Name: SUN RUI
Student ID: 18083229g

Answer 1:

a)

license: have a driver license ^license: have no driver license

wheel: want to own a balance wheel ^wheel: do not want to own a balance wheel

selfie: like selfie ^selfie: dislike selfie

	wheel	selfie	^wheel	^selfie	wheel, selfie	wheel, ^selfie	^wheel, selfie	^wheel, ^selfie
license	3250	3750	1750	1250	2500	750	1250	500
^license	2750	4000	2250	1000	2250	500	1750	500
sum	6000	7750	4000	2250	4750	1250	3000	1000

Support >= 5% means >=500 records, because all counts of itemsets are more than 500, except some unmeaning items which have 0 counts, such as "wheel, ^wheel", we can draw the below strong association according rules above table:

Rule	Confidence
wheel, selfie → license	2500/4750 ≈ 52.6%
^wheel, ^selfie → license	500/1000 = 50%
wheel, ^selfie → license	750/1250 = 60%
~~^wheel, selfie → license~~	~~1250/3000 ≈ 41.7~~

Note: the confidence of last row is less than 50%, so remove it

b)

Interest(wheel, selfie → license) = conf(wheel, selfie → license) * $\frac{1}{P(license)}$ = 0.526 * 2 = 1.052

Interest(^wheel, ^selfie → license) = conf(^wheel, ^selfie → license) * $\frac{1}{P(license)}$ = 0.5 * 2 =1

Interest(wheel, ^selfie → license) = conf(wheel, ^selfie → license) * $\frac{1}{P(license)}$ = 0.6 * 2 = 1.2

c)

In part (a), we can find the confidence of "wheel, ^selfie → license" is the biggest one, so "wheel, ^selfie → license" is most interesting rule in part (a), which can be described that these participants who want to own balance wheels and dislike selfie perhaps have driver licenses.

d)

In part (b), we can find that "wheel, ^selfie → license" have the most lift ratio, because of this, we can consider this rule is most interesting.

Actually, it is obvious that this conclusion is same to **c)**, because the left ratio is proportional to confidence value.

Answer 2:

a)

We use some simple marks instead of these stock names and operations:

	Buy	Sell	No operation
HSBC	bHS	sHS	^HS
BoEA	bBo	sBo	^Bo
China_Mobile	bCM	sCM	^CM
China_Petroleum	bCP	sCP	^CP

According to above table, we can improve Transaction Data to this:

	HSBC	BoEA	China_Mobile	China_Petroleum
# 1	bHS	sBo	bCM	^CP
# 2	bHS	sBo	^CM	^CP
# 3	^HS	bBo	bCM	bCP
# 4	bHS	^Bo	bCM	^CP
# 5	bHS	^Bo	sCM	sCP
# 6	^HS	sBo	bCM	bCP
# 7	bHS	sBo	^CM	^CP
# 8	bHS	bBo	^CM	^CP
# 9	^HS	^Bo	bCM	sCP
# 10	bHS	^Bo	bCM	bCP

Frequent itemsets:

min_sup = 20% means >=2 records

1-itemsets:

1-itemset	Count
bHS	7
^HS	3
sBo	4
bBo	2
^Bo	4
bCM	6
~~sCM~~	~~1~~
^CM	3
bCP	3
sCP	2
^CP	5

2-itemset:

2-itemset	Count	2-itemset	Count	2-itemset	Count
~~bHS, bBo~~	~~1~~	~~bBo, bCM~~	~~1~~	bCM, bCP	3
bHS, sBo	3	~~bBo, ^CM~~	~~1~~	~~bCM, sCP~~	~~1~~
bHS, ^Bo	3	sBo, bCM	2	bCM, ^CP	2
~~^HS, bBo~~	~~1~~	sBo, ^CM	2	~~sCM, sCP~~	~~1~~
~~^HS, sBo~~	~~1~~	^Bo, bCM	3	^CM, ^CP	3
~~^HS, ^Bo~~	~~1~~	~~^Bo, sCM~~	~~1~~		
bHS, bCM	3	~~bBo, bCP~~	~~1~~		
~~bHS, sCM~~	~~1~~	~~bBo, ^CP~~	~~1~~		
bHS, ^CM	3	~~sBo, bCP~~	~~1~~		
^HS, bCM	3	sBo, ^CP	3		
~~bHS, bCP~~	~~1~~	~~^Bo, bCP~~	~~1~~		
~~bHS, sCP~~	~~1~~	~~^Bo, ^CP~~	~~1~~		
bHS, ^CP	5	^Bo, sCP	2		
^HS, bCP	2				
~~^HS, sCP~~	~~1~~				

3-itemset:

3-itemset	Count	3-itemset	Count
~~bHS, sBo, bCM~~	~~1~~	bHS, ^CM, ^CP	3
bHS, sBo, ^CM	2	^HS, bCM, bCP	2
bHS, sBo, ^CP	3	~~sBo, bCM, ^CP~~	~~1~~
bHS, ^Bo, bCM	2	sBo, ^CM, ^CP	2
~~bHS, ^Bo, ^CP~~	~~1~~	~~^Bo, bCM, sCP~~	~~1~~
bHS, bCM, ^CP	2		

4-itemset:

4-itemset	Count
bHS, sBo, ^CM, ^CP	2

b) according 3-itemset of a), we can conclude: (min_conf=70%)

Rule	Confidence	Rule	Confidence	Rule	Confidence
~~bHS, sBo → ^CM~~	~~2/3 ≈ 0.67~~	~~bHS, bCM → ^Bo~~	~~2/3 ≈ 0.67~~	^CM, ^CP → bHS	3/3 = 1
~~bHS, ^CM → sBo~~	~~2/3 ≈ 0.67~~	~~^Bo, bCM → bHS~~	~~2/3 ≈ 0.67~~	~~^HS, bCM → bCP~~	~~2/3 ≈ 0.67~~
sBo, ^CM → bHS	2/2 = 1	~~bHS, bCM → ^CP~~	~~2/3 ≈ 0.67~~	^HS, bCP → bCM	2/2 = 1
bHS, sBo → ^CP	3/3 = 1	~~bHS, ^CP → bCM~~	~~2/5 = 0.4~~	~~bCM, bCP → ^HS~~	~~2/3 ≈ 0.67~~
~~bHS, ^CP → sBo~~	~~3/5 = 0.6~~	bCM, ^CP → bHS	2/2 = 1	sBo, ^CM → ^CP	2/2 = 1
sBo, ^CP → bHS	3/3 = 1	bHS, ^CM → ^CP	3/3 = 1	~~sBo, ^CP → ^CM~~	~~2/3 ≈ 0.67~~
~~bHS, ^Bo → bCM~~	~~2/3 ≈ 0.67~~	~~bHS, ^CP → ^CM~~	~~3/5 = 0.6~~	~~^CM, ^CP → sBo~~	~~2/3 ≈ 0.67~~

Strong rules for conf>70%:

sBo, ^CM → bHS   bHS, sBo → ^CP   sBo, ^CP → bHS   bCM, ^CP → bHS

bHS, ^CM → ^CP   ^CM, ^CP → bHS   ^HS, bCP → bCM   sBo, ^CM → ^CP

Answer 3:

c1:Up, c2:Down, c3:Level

I(c1, c2, c3) = I(3, 2, 2) = $\frac{3}{7} * \log_2 \frac{3}{7} + \frac{2}{7} * \log_2 \frac{2}{7} + \frac{2}{7} * \log_2 \frac{2}{7} \approx$ 1.557

Entropy for 2TDB:

2TDB	c1	c2	c3	I(c1, c2, c3)
Up	1	1	2	1.5
Down	1	0	0	0
Level	1	1	0	1

E(2TDB) = $\frac{4}{7} * 1.5 + \frac{1}{7} * 0 + \frac{2}{7} * 1 \approx 1.14$

Information_Gain (2TDB) = 1.557 – 1.14 = 0.417

Entropy for 1TDB:

1TDB	c1	c2	c3	I(c1, c2, c3)
Up	0	2	1	0.918
Down	2	0	0	0
Level	1	0	1	1

E(1TDB) = $\frac{3}{7} * 0.918 + \frac{2}{7} * 0 + \frac{2}{7} * 1 = 0.6791$

Information_Gain (1TDB) = 1.557 – 0.6791 = 0.878

Entropy for TD:

2TDB	c1	c2	c3	I(c1, c2, c3)
Up	1	1	0	1
Down	0	0	2	0
Level	2	1	0	0.918

E(TD) = $\frac{2}{7} * 1 + \frac{2}{7} * 0 + \frac{3}{7} * 0.918 = 0.6791$

Information_Gain (TD) = 1.557 – 0.6791 = 0.878

According above tables, we can find that the information gains for TD and 1TDB are equal , and bigger than the 2TDB, so we choose TD as root node. **And we can find if TD is down, the all results of NTD are level, so we decide this one branch of root node.** Then, we need to decide child nodes and branches.

If TD is level:

I(c1, c2, c3) = I(2, 1, 0) = $\frac{2}{3} * \log_2 \frac{1}{3} + \frac{2}{3} * \log_2 \frac{2}{3} + 0 \approx$ 0.918

Entropy for 1TDB:

1TDB	c1	c2	c3	I(c1, c2, c3)
Up	0	1	0	0
Down	2	0	0	0
Level	0	0	0	0

$E(1TDB) = \frac{1}{3} * 0 + \frac{2}{3} * 0 + \frac{0}{3} * 0 = 0$

Information_Gain (1TDB) = 0.918 – 0 = 0.918

Entropy for 2TDB:

2TDB	c1	c2	c3	I(c1, c2, c3)
Up	1	1	0	1
Down	0	0	0	0
Level	1	0	0	0

$E(2TDB) = \frac{2}{3} * 1 + 0 + 0 \approx 0.667$

Information_Gain (2TDB) = 0.918 – 0.667 = 0.251

Obviously, the information gain of 1TDB is bigger, so we can choose 1TDB as the child node of root node when TD is level. But we also need to check the other situation if TD is up.

Now, the only branch is when TD is up, we can calculate entropy in this situation:

Entropy for 1TDB:

1TDB	c1	c2	c3	I(c1, c2, c3)
Up	0	1	0	0
Down	0	0	0	0
Level	1	0	0	0

E(1TDB) = 0 + 0 + 0 =0

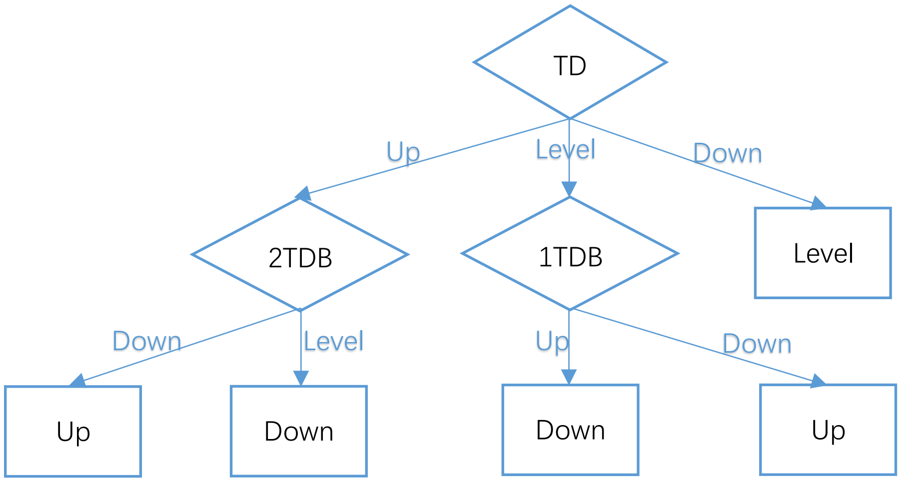
Entropy for 2TDB:

2TDB	c1	c2	c3	I(c1, c2, c3)
Up	0	0	0	0
Down	1	0	0	0
Level	0	1	0	0

E(2TDB) = 0

Because entropy of both of these two attributes is zero, any one of them can be as the child node when TD is up. Then we choose 2TDB as the child node when TD is up.

According to above analysis, we can get this decision tree:



12 Sep: IF TD = Level AND 1TDB = Up THEN NTD = Down
13 Sep: IF TD = Down THEN NTD = Level