ENGG 5103: Techniques for data mining

Wanli Wang
2020fall-ENGG5103-HW2-Wang-Wanli-1155160517

I. Question 1

(Preprocessing)

TABLE I
INITIAL DATA RECORDS

Course ID	Course Name	Instructor Gender	Average GPA	Required Course
CSC236	Theory of Computing	Male	3.2	Yes
CSC263	Data Structures	Female	3.2	Yes
CSC373	Algorithms	Female	3.8	No
CSC463	Complexity Theory	Male		No
CSC258	Computer Organizations	Male	3.8	No
CSC165	Computer Logic	Male	2.0	No
CSC411	Machine Learning	Female	2.0	No
CSC108	Introduction to Computer Science	Male	4.0	No
CSC384	Artificial Intelligence	Female	3.7	Yes

a) Clean data with missing values. Output the new table at this step;
Since CSC463 Complexity Theory has missing value, we can directly remove this record.

TABLE II
DATA RECORDS WITH CLEANING DATA

Course ID	Course Name	Instructor Gender	Average GPA	Required Course
CSC236	Theory of Computing	Male	3.2	Yes
CSC263	Data Structures	Female	3.2	Yes
CSC373	Algorithms	Female	3.8	No
CSC258	Computer Organizations	Male	3.8	No
CSC165	Computer Logic	Male	2.0	No
CSC411	Machine Learning	Female	2.0	No
CSC108	Introduction to Computer Science	Male	4.0	No
CSC384	Artificial Intelligence	Female	3.7	Yes

b)There is a redundant feature. Please remove that feature and select useful features. Output the new table at this step;
We remove redundant feature of Instructor Gender, which is usesless. In fact, we can also remover the feature of course name since the course ID is in corresponding to the course name. But this feature of course name is not suggested to be removed since it is better for students to know the information of the course.

TABLE III
DATA RECORDS WITH REMOVING REDUNDANT FEATURE

Course ID	Course Name	Average GPA	Required Course
CSC236	Theory of Computing	3.2	Yes
CSC263	Data Structures	3.2	Yes
CSC373	Algorithms	3.8	No
CSC463	Complexity Theory		No
CSC258	Computer Organizations	3.8	No
CSC165	Computer Logic	2.0	No
CSC411	Machine Learning	2.0	No
CSC108	Introduction to Computer Science	4.0	No
CSC384	Artificial Intelligence	3.7	Yes

c) Convert categorical values to numerical values and output the new table. Output the new table at this step;
We convert categorical values Male and Female of instructor gender as 1 and 0, respectively. In addition, we convert categorical values Yes and No of required courses as 1 and 0, respectively.

TABLE IV
DATA RECORDS WITH CONVERTING CATEGORICAL VALUES TO NUMERICAL VALUE

Course ID	Course Name	Instructor Gender	Average GPA	Required Course
CSC236	Theory of Computing	1	3.2	1
CSC263	Data Structures	0	3.2	1
CSC373	Algorithms	0	3.8	0
CSC463	Complexity Theory	1		0
CSC258	Computer Organizations	1	3.8	0
CSC165	Computer Logic	1	2.0	0
CSC411	Machine Learning	0	2.0	0
CSC108	Introduction to Computer Science	1	4.0	0
CSC384	Artificial Intelligence	0	3.7	1

II. QUESTION 2

(Backpropagation)
Here is a simple neural network.

$$z = wx + b \tag{1}$$
$$y = \delta(z) \tag{2}$$
$$L = \frac{1}{2}(y-t)^2 \tag{3}$$
$$R = \frac{1}{2}w^2 \tag{4}$$
$$L_{reg} = L + \lambda R \tag{5}$$

Please derive the gradients $\frac{\partial L_{reg}}{\partial w}$ and $\frac{\partial L_{reg}}{\partial b}$ by computing $\frac{\partial L_{reg}}{\partial L_{reg}}$, $\frac{\partial L_{reg}}{\partial R}$, $\frac{\partial L_{reg}}{\partial L}$, $\frac{\partial L_{reg}}{\partial y}$, $\frac{\partial L_{reg}}{\partial z}$.

Since we have the loss function as follows

$$L_{reg} = \frac{1}{2}(y-t)^2 + \lambda R, \tag{6}$$

we have derivation to the weight w as

$$\begin{aligned}\frac{\partial L_{reg}}{\partial w} &= \frac{\partial L_{reg}}{\partial L}\frac{\partial L}{\partial z}\frac{\partial z}{\partial w} + \frac{\partial L_{reg}}{\partial R}\frac{\partial R}{\partial w} \\ &= \frac{\partial L_{reg}}{\partial L}\frac{\partial L}{\partial y}\frac{\partial y}{\partial z}\frac{\partial z}{\partial w} + \frac{\partial L_{reg}}{\partial R}\frac{\partial R}{\partial w} \\ &= (y-t)\frac{\partial y}{\partial z}\frac{\partial z}{\partial w} + \frac{\partial L_{reg}}{\partial R}\frac{\partial R}{\partial w}.\end{aligned} \tag{7}$$

Here,we consider sigmoid function as activation function.

$$\delta(x) = \frac{1}{1+e^{-x}}. \tag{8}$$

Therefore, we have the derivation of δ to x.

$$\frac{\partial \delta(x)}{\partial x} = \delta(x)\left(1-\delta(x)\right). \tag{9}$$

Therefore, (7) can be reformulated as

$$\begin{aligned}\frac{\partial L_{reg}}{\partial w} &= (y-t)\delta(z)\left(1-\delta(z)\right)x + \lambda w \\ &= (y-t)y\left(1-y\right)x + \lambda w.\end{aligned} \tag{10}$$

Similarly, we have derivation to the bias b as follows:

$$\begin{aligned}\frac{\partial L_{reg}}{\partial b} &= \frac{\partial L_{reg}}{\partial L}\frac{\partial L}{\partial z}\frac{\partial z}{\partial b} \\ &= \frac{\partial L_{reg}}{\partial L}\frac{\partial L}{\partial y}\frac{\partial y}{\partial z}\frac{\partial z}{\partial b} \\ &= (y-t)y(1-y).\end{aligned} \tag{11}$$

Therefore, we have

$$\frac{\partial L_{reg}}{\partial w} = (y-t)y\,(1-y)\,x + \lambda w, \tag{12}$$

$$\frac{\partial L_{reg}}{\partial b} = (y-t)y(1-y). \tag{13}$$

III. Question 3

(Bayes Probability) (25%)

We have statistics indicating that about 50% of emails are spam emails. A software claims that it can detect 95% of spam emails, and the probability for a false positive (a non-spam email detected as spam) is 5%. Now if an email is detected as spam, then what is the probability that it is in fact a non-spam email?

Define following events:

A = event that an email is detected as spam,

B = event that an email is spam,

$\bar{B}$ = event that an email is not spam.

In addition, we have

$$P(A|B) = 0.95 \tag{14}$$

$$P(A|\bar{B}) = 0.05 \tag{15}$$

$$P(B) = 0.5 \tag{16}$$

$$P(\bar{B}) = 0.5 \tag{17}$$

Then, the Bayesian theory gives

$$P(\bar{B}|A) = \frac{P(A|\bar{B})P(\bar{B})}{P(A)} \tag{18}$$

where the probability $P(A)$ is

$$P(A) = P(A|\bar{B})P(\bar{B}) + P(A|B)P(B). \tag{19}$$

Therefore, we have

$$\begin{aligned} P(\bar{B}|A) &= \frac{P(A|\bar{B})P(\bar{B})}{P(A)} \\ &= \frac{P(A|\bar{B})P(\bar{B})}{P(A|\bar{B})P(\bar{B}) + P(A|B)P(B)} \\ &= \frac{0.05 \times 0.5}{0.05 \times 0.5 + 0.95 \times 0.5} \\ &= \frac{5}{100} \end{aligned} \tag{20}$$

Therefore, if an email is detected as spam, then the probability that it is in fact a non-spam email is 5%.

IV. Question 4

(Decision Tree) Using the data in table 1, build a decision tree for customers of the online shop *Shein* into *Satisfied* or *Unsatisfied*. Use the Information Gain (IG) as the decision criterion to select which attribute to split on. Please show your detailed calculations.

TABLE V
CUSTOMERS INFORMATION OF THE ONLINE SHOP *Shein*

Person ID	friendly website?	delivery time	Good Service?	Satisfied?
1	Yes	Long	No	Yes
2	No	Short	Yes	Yes
3	Yes	Long	Yes	No
4	No	Long	Yes	Yes
5	Yes	Short	Yes	No

At first, we can change the table in V as the the table in VI. The information gain (IG) $g(D, A)$ is defined as follows:

$$g(D, A) = H(D) - H(D|A) \tag{21}$$

where $H(D)$ and $H(D|A)$ are entropy and conditional entropy, respectively.

TABLE VI
CHANGED CUSTOMERS INFORMATION OF THE ONLINE SHOP *Shein*

Person ID	friendly website?	delivery time	Good Service?	Satisfied?
1	1	1	0	Yes
2	0	0	1	Yes
3	1	1	1	No
4	0	1	1	Yes
5	1	0	1	No

1) If friendly website is node node, i.e., D : friendly website : $\{A_1$: Yes, A_2 : No$\}$

$$H(D) = -\left(\frac{3}{5}log_2\frac{3}{5} + \frac{2}{5}log_2\frac{2}{5}\right) = 0.9710 \tag{22}$$

$$H(A_1) = -\left(\frac{1}{3}log_2\frac{1}{3} + \frac{2}{3}log_2\frac{2}{3}\right) = 0.9183 \tag{23}$$

$$H(A_2) = 0 \tag{24}$$

$$\begin{aligned} g(D,A) &= H(D) - H(D|A) \\ &= 0.9710 - (\frac{3}{5} \times 0.9183 + \frac{2}{5} \times 0) \\ &= 0.4200 \end{aligned} \tag{25}$$

If delivery time is node node, i.e., D : delivery time : $\{A_1$: Long, A_2 : Short$\}$

$$H(D) = -\left(\frac{3}{5}log_2\frac{3}{5} + \frac{2}{5}log_2\frac{2}{5}\right) = 0.9710 \tag{26}$$

$$H(A_1) = -\left(\frac{1}{3}log_2\frac{1}{3} + \frac{2}{3}log_2\frac{2}{3}\right) = 0.9183 \tag{27}$$

$$H(A_2) = -\left(\frac{1}{2}log_2\frac{1}{2} + \frac{1}{2}log_2\frac{1}{2}\right) = 1 \tag{28}$$

$$\begin{aligned} g(D,A) &= H(D) - H(D|A) \\ &= 0.9710 - (\frac{3}{5} \times 0.9183 + \frac{2}{5} \times 1) \\ &= 0.0200 \end{aligned} \tag{29}$$

If good service is node node, i.e., D : good service : $\{A_1$: Yes, A_2 : No$\}$

$$H(D) = -\left(\frac{3}{5}log_2\frac{3}{5} + \frac{2}{5}log_2\frac{2}{5}\right) = 0.9710 \tag{30}$$

$$H(A_1) = -\left(\frac{2}{4}log_2\frac{2}{4} + \frac{2}{4}log_2\frac{2}{4}\right) = 1 \tag{31}$$

$$H(A_2) = 0 \tag{32}$$

$$\begin{aligned} g(D,A) &= H(D) - H(D|A) \\ &= 0.9710 - (\frac{4}{5} \times 1 + \frac{1}{5} \times 0) \\ &= 0.1710 \end{aligned} \tag{33}$$

According to (22), (34) and (38), friendly website is regarded as the root node.
2) The root node is friendly website and the value is No:
There is no child node.
3) The root node is friendly website and the value is Yes:

3.1) If delivery time is child node, i.e., D : delivery time : $\{A_1$: Long, A_2 : Short$\}$

$$H(D) = -\left(\frac{2}{3}log_2\frac{2}{3} + \frac{1}{3}log_2\frac{1}{3}\right) = 0.9183 \tag{34}$$

$$H(A_1) = -\left(\frac{1}{2}log_2\frac{1}{2} + \frac{1}{2}log_2\frac{1}{2}\right) = 1 \tag{35}$$

$$H(A_2) = 0 \tag{36}$$

$$\begin{aligned} g(D,A) &= H(D) - H(D|A) \\ &= 0.9183 - (\frac{2}{3} \times 1 + \frac{1}{3} \times 0) \\ &= 0.2516 \end{aligned} \tag{37}$$

3.2) If good service is child node, i.e., D : good service : $\{A_1$: Yes, A_2 : No$\}$

$$H(D) = -\left(\frac{2}{3}log_2\frac{2}{3} + \frac{1}{3}log_2\frac{1}{3}\right) = 0.9183 \tag{38}$$

$$H(A_1) = 0 \tag{39}$$

$$H(A_2) = 0 \tag{40}$$

$$\begin{aligned} g(D,A) &= H(D) - H(D|A) \\ &= 0.9183 - (\frac{2}{3} \times 0 + \frac{1}{3} \times 0) \\ &= 0.9183 \end{aligned} \tag{41}$$

Therefore, good service is child node.

In summary, the root node is friendly website. There is no child node in corresponding to its value No. However, there is child node good service in corresponding to its value Yes.

Code is as follows: https://github.com/wangwanlitype1/ENGG5103