TITLE OF PROJ. OR STUDY CIS-483/683 PROJ. OR STUDY NO.

SUBJECT Homework 3 WORKS

COMPUTER Answers Fall 2019 DATE

1. a) Bin1 = 105, 110, 115, 120
Bin2 = 140, 146, 150
Bin3 = 156, 160, 165, 171, 171, 172, 172, 173, 175
Bin4 = 180, 184, 185

b) Since there are 19 instances, you should have 3 bins with 5 instances in each bin and 1 bin with 4 instances — it does not matter which bin has only 4 instances

c) Weighted entropy if split at 148

$$= \frac{6}{19}\left(-\frac{1}{6}\log\frac{1}{6} - \frac{1}{6}\log\frac{1}{6} - \frac{4}{6}\log\frac{4}{6}\right)$$
$$+ \frac{13}{19}\left(-\frac{7}{13}\log\frac{7}{13} - \frac{6}{13}\log\frac{6}{13} - \frac{5}{13}\log\frac{5}{13}\right)$$
$$\approx 1.395$$

Weighted entropy if split at 168

$$= \frac{10}{19}\left(-\frac{1}{10}\log\frac{1}{10} - \frac{8}{10}\log\frac{8}{10} - \frac{1}{10}\log\frac{1}{10}\right)$$
$$+ \frac{9}{19}\left(-\frac{2}{9}\log\frac{2}{9} - \frac{2}{9}\log\frac{2}{9} - \frac{5}{9}\log\frac{5}{9}\right)$$
$$\approx 1.165$$

Since splitting at 168 has lower weighted entropy than splitting at 148, 168 is the better split point.

** You could have computed Information Gain instead - then 168 would have the greater Information Gain and so would be the better split point.

TITLE OF PROJ. OR STUDY _____ PROJ. OR STUDY NO. _____

SUBJECT _____ WORKS _____

COMPUTER _____ DATE _____ 20__

d) i) Consider intervals $(148, 168)$ and $(168, \infty)$

| | Reject | Admit | Wait |
|---|---|---|---|
| $(148, 168)$ | 0 | 4 | 0 |
| $(168, \infty)$ | 2 | 2 | 5 |

Degrees of freedom $= (2-1)(3-1) = 2$

Looking in the chi-square table in row 2 and column .90, the critical value is 4.605

$E_{11} = \frac{4*2}{13} = \frac{8}{13}$ $\quad E_{21} = \frac{9*2}{13} = \frac{18}{13}$

$E_{12} = \frac{4*6}{13} = \frac{24}{13}$ $\quad E_{22} = \frac{9*6}{13} = \frac{54}{13}$

$E_{13} = \frac{4*5}{13} = \frac{20}{13}$ $\quad E_{23} = \frac{9*5}{13} = \frac{45}{13}$

$$\chi^2 = \frac{\left(0 - \frac{8}{13}\right)^2}{8/13} + \frac{\left(4 - \frac{24}{13}\right)^2}{\frac{24}{13}} + \frac{\left(0 - \frac{20}{13}\right)^2}{20/13} + \frac{\left(2 - \frac{18}{13}\right)^2}{\frac{18}{13}} + \frac{\left(2 - \frac{54}{13}\right)^2}{\frac{54}{13}} + \frac{\left(5 - \frac{45}{13}\right)^2}{\frac{45}{13}}$$

$\approx 6.74$

Since $6.74 > 4.605$ (critical value), we reject the null hypothesis and do NOT merge the intervals.

ii) critical value is again 4.605

| | Reject | Admit | Wait |
|---|---|---|---|
| $(105, 148)$ | 1 | 4 | 1 |
| $(148, 168)$ | 0 | 4 | 0 |

$E_{11} = \frac{6*1}{10} = .6$ $\quad E_{21} = \frac{4*1}{10} = .4$

$E_{12} = \frac{6*8}{10} = 4.8$ $\quad E_{22} = \frac{4*8}{10} = 3.2$

$E_{13} = \frac{6*1}{10} = .6$ $\quad E_{23} = \frac{4*1}{10} = .4$

$$\chi^2 = \frac{(1-.6)^2}{.6} + \frac{(4-4.8)^2}{4.8} + \frac{(1-.6)^2}{.6} + \frac{(0-.4)^2}{.4} + \frac{(4-3.2)^2}{3.2} + \frac{(0-.4)^2}{.4}$$

$\approx 1.67$

Since $1.67 \not> 4.605$ (critical value), we do not reject the null hypothesis and we merge the intervals.

TITLE OF PROJ. OR STUDY _____ PROJ. OR STUDY NO. _____

SUBJECT _____ WORKS _____

COMPUTER _____ DATE _____ 20 _____

2. a) ii) You can look at what is displayed by Weka and see that the same intervals are not always used. Thus you can judge that Weka is using top-down and local discretization.

b) supervised Discretize: entropy based
unsupervised discretize: simple binning (equal interval)
unsupervised PKI discretize: equal frequency binning