CSE 5243 Autumn’22 Assignment #2 p1-p4 – Intr Data Mining

Neng Shi

2022/09/19

1 Problem 1.

(a) mean: $\frac{200+400+600+800+1300+1500+2000}{7} = 971.43$

variance: $\frac{(200-971.43)^2+(400-971.43)^2+(600-971.43)^2+(800-971.43)^2+(1300-971.43)^2+(1500-971.43)^2+(2000-971.43)^2}{7} =$ 362040.82

(b) $v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$

$max_A = 2000$, $min_A = 200$

So after normalization, the new group of data is $0.00, 2.22, 4.44, 6.67, 12.22, 14.44, 20.00$.

(c) $v' = \frac{v - \mu_A}{\sigma_A}$

So 20 will be transformed to -1.58.

2 Problem 2.

(a)

	a	b	Sum (row)
c_1	0 (1.2)	2 (0.8)	2
c_2	4 (3)	1 (2)	5
c_3	2 (1.8)	1 (1.2)	3
Sum (col.)	6	4	10

(b)

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$$
$$=\frac{(0-1.2)^2}{1.2}+\frac{(2-0.8)^2}{0.8}+\frac{(4-3)^2}{3}+\frac{(1-2)^2}{2}+\frac{(2-1.8)^2}{1.8}+\frac{(1-1.2)^2}{1.2}$$
$$=3.89$$

3 Problem 3.

(a) the mean of X_1 is $\dfrac{8+0+6+5+2+(-2)}{6}=3.17.$

the mean of X_2 is $\dfrac{-20+(-1)+(-10)+(-15)+0+5}{6}=-6.83.$

(b) $E[X_1X_2]=\dfrac{(-160)+0+(-60)+(-75)+0+(-10)}{6}=3.17$

$\sigma_{12}=\sigma_{21}=E[X_1X_2]-E[X_1]E[X_2]=-50.84-3.17\times(-6.83)=-29.20$

$\sigma_{11}=E[X_1^2]-E[X_1]^2=12.14$

$\sigma_{12}=E[X_2^2]-E[X_2]^2=78.47$

So the covariance matrix is

$$\begin{bmatrix} 12.14 & -29.29 \\ -29.20 & 78.47 \end{bmatrix}$$

4 Problem 4.

(a) $Info(D)=I(6,4)=-\frac{2}{6}log_2(\frac{2}{6})-\frac{4}{6}log_2(\frac{4}{6})=0.92$

$Info_{Car}(D)=\frac{3}{6}I(3,1)+\frac{1}{6}I(1,1)+\frac{2}{6}I(2,2)=0.5\times 0.92=0.46$

So $Gain(Car)=Info(D)-Info_{Car}(D)=0.46.$

(b) There are two ways of splitting.

$Car1:\{Sports,Vintage\},\{SUV\},$

$Car2:\{Sports\},\{Vintage,SUV\},$

$Car3:\{Sports,SUV\},\{Vintage\}.$

$Info_{Car1}(D)=\frac{4}{6}I(4,2)+\frac{2}{6}I(2,2)=0.67$

$Info_{Car2}(D)=\frac{3}{6}I(3,1)+\frac{3}{6}I(3,3)=0.46$

$$Info_{Car3}(D) = \frac{5}{6}I(5,3) + \frac{1}{6}I(1,1) = \frac{5}{6} \times 0.97 = 0.81$$

So we would choose $Car2 : \{Sports\}, \{Vintage, SUV\}$ since it is with the maximum information gain.

(c) $SplitInfo_{Car}(D) = -\frac{3}{6}log(\frac{3}{6}) - \frac{1}{6}log(\frac{1}{6}) - \frac{2}{6}log(\frac{2}{6}) = 1.46$

$$gain\_ratio(Car) = \frac{0.92 - 0.46}{1.46} = 0.32$$

$$SplitInfo_{Car1}(D) = -\frac{4}{6}log(\frac{4}{6}) - \frac{2}{6}log(\frac{2}{6}) = 0.92$$

$$gain\_ratio(Car1) = \frac{0.92 - 0.67}{0.92} = 0.27$$

$$SplitInfo_{Car2}(D) = -\frac{3}{6}log(\frac{3}{6}) - \frac{3}{6}log(\frac{3}{6}) = 1.00$$

$$gain\_ratio(Car2) = \frac{0.92 - 0.46}{1.00} = 0.46$$

$$SplitInfo_{Car3}(D) = -\frac{5}{6}log(\frac{5}{6}) - \frac{1}{6}log(\frac{1}{6}) = 0.65$$

$$gain\_ratio(Car3) = \frac{0.92 - 0.81}{0.65} = 0.17$$

So we would choose $Car2 : \{Sports\}, \{Vintage, SUV\}$ since it is with the maximum gain ratio.

(d) For the continuous attribute Age, we first choose the splitting value. We sort the ages and get the list 20, 20, 20, 25, 25, 45, so the possible splitting values are 22.5 and 35.

$$Info_{Age=22.5}(D) = \frac{2}{6}I(2,2) + \frac{4}{6}I(4,2) = 0.67$$

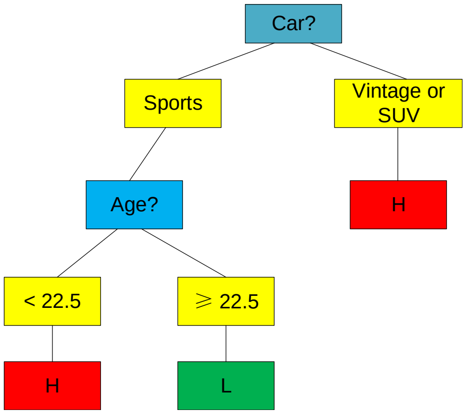
$$Info_{Age=35}(D) = \frac{5}{6}I(5,3) + \frac{1}{6}I(1,1) = 0.81$$

So among all the possible splitting method, the multi-way split for the Car attribute and the binary split of the Car attribute $\{Sports\}, \{Vintage, SUV\}$ are with the same maximum information gain, but the binary split is with a larger gain ratio. Thus, we choose the binary split of the Car attribute $\{Sports\}, \{Vintage, SUV\}$ as the first split for the decision tree.

Second, we consider the sub-tree with $Car = Sports$. We have the only attribute Age left and the possible splitting value is 22.5.

Third, we consider we consider the sub-tree with $Car \in \{Vintage, SUV\}$. There is no need for any further splitting since the output should be high risk.

Overall, the decision tree is shown as below:



(e) The point would go to the $Car \in \{Vintage, SUV\}$ branch and be classified as high risk.