胡帆

2020000885

习题8

2. 将数据规范化到区间 [0, 1]

$$y_i = \frac{x_i - min(x)}{max(x) - min(x)} * (1 - 0)$$

得到结果

	身高	体重
1	0.091	0.222
2	0.227	0.267
3	0.000	0.000
4	0.545	0.444
5	0.591	0.556
6	0.364	0.378
7	0.682	0.333
8	0.909	1.000
9	0.727	0.556
10	1.000	0.667

3. 离散化属性年收入

3.1 分箱法

1. 等距离分箱法

年收入区间为 [10, 120]，则每个区间的间距 $I = 37$

选取 [10, 47), [47, 84), [84, 120) 三个区间进行分箱，分箱结果为三个表：

年龄	性别	年收入	婚姻	车型
25	男	10	单身	普通
32	男	20	离异	普通
27	女	25	单身	普通
30	男	30	单身	高级
35	女	30	离异	普通
28	男	40	已婚	中档

52	**男**	**50**	**已婚**	**中档**
45	女	60	单身	高级

55	**男**	**100**	**已婚**	**高级**
48	女	120	离异	高级

2. 等频率分箱法

   每个箱3个值，则3个区间为

   [10, 30), [30, 60), [60, 120]

年龄	性别	年收入	婚姻	车型
25	男	10	单身	普通
32	男	20	离异	普通
27	女	25	单身	普通

30	**男**	**30**	**单身**	**高级**
35	女	30	离异	普通
28	男	40	已婚	中档
52	男	50	已婚	中档

45	**女**	**60**	**单身**	**高级**
55	男	100	已婚	高级
48	女	120	离异	高级

3.2 基于熵的方法

首先对年收入取值进行升序排列：

年龄	性别	年收入	婚姻	车型
25	男	10	单身	普通
32	男	20	离异	普通
27	女	25	单身	普通
30	男	30	单身	高级
35	女	30	离异	普通
28	男	40	已婚	中档
52	男	50	已婚	中档
45	女	60	单身	高级
55	男	100	已婚	高级
48	女	120	离异	高级

信息熵计算公式：

$$entropy(D) = -\sum_{i=1}^{k} p(c_i) log_2 p(c_i)$$

一个数据集D按 $A \leq v$ 分裂前后信息熵的差值称为信息增益，记为 $gain(D, v)$

$$gain(D, v) = entropy(D) - entropy(D, v)$$

分裂前信息熵为 $entropy(D) = 1.5219$ ，分别按照"年收入 $\leq$ 25"，"年收入 $\leq$ 30"，"年收入 $\leq$ 50"三种情况分裂，其信息增益计算如下：

$$gain(D, 25) = entropy(D) - entropy(D,\ 25) = 1.5219 - 0.9651 = 0.5568$$
$$gain(D, 30) = entropy(D) - entropy(D,\ 30) = 1.5219 - 0.8464 = 0.6755$$
$$gain(D, 50) = entropy(D) - entropy(D,\ 50) = 1.5219 - 0.9651 = 0.5568$$

所以选择信息增益最大的 "年收入 $\leq$ 30" 进行分裂，其中
$entropy(D_1) = 0.3610$，$entropy(D_2) = 0.4855$ ，所以对 "年收入 $>$ 30" 继续分裂。

D' 为"年收入 $>$ 30"数据集，由于车型只在年收入为50时发生改变，所以以50为划分阈值。

综上，离散化的三个区间为：

[10, 30], (30, 50], (50, 120]

3.3 基于 ChiMerge 方法

首先将离散化属性 "年收入" 进行排序，然后以相邻两个值的中点为分界线：

年收入	分界线
10	
20	15
25	22.5
30	27.5
30	30
40	35
50	45
60	55
100	80
120	110

以 [0, 15) 和 [15, 22.5) 为例列出列联表：

	车型=普通	车型=中档	车型=高级	合计
[0, 15)	1	0	0	1
[15, 22.5)	1	0	0	1
合计	2	0	0	2

其卡方的计算公式如下：

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

其中：

$$E_{ij} = \frac{R_i C_j}{R_1 + R_2}$$

如果 $E_{ij} = 0$，则 $E_{ij} = 0.1$.

由列联表可以计算出卡方值 $\chi^2 = 0.2$，查卡方分布表，$\alpha = 0.1$ 时，$\beta = 2.706$, 由于 0.2<2.706，因此合并这两个区间为 [0, 22.5)。

同理，继续计算 [0, 22.5) 和 [22.5, 27.5) 卡方值 $\chi^2 = 0.2$ ，并合并为 [0, 27.5).

	车型=普通	车型=中档	车型=高级	合计
[0, 27.5)	3	0	0	3
[27.5, 35)	1	0	1	2
合计	4	0	1	5

对于 [27.5, 35)，由列联表可以计算出卡方值 $\chi^2 = 0.7$，查卡方分布表，$\alpha = 0.1$ 时，$\beta = 2.706$, 由于 0.7<2.706，因此合并这两个区间为 [0, 35)。

	车型=普通	车型=中档	车型=高级	合计
[0, 35)	4	0	1	5
[35, 45)	0	1	0	1
合计	4	1	1	6

对于 [35, 45)，由列联表可以计算出卡方值 $\chi^2 = 152.7$，查卡方分布表，$\alpha = 0.1$ 时，$\beta = 2.706$, 由于 152.7<2.706，因此不合并。

同理，[35, 45) 和 [45, 55)通过卡方检验，合并为 [35, 55).

[55, 80), [80, 110) 和 [110, 120] 通过卡方检验，合并为 [55, 120]。

综上，离散化的三个区间为：

[0, 35), [35, 55), [55, 120].