Data Mining first assessment (Fall 2021)

Max score possible: 110%

Question 1 (10%)

A network security briefing reports that 94% of networks got compromised have a firewall. Does it mean setting up a firewall is useless to secure a network? Does it mean it's better off without a firewall since there is a high probability of network security being compromised? Explain the reason behind your conclusion with a concrete mathematical example.

Let X={0,1} be the variable accounting for (not) getting comprised
Let Y={0,1} be the variable accounting for (not) setting up a firewall

Given:

$Pr(X:1 \mid Y:1) = 0.94$

Question:

$Pr(X:1\mid Y:1)/Pr(X:1 \mid Y:0) > 1$ to favor (NOT) to have firewall?

Or $Pr(X:1\mid Y:1) > Pr(X:1 \mid Y:0) \Leftrightarrow Pr(X:0\mid Y:0) > Pr(X:0 \mid Y:1)$

Question 2 (10% or 25%)

There are 16 items. One differs from the rest. We do not know whether it is heavier or lighter than the rest. You have a scale that will balance if the weight on both sides is the same. Otherwise the heavier side will go "down".

(10%) Part 1: What is the min. number of measurements you will need to use the scale to identify the odd one AND be able to tell whether it is heavier or lighter? Why?

There are 16x2=32 possibilities. The scale can at best discern 3 outcomes at a time.

Ceiling[Log_2(32)/Log_2(3)] = 3.15 -> 4

(15%) Part 2: Show the steps to identify the odd one and tell whether it is heavier or lighter.
Remark: Skip this part if you attempt Question 4 part 4. You will NOT get extra credit to attempt both.

For 16 objects, there are 32 cases H/L for Obj-1 … H/L Obj-16

5-5-6 6-6-5 4-4-8

5-5 | 6
Not balance | Balance
Left with 10 cases | 12 cases

(1)6-6 | 4
Not balance | Balance
12 cases | **8 cases L13,L14,L15,L16,H13,H14,H15,H16

L1-L6 or H7-H12

(2) L1,L2,L3,H7,H8 <-> L4,L5,L6,H9,H10

Not balance | Balance
L1,L2,L3 or H9 H10 | H11, H12 (Just measure against each other)

(3)L1,L2 <->L3,H9

Either L3, or L1,L2,L9 (Just measure L1 vs L2)

**8 cases L13,L14,L15,L16,H13,H14,H15,H16
(2)13,14<->15,16
Not balance
Say, 13L,14L, or 15H,16H
(3) 13L<-> 14L
If not balance, lighter is the answer
If balance, (4) 15H <-> 16H heavier is the answer

Solve the problem recursively. 6-6-4 or 5-5-6

Putting it together
Step 1 (Equal) Obs: H/L 13-16
O1 … O6 O7 .. O12

Step 1 (Equal) -> Step 2:

```
O13 O14          S1 S2
  If (equal) then {
          if (O15 == S1)  { // step 3
            Measure (O16 S1)        // step 4
          } else O15
  Else {
          Repeat for (O15 O16 vs S1 S2)
  }
```

Step 1 (Right side up) -> 2: Obs: One of the H1-H6 is heavier, or L7-L12 is lighter
H1 H2 L7 L8 H3 H4 L9 L10

Step 1 (Right side up) -> 2 (Right side up) Obs: H1, H2, L9, L10
H3 H4 L7 L8 H1 H2 L9 L10

Step 1 (Right side up) -> 2 (Equal) Obs: H5, H6, L11, L12

Step 1 (Left side up) -> 2: Obs: One of the L1-L6 is lighter, or H7-H12 is heavier
L1 L2 H7 H8 L3 L4 H9 H10

Step 1 (Left side up) -> 2 (Left side up) Obs: L1, L2, H9, H10
L3 L4 H7 H8 L1 L2 H9 H10

Step 1 (Left side up) -> 2 (Equal) Obs: L5, L6, H11, H12

In all cases with a pattern H-x, H-y, L-w, L-z
H-x L-w S1 S2
If equal, measure (H-y vs S) to determine it is H-y or L-z
If not equal, measure (H-x vs S) to determine it is Hx or L-w

Question 3 (40%)
Given the data set below:

	x1	x2	x3	f
Row 1	0.148	8.76	73.201	13.283
Row 2	0.693	5.393	68.224	***v1=? 3872.47***
Row 3	0.427	4.621	72.191	21.053
Row 4	0.967	8.622	12.303	***v2=?17902.73***

Row 5	0.153	7.797	83.466	***v3=?14.5785***
Row 6	0.822	9.968	51.702	***v4=?8831.451***
Row 7	0.191	6.115	42.621	3.507
Row 8	0.156	8.401	54.954	15.006

Table 1

Define a mapping function ***h1*** such that: $h1(x1) \xrightarrow{h1} f$.
In other words, ***h1(0.148)= 13.283***, ***h1(0.427)= 21.053***, … etc.

Define a mapping function ***h*** such that: $(x1,x2,x3) \xrightarrow{h} f$.
In other words, ***h(0.148, 8.76, 73.201)= 13.283***, ***h(0.427, 4.621, 72.191)= 21.076***, … etc.

(12%) Part 1: Write down the expression (mathematical structure) of the one-dimensional Larange polynomial regression shown in the text book (page 50) for ***h1***.

h1(x1) = (x1-0.427)(x1-0.191)(x1-0.156)x13.283/(0.148-0.427)(0.148-0.191)(0.148-0.156) +
(x1-0.148)(x1-0.191)(x1-0.156)x21.053/(0.427-0.148)(0.427-0.191)(0.427-0.156) +
(x1-0.148)(x1-0.427)(x1-0.156)x3.507/(0.191-0.148)(0.191-0.427)(0.191-0.156) +
(x1-0.148)(x1-0.427)(X1-0.191)x15.006/(0.156-0.148)(0.156-0.427)(0.156-0.191)

(12%) Part 2: Derive ***v1, v2, v3*** and ***v4*** using ***h1.***

h(x1)	X1	
3872.47	0.693	V1
17902.73	0.967	V2
14.5785	0.153	V3
8831.451	0.822	V4

(10%) Part 3: Generalize the one-dimensional Larange polynomial regression shown in the text book (page 50) to three dimensions and apply the generalization to derive the mathematical structure of the mapping function ***h***.

(6%) Part 4: Derive ***v1*** using ***h***. Compare this value against the one that you derived using ***h1***. Explain why the derivation for ***v1*** using ***h, and h1*** are similar or not similar. For this question, you do not need to derive ***v2, v3*** and ***v4*** using ***h.***

Question 4 (35% or 50%)

(5%) Part 1: Use the values of ***v1, v2, v3*** and ***v4*** you derived using ***h1*** for this question. Create a new table containing ***(y1, y2, y3, g)*** using table 1 and the following rules for discretization:

$y1 =$ 0 if $0 \le x1 < 0.15$
1 if $0.15 \le x1$

$y2 =$ 0 if $0 \le x2 < 8$
1 if $8 \le x2 < 10$

$y3 =$ 0 if $0 \le x3 < 30$
1 if $30 \le x3 < 60$
2 if $60 \le x3$

$g =$ 0 if $f \le 4$

1 if 4 ≤ f < 15
2 if 15≤ f

	Y1	Y2	Y3	G
Row 1	0	1	2	1
Row 2	1	0	2	***2***
Row 3	1	0	2	2
Row 4	1	1	0	***2***
Row 5	1	0	2	***1***
Row 6	1	1	1	***2***
Row 7	1	0	1	0
Row 8	1	1	1	2

Table 2

(15%) Part 2: Find all 2nd order association patterns involving ***(y1 g)*** that is/are statistically significant using a threshold 0.4

Need to check

Pattern	Frequency	Pass threshold test
(0 1)	1	
(1 1)	1	
(1 2)	5	yes
(1 0)	1	

Pr(y1=1) = 0.875 Pr(g=2)=0.625

Log_2 (0.625/0.875*0.625)=0.192

N=8
Chi-square=(5-4.375)^2/4.375=0.089
Chi-square/2N =0.00558

MI > (Chi-square/2N) -> Significant

(15%) Part 3: Find all 3rd order association patterns involving ***(y1 y2 y3)*** that that is/are statistically significant using a threshold 0.13

Pattern	Frequency	Pass threshold test
(0 1 2)	1	yes
(1 0 2)	3	yes
(1 1 0)	1	yes
(1 1 1)	2	yes
(1 0 1)	1	yes

P(y1=1) = 0.625 Pr(y2=0) = 0.5 Pr(y2=1) = 0.5
Pr(y3=1) = 0.375 Pr(y3=2) = 0.5

E' = 3*(1/8)Log_2 8 + (3/8)Log_2(8/3) + (2/8)Log_2(4) = 9/8 + 0.53 + 0.5 = 2.15564

E^=Log_2(12)=3.585
N=8

(E'/E^)^1.5=0.6

Pattern (1 0 2)
Log_2 Pr(1 0 2)/Pr(y1=1)Pr(y2=0)Pr(y3=2) = Log_2 (0.375/0.625*0.5*0.5) = log_2 (2.4)=1.263

Chi-square (1 0 2) = (3-8*0.625*0.5*0.5)^2/(8*0.625*0.5*0.5) = (3-1.25)^2/1.25 = 2.45

(8/3)(2.45/16)^0.6=0.865

Yes (1 0 2) is significant

Pattern (1 1 1)
Log_2 Pr(1 1 1)/Pr(y1=1)Pr(y2=1)Pr(y3=1) = Log_2(0.25/0.625*0.5*0.375) = log_2(2.133)=1.092

Chi-square(1 1 1) = (2-8*0.625*0.5*0.375)^2/(8*0.625*0.5*0.375) = (2-0.9375)^2/0.9375=1.204

(8/2)(1.204/16)^0.60 = 0.8471

Yes (1 1 1) is significant

(15%) Part 4: Derive the optimal decision tree to predict ***g*** using ***(y1 y2 y3)*** and the following frequency information:
Remark: Skip this part if you attempt Question 2 part 2. You will NOT get extra credit to attempt both.

Pattern	Output	Frequency
(0 1 2)	1	1
(1 0 2)	2	3
(1 1 0)	2	1
(1 1 1)	2	2
(1 0 1)	0	1

x1	x2	x3	f
0	0	2	1
1	0	2	***2***
1	0	2	2
1	1	0	***2***
1	0	2	***1***
1	1	1	***2***
1	0	1	0
1	1	1	2

First question:
I(X1:1 -> f) = 5/7log(7/5)+ 2x (1/7)log(7)
E(X1 -> f) = 5/8log(7.5)+(2/8)log(7) = 1.005

I(X2:0-> f) = (1/3)log(3) + (2/3)log(3/2)
I(X2:1-> f) = (2/5)log(5/2) + (3/5)log(5/3)
E(X2-> f) = 1.19

I(X3:2->f) = (2/4)log(4/2) + (2/4)log(4/2) = 1
I(X3:1->f) = (1/3)log(3) + (2/3)log(3/2)
E(x3->f)= 1/8 + (1/8)log(3) + (2/8)log(3/2) = 0.46936

Winner: x3

I(X3:2 X2:0 -> f) = 0
I(X3:2 X2:1->f) = (1/3)log(3) + (2/3)log(3/2)
I(X3:1 X2:1->f) = (1/3)log(3) + (2/3)log(3/2)
E(X3, X2 -> f) = 2x((1/8)log(3) + (2/8)log(3/2)) = 0.5425

I(X3:2, X1: 1) = (1/3)log(3) + (2/3)log(3/2)
I(X3:1, X1:1) = (1/3)log(3) + (2/3)log(3/2)
E(X3, X1 -> f) = 0.5425

Equally good; x3-> x2 or x3 -> x1. E(X3, X2 -> f)