THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

MSCBDT 5002:

Fall 2018 Midterm Examination

Date: Oct 23rd, 2018

Time: 7:30-10:30pm

This exam contains **10** questions in **17** pages (not including the cover page). Please count the pages.
You have **3** hours to complete this exam.

Problem	Your Points	Max Points
1		6
2		10
3		15
4		5
5		10
6		10
7		12
8		10
9		10
10		12
Total		

Name:	
Student ID:	

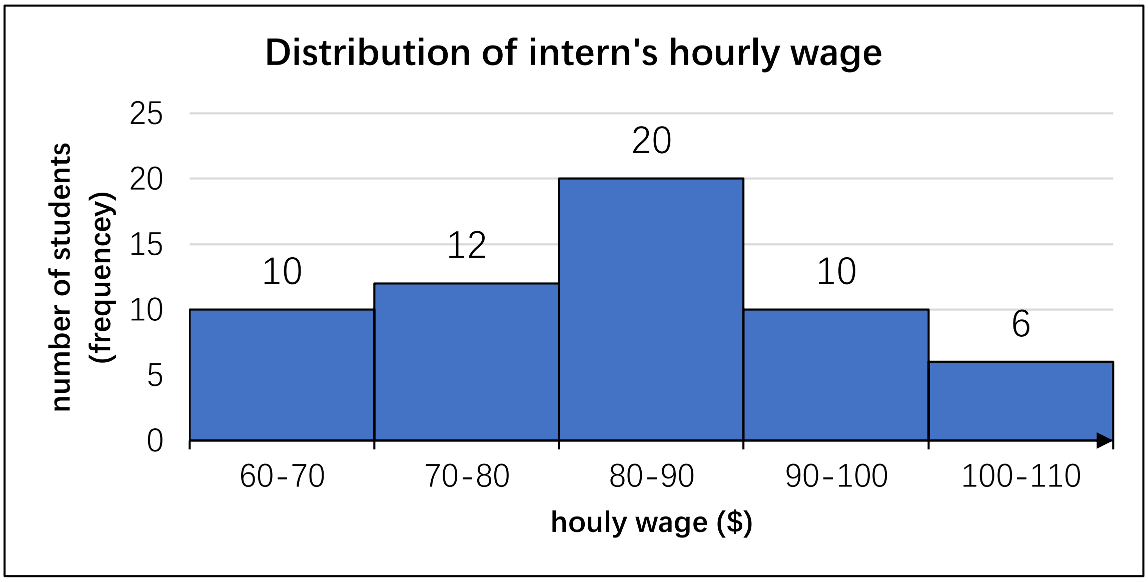
I have neither given nor received any unauthorized aid during this examination. The answers submitted are my own work.

I understand that sanctions will be imposed, if I am found to have violated the University's regulations governing academic integrity.

Signature____

1. Basic statistical description of data (6 marks)

In order to know about how much an information technology intern makes in Hong Kong, the survey collect data from students in a computer science class. The construction of histograms entails grouping data together into class for better visual presentation as shown below:



Please calculate the "best" estimation for the mean and median of hourly wage. (accurate to the second decimal place.)

Answer:

1. mean $= \dfrac{65\times10+75\times12+85\times20+95\times10+105\times6}{10+12+20+10+6} = 83.28$ (assume the data is uniformly spread within each interval)

median $= 80 + \dfrac{58/2-(10+12)}{20}\times10 = 83.5$ (estimated by interpolation)

2. 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.)

a) 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. [6 marks]

b) Please find all maximal and closed frequent itemsets. [4 marks]

Answer:

(a) Categorize the transaction by consumer:

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}

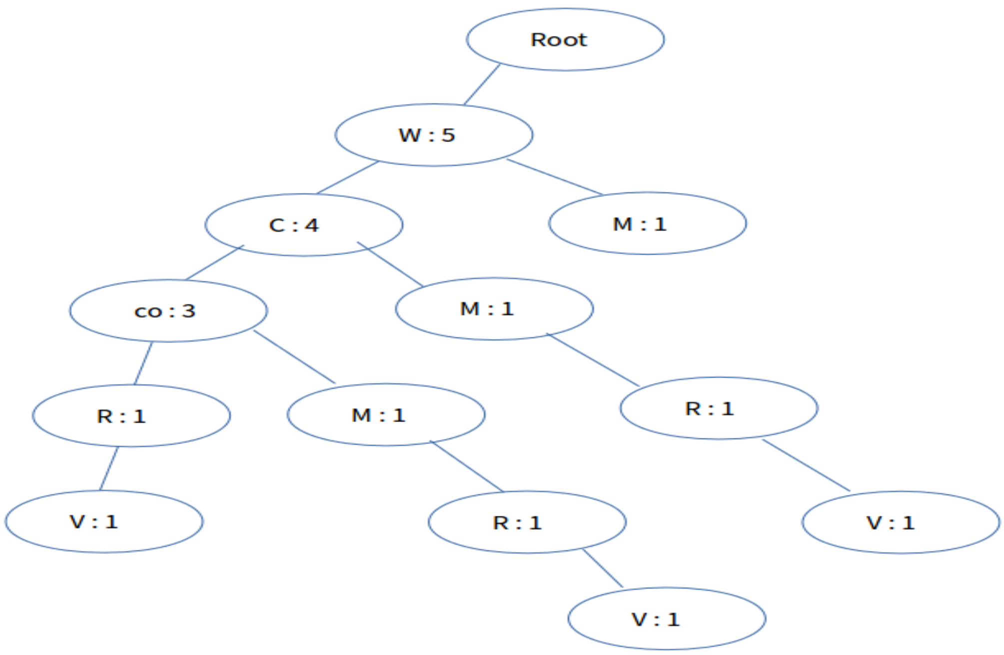
water	5
cider	4
cola	3
milk	3
Red Bull	3
vodka	3
soymilk	2
tea	2
coffee	1
sprite	1

Reorder:

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}



frequent patterns: (21 sets)

{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}

3. Min-Apriori (15 Marks)

E-commerce has become a trend in the future. E-commerce companies like Amazon and Taobao will mine user purchase records to find some useful association rules to improve their future promotion strategies. Suppose we have 10 user purchase records as shown in the following table. Each row represents the purchase amount of A, B, C, D, E by one user during the month. Please complete the following questions:

TID	A	B	C	D	E
1	7	2	6	8	2
2	1	0	8	1	8
3	2	2	5	0	9
4	7	10	4	8	6
5	0	2	8	0	8
6	1	2	7	1	18
7	1	0	8	2	7
8	0	2	15	1	1
9	1	3	7	3	10
10	3	4	8	4	7

* Please round the answers to 3 decimals.
* Min-Support = 0.55

a) Please use Min-Apriori algorithm to find all frequent itemsets and show the major steps (8 Marks)
b) Try to compute the number of candidates you saved in a), comparing with Brute-force approach. (5 Marks)
c) Why you should not simply convert this matrix into 0/1 matrix (1 Marks) or discrete this matrix (1 Marks) then apply the traditional Apriori algorithm.

Answer:

a) **(8 Marks)**

1st : Normalization:

TI0. D	A	B	C	D	E
1	0.304348	0.074074	0.078947	0.285714	0.026316
2	0.043478	0	0.105263	0.035714	0.105263
3	0.086957	0.074074	0.065789	0	0.118421
4	0.304348	0.37037	0.052632	0.285714	0.078947
5	0	0.074074	0.105263	0	0.105263
6	0.043478	0.074074	0.092105	0.035714	0.236842
7	0.043478	0	0.105263	0.071429	0.092105
8	0	0.074074	0.197368	0.035714	0.013158

9	0.043478	0.111111	0.092105	0.107143	0.131579
10	0.130435	0.148148	0.105263	0.142857	0.092105

2nd : Calculate the corresponding supports:

$$\sup(C) = \sum_{i \in T} \min_{j \in C} D(i, j)$$

sup(A) = sup(B) = sup(C) = sup(D) = sup(E) = 1

sup(A,B) = 0.669887 = 0.670
sup(A,C) = 0.476543 = 0.477
sup(A,D) = 0.860247 = 0.860
sup(A,E) =0.458237 = 0.458

sup(B,C) = 0.612085 = 0.612
sup(B,D) =0.681216 = 0.681
sup(B,E) = 0.543859 = 0.544

sup(C,D) = 0.507518 = 0.508
sup(C,E) =0.736841 = 0.737
sup(D,E) =0.460526 = 0.461

sup(A,B,D) = 0.569415 = 0.569

So Frequent item sets:

(A), (B), (C), (D), (E), (A,B),(A,D),(B,C),(B,D) ,(C,E) ,(A,B,D)

b) **(5 Marks)**

Brute-force approach = ${}^5C_1 + {}^5C_2 + {}^5C_3$ = 5+10+10 = 25(According to PPT's example)

Or = ${}^5C_1 + {}^5C_2 + {}^5C_3 + {}^5C_4 + {}^5C_5$ = 31
Or = 2^5 = 32 (including NULL Node)

Min_AP = 5 + 10 + 1 = 16
The number of candidates you saved = 25 – 16 = 9
Or = 31 – 16 = 15
Or = 32 – 16 = 16

c) **(2 Marks)**

- Convert into 0/1 matrix and then apply existing algorithms
    - lose word frequency information **(1 Marks)**
- Discretization does not apply as users want association among goods not ranges of goods **(1 Marks)**

4. Rule Generation (5 Marks)

Please generate the corresponding association rules, According to the result of the Previous question (**Min-Apriori a).)**

* Please round the answers to 3 decimals.
* Min-Confidence = 0.8

Answer:

(A), (B), (C), (D), (E), (A,B),(A,D),(B,C),(B,D) ,(C,E) ,(A,B,D)

Conf(A->B) = Conf (B->A)= 0.669887 = 0.670
Conf (A->D) = Conf (D->A) = 0.860247 = 0.860
Conf (B->C) = Conf (C->B) = 0.612085 = 0.612
Conf (B->D) = Conf (D->B) = 0.681216 = 0.681
Conf (C->E) = Conf (E->C) = 0.736841 = 0.737

Conf (B,D->A) = sup(A,B,D)/ sup(B,D) = 0.569415/0.681216 = 0.83588024943 = 0.836
Conf (A,D->B) = sup(A,B,D)/ sup(A,D) = 0.569415/0.860247 = 0.66192035543 = 0.662
Conf (A,B->D) = sup(A,B,D)/ sup(A,B) = 0.569415/0.669887 = 0.85001649531 = 0.850

Rules : (A->D), (D->A) ,(B,D->A), (A,B->D)

5. Sequence pattern mining (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. If a customer purchases baby lotion, then a new-born blanket, what are they likely to buy next? 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 ≥ 0.8 and show steps.

Answer:

2. custom sequence:

01: {b,d}, {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 × 80% = 4

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}>

Candidate 2-sequences:

	<{bd}>	<{be}>	<{ba}>	<{de}>	<{da}>	<{b},{d}>	<{b},{e}>
sup:	4	0	0	0	0	2	4

	<{b},{a}>	<{d},{b}>	<{d},{e}>	<{d},{a}>	<{e},{b}>	<{e},{d}>
sup:	4	4	3	3	3	2

	<{e},{a}>	<{a},{b}>	<{a},{d}>	<{a},{e}>	<{b},{b}>	<{d},{d}>
sup:	3	3	2	2	5	2

	<{e},{e}>	<{a},{a}>	<{ea}>
sup:	1	3	1

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

candidate: 3-sequences are:

<{bd},{b}>	<{d},{bd}>	<{d},{b},{b}
4	0	1

<{b},{bd}>	<{b},{b},{e}>	<{b},{b},{a}>
0	3	2

After candidate pruning: <{bd},{b}>

∴ All candidates: <{b}>, <{d}>, <{e}>, <{a}>, <{bd}>, <{b},{e}>, <{b},{a}>, <{d},{b}>, <{b},{b}>, <{bd},{b}>

6. KNN (10 marks)

a) State the Pros/Cons to KNN algorithm. (4 marks)

b) A good value for K can be determined by considering a range of K values. State the impact of the range of K values. (4 marks)

c) What is the accuracy on the training data? Why? (2 marks)

Answer:

(a)

Pros:

- Simple and powerful. No need for tuning complex parameters to build a model.
- No training involved ("lazy"). New training examples can be added easily.

Cons:

- Expensive and slow: To determine the nearest neighbor of a new point x, must compute the distance to all m training examples. Runtime performance is slow.
- Hard to determine K.

(b)

K too small: we'll model the noise

K too large: neighbors include too many points from other classes

(c)

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

If K = 1, accuracy is 100%.

Distance to self is zero.

7. Decision tree (12 marks)

The target classification is "should we play baseball?" which can be yes or no. Please construct the decision tree using ID3 to determine the answer. Please write the detailed calculation process and draw the decision tree.

The weather attributes are outlook, temperature, humidity, and wind speed. They can have the following values:

outlook = {sunny, overcast, rain}

temperature = {hot, mild, cool}

humidity = {high, normal}

wind = {weak, strong}

Examples of set S are:

#	outlook	temperature	humidity	wind	answer
0	sunny	hot	high	weak	no
1	sunny	hot	high	strong	no
2	overcast	hot	high	weak	yes
3	rain	mild	high	weak	yes
4	rain	cool	normal	weak	yes
5	rain	cool	normal	strong	no
6	overcast	cool	normal	strong	yes
7	sunny	mild	high	weak	yes
8	sunny	cool	normal	weak	yes
9	rain	mild	normal	weak	yes
10	sunny	mild	normal	strong	yes
11	overcast	mild	high	strong	yes
12	overcast	hot	normal	weak	yes
13	rain	mild	high	strong	no

Answer:

For node root

S =

#	outlook	temperature	humidity	wind	answer
0	sunny	hot	high	weak	no
1	sunny	hot	high	strong	no
2	overcast	hot	high	weak	yes
3	rain	mild	high	weak	yes
4	rain	cool	normal	weak	yes

5	rain	cool	normal	strong	no
6	overcast	cool	normal	strong	yes
7	sunny	mild	high	weak	yes
8	sunny	cool	normal	weak	yes
9	rain	mild	normal	weak	yes
10	sunny	mild	normal	strong	yes
11	overcast	mild	high	strong	yes
12	overcast	hot	normal	weak	yes
13	rain	mild	high	strong	no

For attribute outlook
E(S,overcast) = 0

E(S,rain) = -2/5×log_2(2/5)-3/5×log_2(3/5) = 0.97

E(S,sunny) = -2/5×log_2(2/5)-3/5×log_2(3/5) = 0.97

E(S,outlook) = 4/14×E(S,overcast)+5/14×E(S,rain)+5/14×E(S,sunny) = 0.69

E(S) = -4/14×log_2(4/14)-10/14×log_2(10/14) = 0.86

G(root,outlook) = E(S) - E(S,outlook) = 0.17
For attribute temperature

E(S,cool) = -1/4×log_2(1/4)-3/4×log_2(3/4) = 0.81

E(S,hot) = -2/4×log_2(2/4)-2/4×log_2(2/4) = 1.0

E(S,mild) = -1/6×log_2(1/6)-5/6×log_2(5/6) = 0.65

E(S,temperature) = 4/14×E(S,cool)+4/14×E(S,hot)+6/14×E(S,mild) = 0.8

E(S) = -4/14×log_2(4/14)-10/14×log_2(10/14) = 0.86

G(root,temperature) = E(S) - E(S,temperature) = 0.07
For attribute humidity

E(S,high) = -3/7×log_2(3/7)-4/7×log_2(4/7) = 0.99

E(S,normal) = -1/7×log_2(1/7)-6/7×log_2(6/7) = 0.59

E(S,humidity) = 7/14×E(S,high)+7/14×E(S,normal) = 0.79

E(S) = -4/14×log_2(4/14)-10/14×log_2(10/14) = 0.86

G(root,humidity) = E(S) - E(S,humidity) = 0.07
For attribute wind

E(S,strong) = -3/6×llog_2(3/6)-3/6×log_2(3/6) = 1.0

E(S,weak) = -1/8×log_2(1/8)-7/8×log_2(7/8) = 0.54

E(S,wind) = 6/14×E(S,strong)+8/14×E(S,weak) = 0.74

E(S) = -4/14×log_2(4/14)-10/14×log_2(10/14) = 0.86

G(root,wind) = E(S) - E(S,wind) = 0.12

outlook attribute has the highest gain, therefore it is used as the decision node

....

For node overcast

S =

temperature	humidity	wind	answer
hot	high	weak	yes
cool	normal	strong	yes
mild	high	strong	yes
hot	normal	weak	yes

It has been divided purely, therefore it has no branch. All the item divided to this node belong to yes

....

For node rain

S =

temperature	humidity	wind	answer
mild	high	weak	yes
cool	normal	weak	yes
cool	normal	strong	no
mild	normal	weak	yes
mild	high	strong	no

For attribute temperature

E(S,cool) = -1/2×log_2(1/2)-1/2×log_2(1/2) = 1.0

E(S,mild) = -1/3×log_2(1/3)-2/3×log_2(2/3) = 0.92

E(S,temperature) = 2/5×E(S,cool)+3/5×E(S,mild) = 0.95

E(S) = -2/5×log_2(2/5)-3/5×log_2(3/5) = 0.97

G(rain,temperature) = E(S) - E(S,temperature) = 0.02

For attribute humidity

E(S,high) = -1/2×log_2(1/2)-1/2×log_2(1/2) = 1.0

E(S,normal) = -1/3×log_2(1/3)-2/3×log_2(2/3) = 0.92

E(S,humidity) = 2/5×E(S,high)+3/5×E(S,normal) = 0.95

$E(S) = -2/5 \times log_2(2/5) - 3/5 \times log_2(3/5) = 0.97$

G(rain,humidity) = E(S) - E(S,humidity) = 0.02

For attribute wind

E(S,strong) = 0

E(S,weak) = 0

E(S,wind) = 2/5×E(S,strong)+3/5×E(S,weak) = 0.0

$E(S) = -2/5 \times log_2(2/5) - 3/5 \times log_2(3/5) = 0.97$

G(rain,wind) = E(S) - E(S,wind) = 0.97

wind attribute has the highest gain, therefore it is used as the decision node

....

For node strong

S =

temperature	humidity	answer
cool	normal	no
mild	high	no

It has been divided purely, therefore it has no branch. All the items divided to this node belong to no

....

For node weak

S =

temperature	humidity	answer
mild	high	yes
cool	normal	yes
mild	normal	yes

It has been divided purely, therefore it has no branch. All the items divided to this node belong to yes

....

For node sunny

S =

temperature	humidity	wind	answer
hot	high	weak	no
hot	high	strong	no
mild	high	weak	yes
cool	normal	weak	yes
mild	normal	strong	yes

For attribute temperature

E(S,cool) = 0

E(S,hot) = 0

E(S,mild) = 0

$E(S,temperature) = 1/5 \times E(S,cool)+2/5 \times E(S,hot)+2/5 \times E(S,mild) = 0$

$E(S) = -2/5 \times log_2(2/5)-3/5 \times log_2(3/5) = 0.97$

$G(sunny,temperature) = E(S) - E(S,temperature) = 0.97$

For attribute humidity

$E(S,high) = -2/3 \times log_2(2/3)-1/3 \times log_2(1/3) = 0.92$

$E(S,normal) = 0$

$E(S,humidity) = 3/5 \times E(S,high)+2/5 \times E(S,normal) = 0.55$

$E(S) = -2/5 \times log_2(2/5)-3/5 \times log_2(3/5) = 0.97$

$G(sunny,humidity) = E(S) - E(S,humidity) = 0.42$

For attribute wind

$E(S,strong) = -1/2 \times log_2(1/2)-1/2 \times log_2(1/2) = 1.0$

$E(S,weak) = -1/3 \times log_2(1/3)-2/3 \times log_2(2/3) = 0.92$

$E(S,wind) = 2/5 \times E(S,strong)+3/5 \times E(S,weak) = 0.95$

$E(S) = -2/5 \times log_2(2/5)-3/5 \times log_2(3/5) = 0.97$

$G(sunny,wind) = E(S) - E(S,wind) = 0.02$

temperature attribute has the highest gain, therefore it is used as the decision node

....

For node cool

S =

humidity	wind	answer
normal	weak	yes

It has been divided purely, therefore it has no branch. All the items divided to this node belong to yes

....

For node hot

S =

humidity	wind	answer
high	weak	no
high	strong	no

It has been divided purely, therefore it has no branch. All the items divided to this node belong to no

....

For node mild

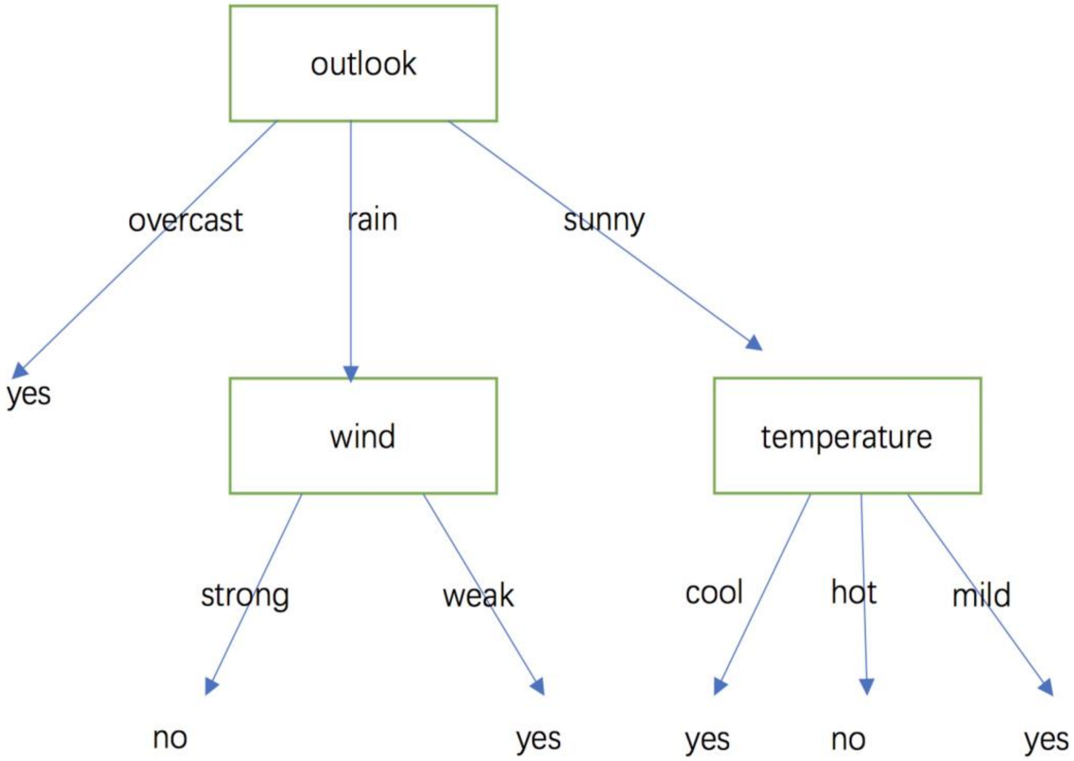
S =

humidity	wind	answer
high	weak	yes

normal	strong	yes

It has been divided purely, therefore it has no branch. All the items divided to this node belong to yes.

The decision tree as follow:



8. Naïve Bayesian Classifier (10 marks)

a) 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	Degree	GPA	EXP	Hired
1	Male	25 ~ 30	Bachelor	Medium	Yes	Yes
2	Female	25 ~ 30	Master	High	Yes	Yes
3	Female	30 ~ 35	Bachelor	High	Yes	Yes
4	Male	20 ~ 25	Master	Low	No	No
5	Female	20 ~ 25	Bachelor	Medium	No	Yes
6	Male	25 ~ 30	Master	Medium	No	No
7	Male	25 ~ 30	Master	Medium	No	No
8	Male	20 ~ 25	Bachelor	Low	No	No
9	Male	25 ~ 30	Master	Medium	No	Yes
10	Male	20 ~ 25	Bachelor	Low	No	No
11	Female	25 ~ 30	Master	Medium	Yes	Yes
12	Male	25 ~ 30	Master	Medium	No	No
13	Male	20 ~ 25	Bachelor	Medium	No	Yes
14	Male	30 ~ 35	Master	Low	No	No
15	Female	20 ~ 25	Bachelor	Medium	No	No

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:

$$X = (Gender = Female, age = 26, Degree = Bachelor, GPA = Medium, EXP = Yes).$$

[7 marks]

b) Please briefly describe why Naïve Bayesian Classifier is called "naïve". [3 marks]

Answer:

(a)

$$P(Hired = Yes) = \frac{7}{15}$$

$$P(Hired = No) = \frac{8}{15}$$

$$P(Female|Hired = Yes) = \frac{4}{7}$$

$$P(Female|Hired = No) = \frac{1}{8}$$

$$P\left(age = 25 \sim 30|Hired = Yes\right) = \frac{4}{7}$$

$$P\left(age = 25 \sim 30|Hired = No\right) = \frac{3}{8}$$

$$P(Bachelor|Hired = Yes) = \frac{4}{7}$$

$$P(Bachelor|Hired = No) = \frac{3}{8}$$

$$P(GPA = Medium|Hired = Yes) = \frac{5}{7}$$

$$P(GPA = Medium|Hired = No) = \frac{4}{8}$$

$$P(EXP = Yes|Hired = Yes) = \frac{4}{7}$$

$$P(EXP = Yes|Hired = No) = 0$$

For Naïve Bayesian Classifier, each conditional probability can't be zero. For samples with $Hired = No$, there are 0 $EXP = Yes$ and 8 $EXP = No$, we add 0.01 to them respectively, then we have 0.01 $EXP = Yes$, 8.01 $EXP = No$ we can get:

$$P(EXP = Yes|Hired = No) = \frac{0.01}{0.01 + 8.01} = 0.0012$$

$$P(X, Hired = Yes) = P(X|Hired = Yes)P(Hired = Yes) = \frac{4}{7} * \frac{4}{7} * \frac{4}{7} * \frac{5}{7} * \frac{4}{7} * \frac{7}{15} = 0.0355$$

$$P(X, Hired = No) = P(X|Hired = No)P(Hired = No) = \frac{1}{8} * \frac{3}{8} * \frac{3}{8} * \frac{4}{8} * 0.0012 * \frac{8}{15} = 0.000005625$$

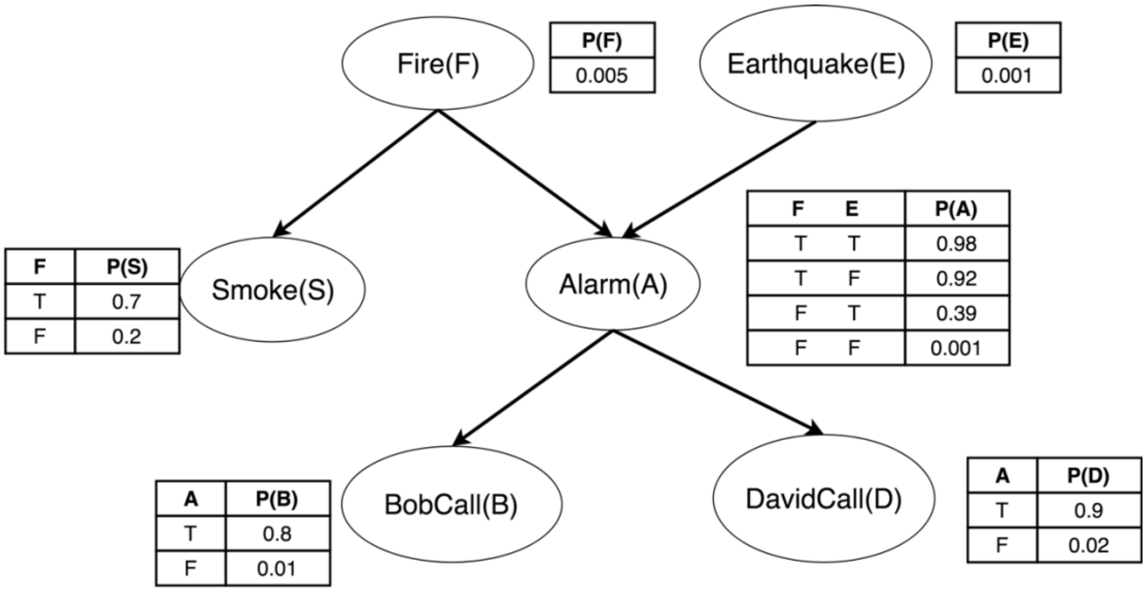
Thus, candidate X should be hired.

(b)

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.

9. Bayesian Network (10 marks)

Below is a Bayesian network and its conditional probability tables.



P(F)
0.005

P(E)
0.001

F	P(S)
T	0.7
F	0.2

F	E	P(A)
T	T	0.98
T	F	0.92
F	T	0.39
F	F	0.001

A	P(B)
T	0.8
F	0.01

A	P(D)
T	0.9
F	0.02

a) Please compute the probability of BobCall (B) given DavidCall(D): $P(B|D)$. Please round the answers to 4 decimals. [6 marks]

b) Please compute the probability of DavidCall(D) given Fire(F): $P(D|F)$. Please round the answers to 4 decimals. [4 marks]

Answer:

(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]

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

10.Ensemble Methods (12 marks)

Given the following 10 training samples, sample X are 2-dimensional data points. Label Y can be either -1 or 1, where -1 and 1 represent two classes.

No.	1	2	3	4	5	6	7	8	9	10
Sample X	(1,3)	(2,8)	(3,2)	(4,5)	(5,4)	(6,9)	(7,8)	(8,7)	(9,8)	(10,6)
Label Y	1	1	-1	-1	-1	1	1	1	-1	-1

Suppose we have two weak classifiers h1 and h2 for the following questions, which

$$h1 = \begin{cases} 1, & x1 < 2.5 \\ -1, & x1 > 2.5 \end{cases}, \quad h2 = \begin{cases} -1, & x2 < 6.5 \\ 1, & x2 > 6.5 \end{cases}$$

where X1 and X2 represent for two attributes of a data point X.

a) Give out the classification results of h1 and h2, and calculate their error rates respectively. [4 marks]

b) Please use Adaboost method to get a strong classifier by using these two weak classifiers. [8 marks]

Answer:

(a) h1 [1,1,-1,-1,-1,-1,-1,-1,-1,-1], h2 [-1,1,-1,-1,-1,1,1,1,1,-1]

error rate(h1) = 3/10 = 0.3, error rate(h2) = 2/10 = 0.2

or error rate(h1) = 0.1*3/10 = 0.03, error rate(h2) = 0.1*2/10 = 0.02

(b) **round 1**:

initial weights D1 = [0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1]

Since 0.2<0.3, we use h2 as the first base classifier, H1(x):

e1 = (0.1 + 0.1)/10 = 0.02

importance of H1(x): a1 = 0.5 * ln((1-e1)/e1) = 1.9459

update weights: D2(correct) = 0.1/[2(1-e1)] = 0.051

D2(wrong) = 0.1/[2(e1)] = 2.5

we can get the first classifier f1(x) = a1H1(x) = 1.9459H1(x)

round2:

D2 = [2.5, 0.051, 0.051, 0.051, 0.051, 0.051, 0.051, 0.051,2.5, 0.051]

error rate(h1) = 0.051 * 3/10 = 0.0153

error rate(h2) = 2.5*2/10 = 0.5

Since 0.0153<0.5, we use h1 as the first base classifier, H2(x):

e2 = 0.051 * 3/10 = 0.0153

importance of H2(x): a2 = 0.5 * ln((1-e2)/e2) = 2.0822

we can get the final classifier f2(x) = 1.9459H1(x) + 2.0822H2(x)

H(final) = sign(1.9459H1(x) + 2.0822H2(x))