CS412: Introduction to Data Mining, Fall 2023, Homework 2

Name: Teng Hou (tenghou2)

Q1

a. *We could assume that the dataset X represents the 20 sales price records. So the*

$$X = \{X_1, X_2, X_3 ... X_{20}\} = \{5, 7, 7, 9 ... 92\}$$

So we could get the four bins using the equal-frequency method, which is

$$Bin_1 = \{5, 7, 7, 9, 10\}$$

$$Bin_2 = \{11, 13, 15, 20, 29\}$$

$$Bin_3 = \{35, 47, 50, 55, 60\}$$

$$Bin_4 = \{65, 72, 80, 87, 92\}$$

Then we could get mean value by $X_{mean} = \frac{X_1+X_2+X_3+...+X_n}{n}$.
So mean values for four bins are:

$$\mu_1 = \frac{5 + 7 + 7 + 9 + 10}{5} = 7.6$$

$$\mu_2 = \frac{11 + 13 + 15 + 20 + 29}{5} = 17.6$$

$$\mu_3 = \frac{35 + 47 + 50 + 55 + 60}{5} = 49.4$$

$$\mu_4 = \frac{65 + 72 + 80 + 87 + 92}{5} = 79.2$$

b. *To find the equal width, we need to find out the maximum and minimum of dataset X. The maximum is 92, the minimum is 5. So the whole width is 92-5=87. Then the equal width is* $\frac{87}{4} = 21.75$.*So the three boundary points are*

$$Point_1 = 5 + 21.75 = 26.75$$

$$Point_2 = 5 + 21.75 * 2 = 48.5$$

$$Point_3 = 5 + 21.75 * 3 = 70.25$$

So the four bins are:

$$Bin_1 = \{5, 7, 7, 9, 10, 11, 13, 15, 20\}$$

$$Bin_2 = \{29, 35, 47\}$$

$$Bin_3 = \{50, 55, 60, 65\}$$

$$Bin_4 = \{72, 80, 87, 92\}$$

Then we could get mean value by $X_{mean} = \frac{X_1+X_2+X_3+...+X_n}{n}$*. So mean values for four bins are:*

$$Bin_1 = \frac{5+7+7+9+10+11+13+15+20}{9} = 10.7778$$

$$Bin_2 = \frac{29+35+47}{3} = 37$$

$$Bin_3 = \frac{50+55+60+65}{4} = 57.5$$

$$Bin_4 = \frac{72+80+87+92}{4} = 82.75$$

c. *To get variances of each bin, we could use* $var = \frac{\Sigma_{i=1}^{n}(X_i-\mu)^2}{n}$.
For equal-frequency bins, the variances are:

$$Var_1 = \frac{\Sigma_{i=1}^{5}(X_i - 7.6)^2}{5} = 3.04$$

$$Var_2 = \frac{\Sigma_{i=6}^{10}(X_i - 17.6)^2}{5} = 41.44$$

$$Var_3 = \frac{\Sigma_{i=11}^{15}(X_i - 49.4)^2}{5} = 71.44$$

$$Var_1 = \frac{\Sigma_{i=16}^{20}(X_i - 79.2)^2}{5} = 95.76$$

For equal-width bins, the variances are:

$$Var_1 = \frac{\Sigma_{i=1}^{9}(X_i - 10.7778)^2}{9} = 19.284$$

$$Var_2 = \frac{\Sigma_{i=10}^{12}(X_i - 37)^2}{3} = 56$$

$$Var_3 = \frac{\Sigma_{i=13}^{16}(X_i - 57.5)^2}{4} = 31.25$$

$$Var_1 = \frac{\Sigma_{i=17}^{20}(X_i - 82.75)^2}{4} = 56.6875$$

So the average variance for equal-frequency and equal-width are

$$Var_{freq} = \frac{3.04+41.44+71.44+95.76}{4} = 52.92$$

$$Var_{width} = \frac{19.284+56+31.25+56.6875}{4} = 40.8054$$

So the variance of equal-width is smaller than that of equal-frequency, which means equal-width's effect is better.

d. *i. The formula is* $x_i^{'} = \frac{X_i - X_{min}}{X_{max} - X_{min}} * (X_{max}^{'} - X_{min}^{'}) + X_{min}^{'}$ *So the dataset after Min-max normalization is:*

$$X^{'} = \{500, 511.4943, 511.4943, 522.9885, 528.7356, 534.4828, 545.9770, 557.4713, 586.2069, 637.9310, 672.4138\}$$

$$\{741.3793, 758.6207, 787.3563, 816.0920, 844.8276, 885.0575, 931.0345, 971.2644, 1000\}$$

ii. The maximum value is 92, so we need to divide all the data by 10^j*. So the new dataset is*

$$X^{'} = \{0.05, 0.07, 0.07, 0.09, 0.10, 0.11, 0.13, 0.15, 0.20, 0.29, 0.35, 0.47, 0.50, 0.55, 0.60, 0.65, 0.72, 0.80, 0.87, 0.92\}$$

Q2

a. *We can calculate the information gain by*

$$Gain(A) = Info(D) - Info_A(D)$$

where

$$Info(D) = -\Sigma_{i=1}^{m} p_i \log_2 p_i$$

$$Info_A(D) = \Sigma_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

For whether passing the final exam:

$$Info(final) = -\frac{8}{13} * log_2(\frac{8}{13}) - \frac{5}{13} * log_2(\frac{5}{13}) = 0.9612$$

For hobby, there are four students who are interested in painting and all of them passed the exam. And five students like music which two of them passed while the left failed. Then four students' hobby is swimming, half of them passed and half failed. So

$$Info_{hobby}(final) = \frac{4}{13}\times(-\frac{4}{4}log_2(\frac{4}{4})+\frac{5}{13}\times(-\frac{2}{5}log_2\frac{2}{5}-\frac{3}{5}log_2\frac{3}{5})+\frac{4}{13}\times(-\frac{2}{4}log_2\frac{2}{4}-\frac{2}{4}log_2\frac{2}{4}) = 0.6811$$

Then

$$Gain(hobby) = Info(final) - Info_{hobby}(final) = 0.9612 - 0.6811 = 0.2801$$

$$Info_{hobby}(final) = \frac{4}{13}\times(-\frac{4}{4}log_2(\frac{4}{4})+\frac{5}{13}\times(-\frac{2}{5}log_2\frac{2}{5}-\frac{3}{5}log_2\frac{3}{5})+\frac{4}{13}\times(-\frac{2}{4}log_2\frac{2}{4}-\frac{2}{4}log_2\frac{2}{4}) = 0.6811$$

For color, there are five students like red, three of them passed the exam while two failed. One likes green and two like purple, all of them passed. Two like blue, one passed one failed. Three like yellow, one passed two failed. So:

$$Info_{color}(final) = \frac{1}{13} \times (-\frac{1}{1}log_2(\frac{1}{1}) + \frac{2}{13} \times (-\frac{2}{2}log_2(\frac{2}{2}) + \frac{2}{13} \times (-\frac{1}{2}log_2\frac{1}{2}-$$

$$\frac{1}{2}log_2\frac{1}{2}) + \frac{3}{13} \times (-\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3}) + \frac{5}{13} \times (-\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5}) = 0.7392$$

Then

$$Gain(color) = Info(final) - Info_{color}(final) = 0.9612 - 0.7392 = 0.2220$$

For practiced hours, there are three students pay 5 hours, two passed one failed. Two pay 9 hours, both passed. Three pay 6 hours, two passed one failed. Two pay 7 hours, one passed one failed. Three pay 8 hours, one passed, two failed. So:

$$Info_{time}(final) = \frac{3}{13}\times(-\frac{2}{3}log_2\frac{2}{3}-\frac{1}{3}log_2\frac{1}{3})+\frac{3}{13}\times(-\frac{2}{3}log_2(\frac{2}{3}-\frac{1}{3}log_2(\frac{2}{3})+\frac{2}{13}\times(-\frac{1}{2}log_2\frac{1}{2}-$$

$$\frac{1}{2}log_2\frac{1}{2}) + \frac{2}{13} \times (-\frac{2}{2}log_2\frac{2}{2}) = 0.7896$$

Then

$$Gain(time) = Info(final) - Info_{time}(final) = 0.9612 - 0.7896 = 0.1716$$

For homework, there are six students submitted and five passed one failed. Seven students submitted and 3 passed 4 failed. So:

$$Info_{homework}(final) = \frac{6}{13} \times (-\frac{5}{6}log_2\frac{2}{6} - \frac{1}{6}log_2\frac{1}{6}) + \frac{7}{13} \times (-\frac{3}{7}log_2\frac{3}{7} - \frac{4}{7}log_2\frac{4}{7}) = 0.8305$$

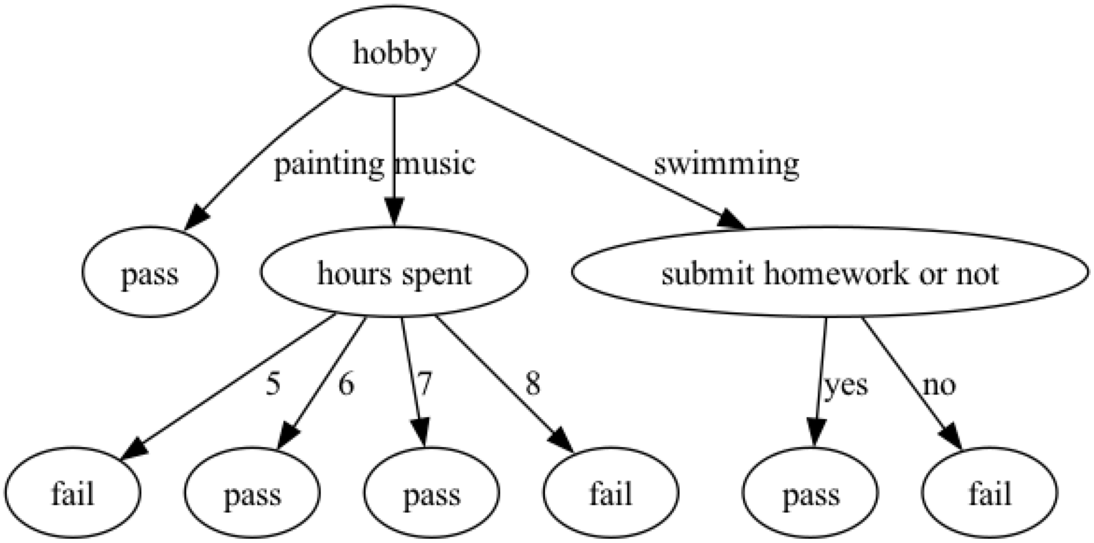
Then

$$Gain(homework) = Info(final) - Info_{homework}(final) = 0.9612 - 0.8305 = 0.1307$$

So for hobby, information gain is 0.2801. For color, information gain is 0.2220. For practiced hours, information gain is 0.1716. For submitted homework, information gain is 0.1307.

b. *Who has the largest information gain will be the root of decision tree. Based on the calculation of a, it is really easy to find that hobby has the largest information gain. So the root of decision tree should be hobby.*

c.



Q3

	Covid19(0.001)	No Covid19(0.999)
Test Positive	1	0.05
Test Negative	0	0.95

1. *From the text, we know that if a person is Covid-19 positive, then the result will be positive in 100 percent. If a person doesn't have Covid-19, then there are 5 percent probability to get the result positive. 95 percent to get the result negative. And from the text:*

$$P(Covid19) = 0.001$$

$$P(Positive|Covid19) = 1$$

$$P(Positive|NotCovid19) = 0.05$$

From the Baye's theorem:

$$P(Covid19|Positive) = \frac{P(Positive|Covid19) * P(Covid19)}{P(Positive)} = \frac{1 \times 0.001}{1 \times 0.001 + 0.05 \times 0.999}$$

The answer is 0.0196.

Q4

a. *here is the likelihood table.*

		Covid19	Outcome
		Positive	Negative
Fever	Yes	$\frac{3}{4}$	$\frac{1}{3}$
	No	$\frac{1}{4}$	$\frac{2}{3}$

		Covid19	Outcome
		Positive	Negative
Cough	Yes	$\frac{1}{2}$	$\frac{2}{3}$
	No	$\frac{1}{2}$	$\frac{1}{3}$

		Covid19	Outcome
		Positive	Negative
Headache	Yes	$\frac{1}{2}$	$\frac{1}{3}$
	No	$\frac{1}{2}$	$\frac{2}{3}$

b. *From the table, we could get:*
For positive part:

$$P(Fever = no|Positive) = 0.25$$

$$P(Cough = no|Positive) = 0.5$$

$$P(Headache = yes|Positive) = 0.5$$

So

$$P(X|Positive) = 0.25 * 0.5 * 0.5 = 0.0625$$

$$P(Positive) = \frac{4}{7}$$

Then

$$P(Positive|X) \propto P(X|Positive) \times P(Positive) = 0.0357$$

For negative part:

$$P(Fever = no|Negative) = \frac{2}{3}$$

$$P(Cough = no|Negative) = \frac{1}{3}$$

$$P(Headache = yes|Negative) = \frac{1}{3}$$

So

$$P(X|Negative) = \frac{2}{3} * \frac{1}{3} * \frac{1}{3} = \frac{2}{27}$$

$$P(Negative) = \frac{3}{7}$$

Then

$$P(Negative|X) \propto P(X|Negative) \times P(Negative) = 0.0317$$