VeriMadenciliği – Data Mining

HW – 3 -SOLUTION

1. A computer company made a survey over 10.000 people on computer game playing habits with asking asking favorite gaming portal (PC or PS or Xbox). The following table provides the survey results.

   Calculate the support, confidence and lift values for the following rules.
   Gender(Male) → Like(PC)
   Gender(Female) → Like(PC)

	PC	PS	XBox	total
Male	1500	4500	1000	7000
Female	1500	1000	500	3000
total	3000	5500	1500	10.000

solution:

Gender(Male) → Like(PC)

destek (support) : Gender(Male) AND Like(PC) / total
1500 / 10000 = %15
guven (confidence) : Gender(Male) AND Like(PC) / Gender(Male)
= 1500 / 7000 = % 21.4
kaldırac (lift) : Prob (Male AND PC) /Prob(Male)Prob(PC)

= (1500 / 10000) / (7000/10000)(3000/10000)
= 15 / 21 = 0.71 → yani negatif korelasyon (which means negative correlation)

Gender(Female) → Like(PC)

destek (support) : Gender(Female) AND Like(PC) / total
1500 / 10000 = %15
guven (confidence) : Gender(Female) AND Like(PC) / Gender(Female)
= 1500 / 3000 = % 50
kaldırac (lift) : Prob (Female AND PC) /Prob(Female)Prob(PC)

= (1500 / 10000) / (3000/10000)(3000/10000)
= 15 / 9 = 1.66 → yani positif korelasyon (which means positive correlation)

2. Daha once onişleme konusunda gorduğumuz X^2 (Chi-square) değerini yukardaki tablo için hesaplayın. 0.001 significance level için (preprocessing slaytlarındaki (26ıncı slayt) X^2 dağılım tablosunu kullanarak) buldugunuz değeri yorumlayın.

(ENG) Calculate the X^2 (Chi-square) value that we had seen previously on preprocessing section for the data table in question 1. Comment on your result using 0.001 significance rate (using the X^2 distribution table in the preprocessing slides (26th slide))

solution:

once beklenen degerler hesaplanır (first calculate the expected values)

beklenen(Erkek,PC) = 7000 * 3000 / 10000 = 2100
beklenen(Erkek,PS)= 7000 * 5500 / 10000 = 3850
beklenen(Erkek,XBox)= 7000 * 1500 / 1000 = 1050

beklenen(Kadın,PC) = 3000 * 3000 / 10000 = 900
beklenen(Kadın,PS) = 3000 *5500 / 10000 = 1650
beklenen(Kadın,Xbox) 3000 * 1500 / 10000 = 450

beklenen değerler tablosu, (expected value table)

	PC	PS	XBox	total
Male	2100	3850	1050	7000
Female	900	1650	450	3000
total	3000	5500	1500	10.000

$$X^2 = \frac{(1500-2100)^2}{2100} + \frac{(4500-3850)^2}{3850} + \frac{(1000-1050)^2}{1050} + \frac{(1500-900)^2}{900} + \frac{(1000-1650)^2}{1650} + \frac{(500-450)^2}{450} = 943$$

degree of freedom = (2-1)(3-1) = 2, significance level 0.001
tablodan bakarsak değer 13.81

943 > 13.8 oldugu icin korelasyon var.

3. Sınıflandırma konusundaki slaytlarda kullandıgımız (13. Slayt) buy_computer öğrenme verisini kullanarak $\text{Info}_{\text{student}}$, Gain(student), $\text{splitInfo}_{\text{student}}$ ve $\text{Gini}_{\text{student}}$ değerlerini hesaplayın

(ENG) Calculate the $\text{Info}_{\text{student}}$, Gain(student), $\text{splitInfo}_{\text{student}}$ and $\text{Gini}_{\text{student}}$ values for the training set used in the classification slides for buys_computer (14th slide)

solution:

Info(D) = Info(9,5) = -9/14 log(9/14) – 5/14 log(5/14) = 0.94

Student / buys_comp	Yes	No	toplam
Yes	6	1	7
No	3	4	7

Info_stu(yes)= Info(6,1) = -1/7 log(1/7) – 6/7 log(6/7) = 0.589
Info_stu(no) = Info(3,4) = -3/7 log(3/7) – 4/7 log(4/7) = 0.98

Info_stu = 7/14 * 0.589 + 7/14 * 0.98 = 0.786

GAIN (student) = 0.94 – 0.786 = 0.154

SplitInfo(Student)= -7/14 log(7/14) - 7/14 log(7/14) = 1

GainRatio(student) = Gain(student) / SplitInfo(student)
= 0.154 / 1 = 0.154

Gini(D) = $1-\left(\left(\frac{9}{14}\right)^2+\left(\frac{5}{14}\right)^2\right)=0.459$

Gini(student) = $\frac{7}{14}\left(1-\left(\left(\frac{1}{7}\right)^2+\left(\frac{6}{7}\right)^2\right)\right)+\frac{7}{14}\left(1-\left(\left(\frac{4}{7}\right)^2+\left(\frac{3}{7}\right)^2\right)\right)=0.366$

$$\Delta Gini = Gini(D) - Gini(student) = 0.459 - 0.366 = 0.093$$