的threshold regression问题还是有一些差别的。

- CART会涉及选变量，在高维的时候实际生成的树层次不是很深
- CART算法不会保证在不同的分叉下，同一个t_i是相同的，所以若用估计出来的c_i去反解β_j时，会出现过度识别的情况。
- CART算法是对二叉树的，不过也有别的理论方法去处理multi-way decision trees的情况