说明文档

Stephen CUI

2023 年 7 月 18 日

0.1 Data View

0.2 Data Wrangling

0.3 Feature Filter

Chapter 1

特征处理

1.1 χ^2分箱

1.1.1 χ^2的计算

$$\chi^2 = \sum_{i=i}^{2}\sum_{j=1}^{k}\frac{(A_{ij}-E_{ij})^2}{E_{ij}} \tag{1.1}$$

这里 k 是类的数量，A_{ij} 是 i 区间，j 类中的数量，$R_i = \sum_{j=1}^{k} A_{ij}$是 i 区间内的数量，$C_j = \sum_{i=1}^{2} A_{ij}$是 j 类的数量，$N = \sum_{i=1}^{2} R_i$ 是所有样本的数量，$E_{ij} = R_i * C_j/N$ 是 A_{ij} 期望频数。如果 R_i 或 C_j 为 0，则设置 $E_{ij} = 0.1$。

$$\begin{aligned} E_{ij} &= \begin{bmatrix} \frac{(a+b)(a+c)}{a+b+c+d} & \frac{(a+b)(b+d)}{a+b+c+d} \\ \frac{(c+d)(a+c)}{a+b+c+d} & \frac{(c+d)(b+d)}{a+b+c+d} \end{bmatrix} = \begin{bmatrix} \frac{(a+b)(a+c)}{N} & \frac{(a+b)(b+d)}{N} \\ \frac{(c+d)(a+c)}{N} & \frac{(c+d)(b+d)}{N} \end{bmatrix} \\ &= \frac{1}{N}\begin{bmatrix} a+b \\ c+d \end{bmatrix}\begin{bmatrix} a+c & b+d \end{bmatrix} = \frac{1}{N}\boldsymbol{RJ} \end{aligned} \tag{1.2}$$

这样代码中可以直接使用广播机制来计算 E_{ij}。

1.1.2 增量更新χ^2

如果分组特别的多，每次都对相邻的组做 χ^2 的计算，会耗费计算资源，可以考虑只用增量更新的方法来更新 χ^2 列表，Table 1.2。

表 1.1: 两个区间、两个类的 χ^2 合并示意图

	y_0	y_1	合计
x_0	$A_{11} = a$	$A_{12} = b$	$R_1 = a+b$
x_1	$A_{21} = c$	$A_{22} = d$	$R_2 = a+b$
合计	$C_1 = a+c$	$C_2 = b+d$	$N = a+b+c+d$

表 1.2: 增量更新 χ^2 列表

原分组	χ^2	一次合并后	合并后χ^2
1.1	0.1	1.1	0.1
2.2	0.3	2.2	0.3
3.2	0.4	3.2	changed
3.5	0.0	3.5,3.7	changed
3.7	0.4	5.5	0.6
5.5	0.6	6.7	0.12
6.7	0.12	7.1	0.11
7.1	0.11	7.8	0.12
7.8	0.12	7.9	0.16
7.9	0.16	10.0	
10.0			

1.1.3 停止条件

卡方分箱的停止条件有如下两种选择：

1. 分箱个数等于指定的分箱数目（`max_bins`）：限制最终的分箱个数结果，每次将样本中具有最小卡方值的区间与相邻的最小卡方区间进行合并，直到分箱个数达到限制条件为止。
2. 最小卡方值大于卡方阈值（`chi2_threshold`）：根据自由度和显著性水平得到对应的卡方阈值，如果分箱的各区间最小卡方值小于卡方阈值，则继续合并，直到最小卡方值超过设定阈值为止。

可以两个同时用，也可以只用一个。看实际需求调整即可。

阈值的意义

类别和属性独立时,有 90% 的可能性,计算得到的卡方值会小于 4.6。大于阈值 4.6 的卡方值就说明属性和类不是相互独立的，不能合并。如果阈值选的大,区间合并就会进行很多次，离散后的区间数量少、区间大。