ASSIGNMENT 1 SUBHAM ANAND Z5151878

Question 1

1

Location	Time	Item	Sum
ALL	2005	PS2	1400
ALL	2006	PS2	1500
ALL	2006	Wii	500
ALL	2005	XBOX 360	1700
ALL	2005	ALL	3100
ALL	2006	ALL	2000
ALL	ALL	PS2	2900
ALL	ALL	Wii	500
ALL	ALL	XBOX 360	1700
ALL	ALL	ALL	5100
Sydney	2005	PS2	1400
Sydney	2006	PS2	1500
Sydney	2006	Wii	500
Sydney	2005	ALL	1400
Sydney	2006	ALL	2000
Sydney	ALL	PS2	2900
Sydney	ALL	Wii	500
Sydney	ALL	ALL	3400
Melbourne	2005	XBOX 360	1700
Melbourne	2005	ALL	1700
Melbourne	ALL	XBOX 360	1700
Melbourne	ALL	ALL	1700

2

Select location, time, item, sum(m) from R group by location, time, item UNION ALL

Select location, time, ALL, sum(m) from R group by location, time UNION ALL

Select location, ALL, items, sum(m) from R group by location, item UNION ALL

Select ALL, time, ALL, sum(m) from R group by time UNION ALL

Select ALL, time, item, sum(m) from R group by time, item UNION ALL

Select location, ALL, ALL, sum(m) from R group by location UNION ALL

Select ALL, ALL, item, sum(m) from R group by item UNION ALL

Select ALL, ALL, ALL, sum(m) from R

3

LOCATION	TIME	ITEM	QUANTITY
ALL	ALL	PS2	2900
ALL	ALL	ALL	5100
Sydney	ALL	PS2	2900
Sydney	ALL	ALL	3400
ALL	2005	ALL	3100
Sydney	2006	ALL	2000
ALL	2006	ALL	2000

4

Location	Time	Item	Sum	Func(l,t,i) Index
1	1	1	1400	17
1	2	1	1500	21
1	2	3	500	23
1	1	0	1400	16
1	2	0	2000	20
1	0	1	2900	13
1	0	3	500	15
1	0	0	3400	12
2	1	2	1700	30
2	1	0	1700	28
2	0	2	1700	26
0	1	1	1400	5
2	0	0	1700	24
0	2	1	1500	9
0	2	3	500	11
0	1	2	1700	6
0	1	0	3100	4
0	2	0	2000	8
0	0	1	2900	1
0	0	3	500	3
0	0	2	1700	2
0	0	0	5100	0

12 location + 4 time + 1 item

Question 2

1

The general Naïve Bayes classifier is $y' = \text{argmax } p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)$. It can be written as :

logClass	=	$\log (Prob[C_+ \mid u] / Prob[C_- \mid u])$
	=	$\log Prob[C_+ \mid u] . Prob[C_+] - \log Prob[C_- \mid u] . Prob[C_-] . Prob[C_-]$
	=	$\log \prod Prob[x_i \mid C_+] + \log Prob[C_+] - \log \prod Prob[x_i \mid C_-] + \log Prob[C_-]$
	=	$\sum \{ \log (Prob[x_i \mid C_+] / Prob[x_i \mid C_-]) + \log (Prob[C_+] / Prob[C_-]) \}$
	=	$\log Prob[C_+] / Prob[C_-]$

logclass = $b + \sum \alpha(i, u_i)$

= $b + \sum (\alpha (i, 1) . u_i + \alpha (i, 0) . (1- u_i))$

It will be a linear classifier for the feature vector [1,x] and so on.

2

Both models use MLE to find parameters. NB model uses the conditional independence assumption and there is no relation between class labels. Logistic regression has nothing like that and all MLE parameters are together. Logistic regression is mainly used in cases where the output is Boolean.

Question 3

1

We take three samples with different probability of quantity of q1 and q2.

First sample, taking both q1 and q2 of same quantity as 0.5.

Calculating u1, u2 and u3

u1 = 25 / 100 = 0.25

u2 = 35 / 100 = 0.35

u3 = 40 / 100 = 0.4

P ({u1, u2, u3} | {q1, q2})

$\log P(u1|q1) = \log \binom{3}{0.25*3}^{0.5*3} . (1-0.5)^{(1-0.25)3}$

log P(u1|q2) = it will be like the above one because the sample is evenly distributed.

$\log P(u2|q1) = \log \binom{3}{0.35*3}^{0.5*3} . (1-0.5)^{(1-0.35)3}$

log P(u2|q2) = it will be like the above one because the sample is evenly distributed.

$\text{Log } P(u3|q1) = \log \binom{3}{0.4*3}^{0.5*3} . (1-0.5)^{(1-0.4)3}$

Similarly, we can find the MLE for other samples which have different probability distribution of q1 and q2.

2

The MLE for q1 and q2 for the given values of u1, u2, u3 are :

log P(u1|q1) = log $\binom{3}{0.3*3}^{0.5*3}$. $(1\text{-}0.5)^{(1\text{-}0.3)3}$

log P(u1|q2) = log $\binom{3}{0.25*3}^{0.5*3}$. $(1\text{-}0.5)^{(1\text{-}0.25)3}$

log P(u2|q1) = log $\binom{3}{0.2*3}^{0.5*3}$. $(1\text{-}0.5)^{(1\text{-}0.2)3}$

log P(u2|q2) = log $\binom{3}{0.2*3}^{0.5*3}$. $(1\text{-}0.5)^{(1\text{-}0.2)3}$

log P(u3|q1) = log $\binom{3}{0.5*3}^{0.5*3}$. $(1\text{-}0.5)^{(1\text{-}0.5)3}$

The percentage of each component are:

O1 => o1/1 = .3

= 0.3

O2 => o2/1 = .2

= 0.2

O3 => 03/1 = .5

= 0.5