MSc DSAA5002
2023 Fall
Midterm Solutions

1. Apriori (10 marks)

Given a transaction database as below,

TID	Item
T1	A,B,C
T2	A,D,E
T3	A,C,F
T4	D,E,F
T5	B,C,F
T6	A,C,D,E
T7	C,D,E
T8	B,C
T9	A,C,D,E
T10	B,F

(1) (6 marks) Set minsupport = 40%, please utilize Apriori to obtain all frequent itemsets, and point out the maximal frequent itemsets.

(2) (4 marks) Set minconfidence = 60%, please utilize Apriori to find all strong association rules.

Sol:

1)

Step 1: Introduce a new item for each distinct attribute value pair.

A: Whether Condition=Good

B: Whether Condition=Bad

C: Driver Status=Alcohol

D: Driver Status=Sober

E: $20K \le$Annual Income$\le 25K$

F: $25K <$Annual Income$\le 30K$

G: $30K <$Annual Income$\le 35K$

H: $35K <$Annual Income$\le 40K$

I: Safety Belt=YES

J: Safety Belt=NO

K:Damaged Condition=Major

L: Damaged Condition=Minor

Step 2:Construct the transaction DB.

TID	Item
T1	A,C,H,J,K

T2	B,D,E,I,L
T3	A,D,F,I,L
T4	A,D,H,I,K
T5	B,D,E,J,K
T6	A,C,G,I,L
T7	B,C,E,I,K
T8	A,D,E,I,K
T9	A,C,G,I,K
T10	B,D,E,J,K

Step 3: Scan the transaction database, generate the candidate 1-itemset C_1, and calculate the support of each C_1 itemset. Determine the frequent 1-itemset L_1.

C_1

Itemset	Support Count	Support degree
{A}	6	60%
~~{B}~~	~~4~~	~~40%~~
~~{C}~~	~~4~~	~~40%~~
{D}	6	60%
{E}	5	50%
~~{F}~~	~~1~~	~~10%~~
~~{G}~~	~~2~~	~~20%~~
~~{H}~~	~~2~~	~~30%~~
{I}	7	70%
{J}	3	40%
{K}	7	70%
~~{L}~~	~~3~~	~~30%~~

L_1

Itemset	Support Count	Support degree
{A}	6	60%
{D}	6	60%
{E}	5	50%
{I}	7	70%
{K}	7	70%

Step 4: Generate the candidate 2-itemset C_2, $C_2 = L_1 * L_1$. Scan the transaction database and calculate the support of each C_2 itemset. Determine the frequent 2-itemset L_2.

C_2

Itemset	Support Count	Support degree
~~{A,D}~~	~~3~~	~~30%~~
~~{A,E}~~	~~1~~	~~10%~~

{A,I}	5	50%
~~{A,K}~~	~~4~~	~~40%~~
~~{D,E}~~	~~4~~	~~40%~~
{D,I}	5	40%
~~{D,K}~~	~~4~~	~~40%~~
~~{E,I}~~	~~3~~	~~30%~~
~~{E,K}~~	~~4~~	~~40%~~
~~{I,K}~~	~~4~~	~~40%~~

L_2

Itemset	Support Count	Support degree
{A,I}	5	50%

C_3 = null

So, all frequent itemsets: {A},{D},{E},{I},{K},{A,I}.
maximal frequent itemsets:{D},{E},{K},{A,I}.

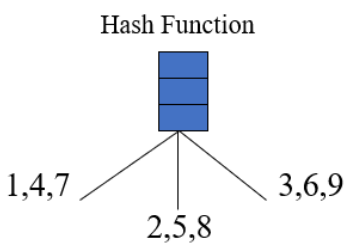
2)
Generate the association rules from frequent itemsets and calculate confidence levels.
confidence(S->(I-S)) = support(I)/support(S)

Frequent Itemset	Subset	Association rules	Support(I)	Support(S)	Confidence	Confidence(%)
{A,I}	{A}	{A}->{I}	5	6	5/6	83.3%
	{I}	~~{I}->{A}~~	5	7	5/7	71.4%

So, all association rules: {A}->{I}.

2. Hash Tree (10 marks)

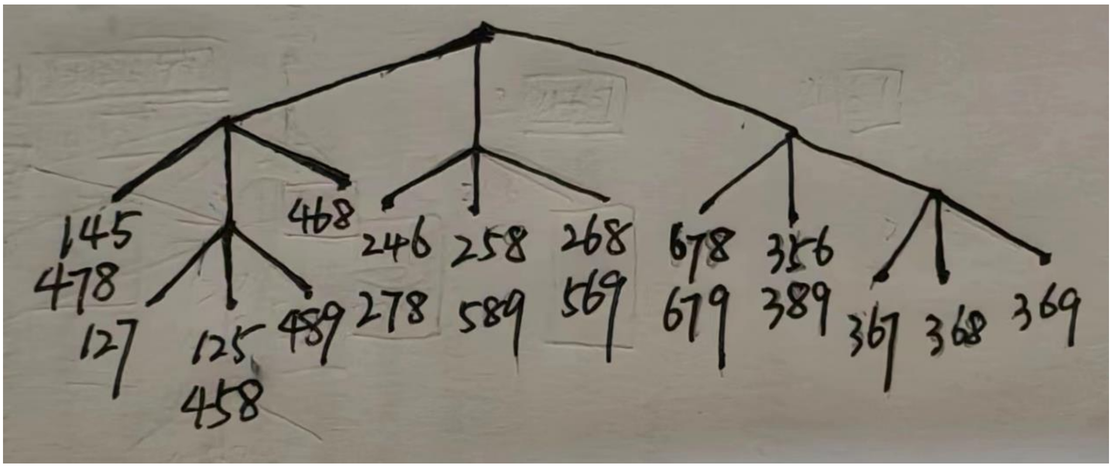
(1) (6 marks) Suppose we have 20 candidate itemsets of length 3:
{1 2 5}, {1 2 7}, {1 4 5}, {2 4 6}, {2 5 8}, {2 6 8}, {2 7 8}, {3 5 6}, {3 6 7}, {3 6 8}, {3 6 9}, {3 8 9}, {4 5 8}, {4 6 8}, {4 7 8}, {4 8 9}, {5 6 9}, {5 8 9}, {6 7 8}, {6 7 9}
Please generate a hash tree with **max leaf size 2**. The hash function is shown in the figure below.



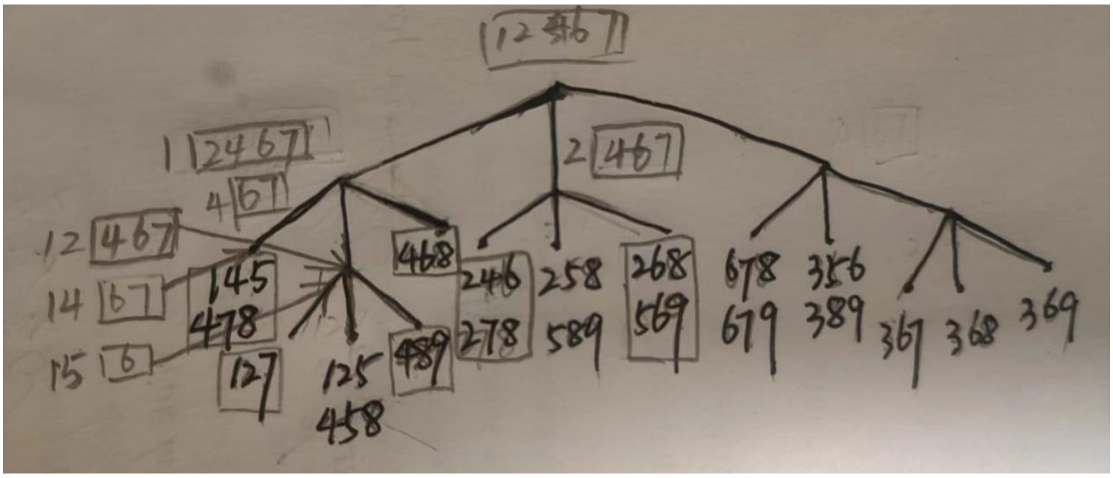
(2) (4 marks) Given a transaction that contains items {1, 2, 4, 6, 7}, Compared to the naïve comparison (20 comparisons), how many comparisons are saved after using the hash tree which you generate above?

Sol:

a)



b)



Match transaction against 9 out of 20 candidates, so save 11 comparisons.

3. FP-Growth (10 marks)

Transaction	Drinks	Name
1	cola, cider	Amy
2	milk, soymilk	Edward
3	Red Bull	Bob
4	vodka, Red Bull	Davis
5	water	Edward
6	cider, cola	Cindy
7	tea	Amy
8	vodka, Red Bull	Cindy
9	cider, water	Davis
10	sprite, vodka, cider	Bob
11	milk	Davis
12	tea, coffee	Edward
13	water	Amy
14	cola, water	Bob
15	milk, water	Cindy
16	soymilk	Amy

The table above is a transaction record of a drinks store. There are five consumers (Amy, Bob, Cindy, Davis, Edward) who used to buy drinks at this store. Do not consider time information. (You could categorize the transaction by consumer.)

(1) (6 marks) The manager of this drinks store wants to find frequent patterns of this transaction record. Suppose the minimum support count is 3 for following questions. Please use FP-Growth to find all frequent patterns and show the major steps.

(2) (4 marks) Please find all maximal and closed frequent itemsets.

Sol:

a)

Step 1: Categorize the transaction by consumer. Scan the transaction database to get the support count of all itemsets.

Amy {cola, tea, water, soymilk, cider}

Bob {Red Bull, sprite, vodka, cider, cola, water}

Cindy {cider, vodka, Red Bull, milk, water, cola}

Davis {vodka, cider, water, milk, Red Bull}

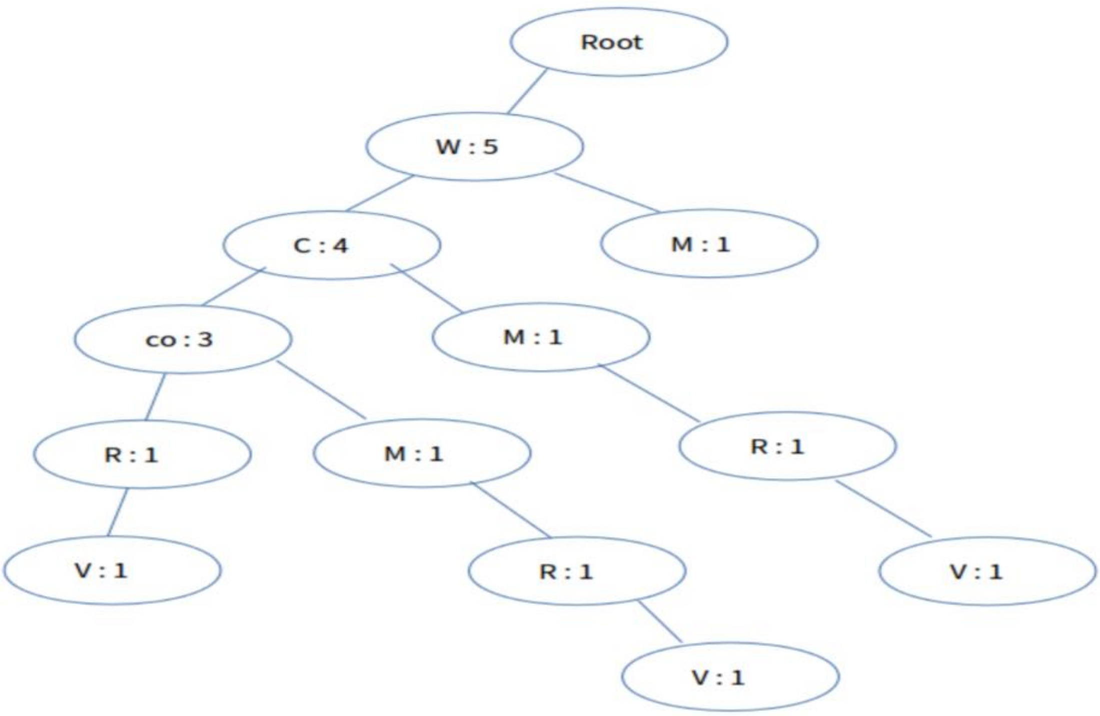
Edward {milk, soymilk, water, tea, coffee}

Itemset	Support Count
water	5
cider	4
cola	3
milk	3
Red Bull	3
vodka	3
~~soymilk~~	~~2~~
~~tea~~	~~2~~

~~coffee~~	~~1~~
~~sprite~~	~~1~~

Step 2: Reorder the frequent itemsets.
Amy {water, cider, cola}
Bob {water, cider, cola, Red Bull, vodka}
Cindy {water, cider, cola, milk, Red Bull, vodka}
Davis{water, cider, milk, Red Bull, vodka}
Edward {water, milk}

Step 3:Construst the FP-tree



Step 4:Find the conditional pattern base for all the node in the head pointer table

Head-Node	Conditional pattern base(Prefix path)	Conditional FP-tree
W	{}5	null
Ci	{W}4	{W}4
Co	{W,Ci}3	{W,Ci}3
R	{W,Ci,Co}1, {W,Ci,Co,M}1, {W,Ci,M}1	{W,Ci}3
M	{W,Ci,Co}1, {W,Ci}1, {W}1	{W}3
V	{W,Ci,Co,R}1, {W,Ci,Co,M,R}1, {W,Ci,M,R}1	{W,Ci,R}3

Step 5: Find frequent itemsets.
1. For W, the cond FR-tree is ϕ(mean null), so all frequent itemsets are {W}.
2. For Ci, the cond FR-tree is $\phi - W{:}\,3$, so all frequent itemsets are {Ci}{W,Ci}.

3. For Co, the cond FR-tree is $\phi - W:3 - Ci:3$, add W, Ci to Co separately.
   3.1 Add W to Co, W's cond FP-tree is ϕ, then {W,Co}.
   3.2 Add Ci to Co, then {Ci,Co}. Since Ci's cond FP-tree is $\phi - W:3$, add W to {Ci,Co}, then {W,Ci,Co}.
   So all frequent itemsets are {Co}{W,Co},{Ci,Co}{W,Ci,Co}.
4. For R, the cond FR-tree is $\phi - W:3 - Ci:3$, add W, Ci to R separately.
   4.1 Add W to R, W's cond FP-tree is ϕ, then{W,R}.
   4.2 Add Ci to R, then {Ci,R}. Since Ci's cond FP-tree is $\phi - W:3$, add W to {Ci,R}, then {W,Ci,R}.
   So all frequent itemsets are {R}{W,R},{Ci,R}{W,Ci,R}.
5. For M, the cond FR-tree is $\phi - W:3$, so all frequent itemsets are {M}{W,M}.
6. For V, the cond FR-tree is $\phi - W:3 - Ci:3 - R:3$, add W, Ci,R to V separately.
   6.1 Add W to V, W's cond FP-tree is ϕ, then {W,V}.
   6.2 Add Ci to V, then {Ci,V}. Since Ci's cond FP-tree is $\phi - W:3$, add W to {Ci,V}, then {W,Ci,V}.
   6.3 Add R to V, then {R,V}. Since R's cond FP-tree is $\phi - W:3 - Ci:3$,
   add W to {R,V}, then {W,R,V},
   add Ci to {R,V}, then {Ci,R,V}.
   So all frequent itemsets are
   {V}{W,V},{Ci,V}{W,Ci,V},{R,V}{W,R,V},{Ci,R,V},{W,Ci,R,V}.

Above all, frequent itemsets:
{V}{W,V},{Ci,V}{W,Ci,V},{R,V}{W,R,V},{Ci,R,V},{W,Ci,R,V}
{R}{W,R},{Ci,R}{W,Ci,R}
{M}{W,M}
{Ci}{W,Ci}
{Co}{W,Co},{Ci,Co}{W,Ci,Co}
{W}

Or
{vodka},{Red Bull, vodka},{vodka, cider},{water, vodka},{water, Red Bull, vodka},
{water, vodka, cider},{Red Bull, vodka, cider},{Red Bull, water, vodka, cider}
{Red Bull}, {Red Bull, water},{Red Bull, cider},{Red Bull, water, cider}
{milk}, {water, milk}
{cider},{water, cider}
{cola}, {cola, cider},{cola, water},{cola,cider, water}
{water}

b) maximal: {water, milk}, {Red Bull, water, vodka, cider}, {water, cider, cola}
close: maximal + {water},{water, cider}

4. Sequence pattern mining with Timing Constraints (10 marks)

If you are a store owner and you want to learn about your customer's buying behavior, you may not only be interested in what they buy together during one shopping trip. You might also want to know about patterns in their purchasing behavior over time. Assume that you have a database full of transactions that looks like this:

Transaction Date	Customer ID	Item Purchased
1	01	b,d
1	02	a
1	05	a
2	01	e
2	02	b
2	03	a,h
2	04	b,d
2	05	b,d
3	02	e
3	03	b,d
3	04	b
3	05	b
4	01	b
4	02	c,a
4	04	e
4	05	e
5	01	a,e
5	02	b
5	03	a
5	05	b
6	02	g
6	03	b
6	04	d
6	05	a,d

Use Generalized Sequential pattern algorithm to find all sequences with $support \geq 0.6$, $x_g = 1$, and show steps (i.e., candidate generation, candidate pruning, support counting and elimination). Note that x_g represents the max-gap.

Sol:

4. custom sequence

01 : < {bd} {e} {b} {a,e} >

02 : < {a} {b} {e} {c,a} {b} {g} >

03 : < {a,h} {b,d} {a} {b} >

04 : < {b,d} {b} {e} {d} >

05 : < {a} {b,d} {b} {e} {b} {a,d} >

min-sup = 5 × 60% = 3

① candidate 1-sequences are:

	< {b} >	< {d} >	< {e} >	< {a} >	~~< {c} >~~	~~< {g} >~~	~~< {h} >~~
sup:	5	4	4	4	1	1	1

candidate pruning remain: < {b} >, < {d} >, < {e} >, < {a} >

② Generate 2-sequences:

< {bd} >, < {be} >, < {ba} >, < {de} >, < {da} >, < {ea} >,

< {b}, {b} >, < {b}, {d} >, < {b} {e} >, < {b} {a} >

< {d}, {d} >, < {d}, {b} >, < {d}, {e} >, < {d}, {a} >

< {e}, {e} >, < {e}, {b} >, < {e}, {d} >, < {e}, {a} >

< {a}, {a} >, < {a}, {b} >, < {a}, {d} >, < {a}, {e} >

Support counting:

	< {bd} >	< {be} >	< {ba} >	< {de} >	< {da} >	< {ea} >
sup	4	0	0	0	0	1
	< {b}, {b} >	< {b}, {d} >	< {b} {e} >	< {b} {a} >		
sup	2	0	4	3		
	< {d}, {d} >	< {d}, {b} >	< {d}, {e} >	< {d}, {a} >		
sup	1	2	1	2		
	< {e}, {e} >	< {e}, {b} >	< {e}, {d} >	< {e}, {a} >		
sup	0	1	2	1		
	< {a}, {a} >	< {a}, {b} >	< {a}, {d} >	< {a}, {e} >		
sup	0	3	2	0		

candidate elimination

< {bd} >, < {b} {e} >, < {a}, {b} >, < {b}, {a} >

③ generate 3-sequences

< {bd}, {e} >, < {a}, {bd} >, < {a}, {b}, {e} >

< {bd}, {a} >, < {a}, {b}, {a} >, < {b}, {a}, {b} >

pruning

< {bd}, {e} > is pruned ← < {d}, {e} > is infrequent

< {a}, {bd} > is pruned ← < {a} {d} > is infrequent

< {bd}, {a} > is pruned ← < {d} {a} > is infrequent.

support counting:

σ(< {a}, {b}, {e} >) = 1, σ(< {b}, {a}, {b} >) = 1

σ(< {a}, {b}, {a} >) = 1

∴ there is no frequent 3-sequences.

5. Decision Tree (15 marks)

ID	Attributes			Label
	Race	Income	Weight	Insurance
1	White	3500	Low	No
2	Black	3200	High	Yes
3	White	2300	High	No
4	White	2500	Low	Yes
5	White	1400	Low	Yes
6	Black	1000	Mid	Yes
7	White	4200	High	No
8	White	6000	Mid	Yes
9	Black	2100	Low	No
10	Black	5000	High	Yes

(1) (7 marks) Try threshold 2000, 3000, and 4000 for attributes Income (that is, use the "Income > 2000, Income < 2000", "Income > 3000, Income < 3000", and "Income > 4000, Income < 4000" respectively). Use the Gini score to determine the best one θ_a among them. Recall

$$Gini(t) := 1 - \sum_{i=1}^{c} [p(i|t)]^2$$

(2) (8 marks) Use θ_a obtained above, and the Gini score, determine which attributes should firstly be used for developing a decision tree.

Sol:

5. (1)

For $\theta_a = 2000$

	< 2000	> 2000
No	0	4
Yes	2	4

$I(\text{Income} < 2000) = 1 - 0 - (\frac{2}{2})^2 = 0$

$I(\text{Income} > 2000) = 1 - (\frac{1}{2})^2 - (\frac{1}{2})^2 = \frac{1}{2}$

weighted Gini $= \frac{2}{10} \times 0 + \frac{8}{10} \times \frac{1}{2} = 0.4$

For $\theta_a = 3000$

	< 3000	> 3000
No	2	2
Yes	3	3

$I(\text{Income} < 3000) = 1 - (\frac{2}{5})^2 - (\frac{3}{5})^2 = \frac{12}{25}$

$I(\text{Income} > 3000) = 1 - (\frac{2}{5})^2 - (\frac{3}{5})^2 = \frac{12}{25}$

weighted Gini $= \frac{5}{10} \times \frac{12}{25} + \frac{5}{10} \times \frac{12}{25} = 0.48$

For $\theta_a = 4000$

	< 4000	> 4000
No	3	1
Yes	4	2

$I(\text{Income} < 4000) = 1 - (\frac{3}{7})^2 - (\frac{4}{7})^2 = \frac{24}{49}$

$I(\text{Income} > 4000) = 1 - (\frac{1}{3})^2 - (\frac{2}{3})^2 = \frac{4}{9}$

weighted Gini $= \frac{7}{10} \times \frac{24}{49} + \frac{3}{10} \times \frac{4}{9} \approx \boxed{0.4762}$

$\therefore$ we choose $\theta_a = 2000$

(2) For Race:

	W	B
Yes	3	3
No	3	1

$I(\text{Race} = \text{white}) = 1 - (\frac{1}{2})^2 - (\frac{1}{2})^2 = \frac{1}{2}$

$I(\text{Race} = \text{Black}) = 1 - (\frac{3}{4})^2 - (\frac{1}{4})^2 = \frac{3}{8}$

weighted Gini $= \frac{6}{10} \times \frac{1}{2} + \frac{4}{10} \times \frac{3}{8} = \boxed{0.45}$

For Income: From (1), we know weighted Gini $= \boxed{0.4}$

For weight:

	Low	High	Mid
Yes	2	2	2
No	2	2	0

$I(\text{weight} = \text{Low}) = 1 - (\frac{1}{2})^2 - (\frac{1}{2})^2 = \frac{1}{2}$

$I(\text{weight} = \text{High}) = 1 - (\frac{1}{2})^2 - (\frac{1}{2})^2 = \frac{1}{2}$

$I(\text{weight} = \text{Mid}) = 1 - (\frac{2}{2})^2 - 0 = 0$

weighted Gini $= \frac{4}{10} \times \frac{1}{2} + \frac{4}{10} \times \frac{1}{2} + \frac{2}{10} \times 0 = \boxed{0.4}$

Therefore, Income or Weight can be choosed.

6. Naïve Bayesian Classifier (15 marks)

Consider a company that is recruiting staff, you’re the HR of this company, and you are given the following examples from the company’s hiring record:

✧ EXP: Yes if the applicant had prior experience in related jobs.

ID	Gender	Age	GPA	EXP	Hired
1	Male	25~30	Medium	Yes	Yes
2	Male	25~30	High	Yes	No
3	Male	25~30	High	Yes	Yes

4	Female	30~35	Medium	Yes	Yes
5	Female	20~25	Medium	Yes	Yes
6	Female	20~25	Medium	Yes	Yes
7	Male	25~30	Medium	No	No
8	Male	25~30	Low	No	No
9	Female	25~30	Low	No	Yes
10	Female	25~30	High	Yes	Yes
11	Female	20~25	Low	Yes	Yes
12	Male	20~25	Medium	Yes	No
13	Male	20~25	Low	No	No
14	Male	20~25	Medium	No	No
15	Male	30~35	Medium	No	Yes

(1) (12 marks) Now you are asked to use the above data and the Naïve Bayes classifier to infer whether the candidate X should be hired. The information of candidate X is: $X = (Gender = Female,\ age = 26,\ GPA = Medium,\ EXP = Yes)$. **Please apply Laplacian correction to each conditional probability.**

(2) (3 marks) Please briefly describe why Naïve Bayesian Classifier is called “naïve”.

Sol:

(1)

$$P(H=Yes \mid \mathbf{X}) \propto P(Gender=Female \mid H=Yes)\, P(age \in (25,30] \mid H=Yes)\, P(GPA=Medium \mid H=Yes)$$
$$\cdot P(EXP=Yes \mid H=Yes) \cdot P(H=Yes)$$
$$= \frac{6+1}{9+2} \times \frac{4+1}{9+3} \times \frac{5+1}{9+3} \times \frac{7+1}{9+2} \times \frac{9}{15} \approx 0.0579$$

$$P(H=No \mid X) \propto P(Gender=Female \mid H=No)\, P(age \in (25,30] \mid H=No)\, P(GPA=Medium \mid H=No)$$
$$P(EXP=Yes \mid H=No) \cdot P(H=No)$$
$$= \frac{0+1}{6+2} \times \frac{3+1}{6+3} \times \frac{3+1}{6+3} \times \frac{2+1}{6+2} \times \frac{6}{15} \approx 0.0037$$

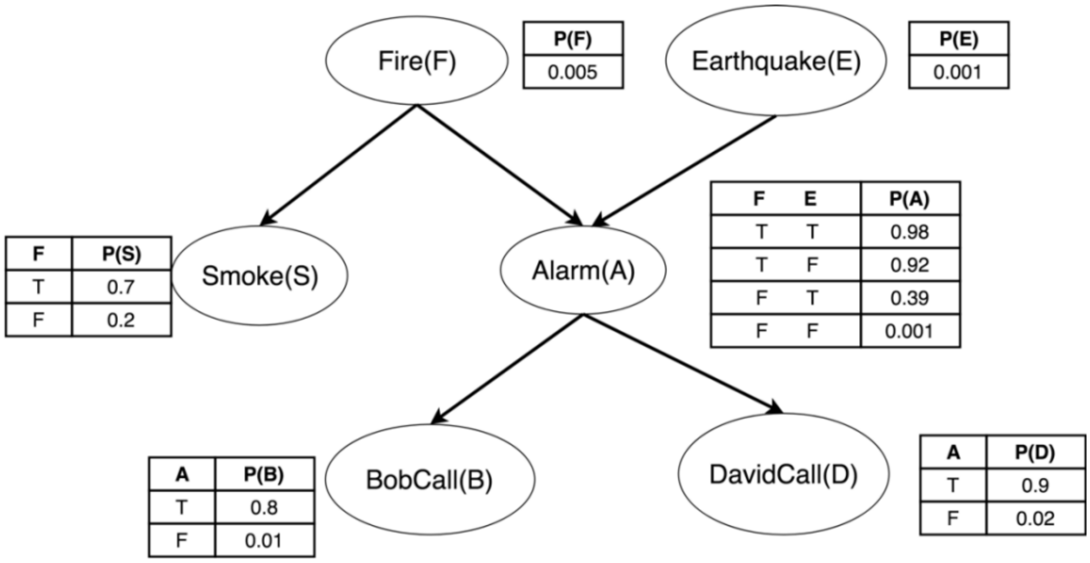
$\therefore$ should hire.

(2)

For Naïve Bayesian Classifier, there is an independent assumption that each attribute are independent. This greatly reduces the computation cost: Only counts the class distribution. Its idea is very simple, the posterior probability is calculated from some prior probability

7. Bayesian Network (10 marks)

Below is a Bayesian network and its conditional probability tables.



(1) (6 marks) Please compute the probability of BobCall(B) given DavidCall(D): $P(B|D)$.

(2) (4 marks) Please compute the probability of DavidCall(D) given Fire(F): $P(D|F)$.

Sol:

(a)

$$P(A) = P(A|FE)P(F)P(E) + P(A|F\bar{E})P(F)P(\bar{E}) + P(A|\bar{F}E)P(\bar{F})P(E) + P(A|\bar{F}\bar{E})P(\bar{F})P(\bar{E}) = 0.0060$$

$$P(\bar{A}) = 1 - P(A) = 1 - 0.0060 = 0.9940$$

$$P(D) = P(D|A)P(A) + P(D|\bar{A})P(\bar{A}) = 0.0253$$

$$P(B|D) = \frac{P(BD)}{P(D)} = \frac{P(BDA) + P(BD\bar{A})}{P(D)} = \frac{P(BD|A)P(A) + P(BD|\bar{A})P(\bar{A})}{P(D)}$$

$$= \frac{P(B|A)P(D|A)P(A) + P(B|\bar{A})P(D|\bar{A})P(\bar{A})}{P(D)}$$

[when given A, B and D are conditionally independent]

$$= \frac{0.8 * 0.9 * 0.006 + 0.01 * 0.02 * 0.994}{0.0253} = 0.1786$$

(b)

$$P(A|F) = P(A|FE)P(E) + P(A|F\bar{E})P(\bar{E}) = 0.9201$$

$$P(\bar{A}|F) = 1 - P(A|F) = 0.0799$$

$$P(D|F) = P(DA|F) + P(D\bar{A}|F) = \frac{P(DAF)}{P(AF)} * \frac{P(AF)}{P(F)} + \frac{P(D\bar{A}F)}{P(\bar{A}F)} * \frac{P(\bar{A}F)}{P(F)}$$

$$= P(D|AF)P(A|F) + P(D|\bar{A}F)P(\bar{A}|F)$$

$= P(D|A)P(A|F) + P(D|\bar{A})P(\bar{A}|F)$ [when given A, D and F are conditionally independent]

$$= 0.9 * 0.9201 + 0.02 * 0.0799 = 0.8297$$

8. Neural Networks (20 marks)

Consider a simple neural network with two hidden layers. The input layer consists of three dimensional $\boldsymbol{x} = (x1, x2, x3)^T$. The hidden layer includes two dimensional $\boldsymbol{h} = (h1, h2)$. The output layer includes one scalar o. We ignore bias terms for simplicity.

We use linear rectified (ReLU) as activation function **for all hidden and output layers**.

$$ReLU(x) = \max(0, x)$$
$$ReLU'(x) = \begin{cases} 1, & x > 0 \\ 0, & x \le 0 \end{cases}$$

Moreover, denote the loss function (also called $error$ in slides) by $\boldsymbol{J}(o, t) = \frac{1}{2}|o - t|^2$ where t is the associated label (target) value for scalar output o.

Represented by weight matrices $W_1,\ W_2\ and\ V$, which connect the input to the first hidden layer, the first hidden layer to the second hidden layer, and the second layer to the output, respectively. They are **initialized** (i.e., the initial condition before first updating round) as follows:

$$W_1 = \begin{bmatrix} 2 & 0 & 1 \\ -2 & -1 & 0 \end{bmatrix}, W_2 = \begin{bmatrix} 1 & 2 \\ -1 & 0 \end{bmatrix}, V = \begin{bmatrix} 1 & 1 \end{bmatrix}$$

Now, try to solve the following parts.

(1) (3 marks) Write out symbolically (thus, no need to plug in the specific values of $W_1,\ W_2\ and\ V$ just yet) the mapping $x \to o$ using ReLU, $W_1, W_2\ and\ V$.

(2) (5 marks) Given the condition $\boldsymbol{x} = (1, 2, 1)^T$, $t = 1$, compute the numerical output value o, clearly show all intermediate steps. You can reuse the results of the previous question.

(3) (12 marks) Compute the gradient of the loss function with respect to the W_2 weights, and evaluate the gradients at specific $\boldsymbol{x} = (1, 2, 1)^T$, $t = 1$. You can reuse the results of the previous question.

Sol:

(1) $$O = \mathrm{ReLU}(\underset{1\times 2}{V}(\mathrm{ReLU}(\underset{2\times 2}{W_2} \cdot \mathrm{ReLU}(\underset{2\times 3}{W_1} \underset{3\times 1}{x}))))$$

(2) $$h_1 = \mathrm{ReLU}(W_1 x)$$
$$h_2 = \mathrm{ReLU}(W_2 h_1)$$
$$O = \mathrm{ReLU}(V \cdot h_2)$$

$$\Rightarrow h_1 = \mathrm{ReLU}\left(\begin{bmatrix} 2 & 0 & 1 \\ -2 & -1 & 0 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}\right) = \mathrm{ReLU}\left(\begin{bmatrix} 3 \\ -4 \end{bmatrix}\right)$$
$$= \begin{bmatrix} 3 \\ 0 \end{bmatrix}$$

$$h_2 = \mathrm{ReLU}\left(\begin{bmatrix} 1 & 2 \\ -1 & 0 \end{bmatrix}\begin{bmatrix} 3 \\ 0 \end{bmatrix}\right) = \mathrm{ReLU}\left(\begin{bmatrix} 3 \\ -3 \end{bmatrix}\right) = \begin{bmatrix} 3 \\ 0 \end{bmatrix}$$
$$O = \mathrm{ReLU}\left(\begin{bmatrix} 1 & 1 \end{bmatrix}\begin{bmatrix} 3 \\ 0 \end{bmatrix}\right) = 3$$

(3) Rewrite $$\underset{2\times 1}{a} = \underset{2\times 2}{W_2}\underset{2\times 1}{h_1}$$
$$\underset{2\times 1}{h_2} = \mathrm{ReLU}(\underset{2\times 1}{a})$$
$$O = \mathrm{ReLU}(\underset{1\times 2}{V} \cdot \underset{2\times 1}{h_2})$$

$$\therefore \frac{\partial J}{\partial W_2} = \frac{\partial J}{\partial O} \cdot \frac{\partial O}{\partial h_2} \cdot \frac{\partial h_2}{\partial a} \cdot \frac{\partial a}{\partial W_2}$$

$$= \underset{1\times 1}{(O-t)} \cdot \left(\underset{2\times 1}{\frac{\partial O}{\partial h_2}} \odot \underset{2\times 1}{\frac{\partial h_2}{\partial a}}\right) \cdot \underset{1\times 2}{\frac{\partial a}{\partial W_2}}$$

$$= (O-t) \cdot \left(\frac{\partial O}{\partial h_2} \odot \frac{\partial h_2}{\partial a}\right) \cdot h_1^T$$

$$= 2 \times \begin{bmatrix} 1 \\ 1 \end{bmatrix} \odot \begin{bmatrix} 1 \\ 0 \end{bmatrix} \times \begin{bmatrix} 3 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 2 \\ 2 \end{bmatrix} \odot \begin{bmatrix} 1 \\ 0 \end{bmatrix} \times \begin{bmatrix} 3 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 2 \\ 0 \end{bmatrix} \times \begin{bmatrix} 3 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 6 & 0 \\ 0 & 0 \end{bmatrix}$$