说明文档

Stephen CUI

2023 年 7 月 17 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}

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	$C_1 = a+c$	$C_2 = b+d$	$N = a+b+c+d$