## ZID: z5143964
## Name: PEIGUO GUAN
## Q1:

(1) Answer:

| 0 | *Location* | *Time* | *Item* | *SUM(Quantity)* |
|---|---|---|---|---|
| 1 | Sydney | 2005 | PS2 | 1400 |
| 2 | Sydney | 2006 | PS2 | 1500 |
| 3 | Sydney | 2006 | Wii | 500 |
| 4 | Melbourne | 2005 | Xbox 360 | 1700 |
| 5 | Sydney | 2005 | ALL | 1400 |
| 6 | Sydney | 2006 | ALL | 2000 |
| 7 | Sydney | ALL | PS2 | 2900 |
| 8 | Sydney | ALL | Wii | 500 |
| 9 | Sydney | ALL | ALL | 3400 |
| 10 | Melbourne | 2005 | ALL | 1700 |
| 11 | Melbourne | ALL | Xbox 360 | 1700 |
| 12 | Melbourne | ALL | ALL | 1700 |
| 13 | ALL | 2005 | PS2 | 1400 |
| 14 | ALL | 2005 | Xbox 360 | 1700 |
| 15 | ALL | 2005 | ALL | 3100 |
| 16 | ALL | 2006 | PS2 | 1500 |
| 17 | ALL | 2006 | Wii | 500 |
| 18 | ALL | 2006 | ALL | 2000 |
| 19 | ALL | ALL | PS2 | 2900 |
| 20 | ALL | ALL | Xbox 360 | 1700 |
| 21 | ALL | ALL | Wii | 500 |
| 22 | ALL | ALL | ALL | 5100 |

(2) Answer:

```
SELECT Location, Time, Item, SUM(Quantity)
FROM SALES
GROUP BY Location, Time, Item
UNION
SELECT Location, Time, “ALL”, SUM(Quantity)
FROM SALES
GROUP BY Location, Time
UNION
SELECT Location, “ALL”, Item, SUM(Quantity)
FROM SALES
GROUP BY Location, Item
UNION
SELECT “ALL”, Time, Item, SUM(Quantity)
FROM SALES
GROUP BY Time, Item
UNION
```

```
SELECT “ALL”, “ALL”, Item, SUM(Quantity)
FROM SALES
GROUP BY Item
UNION
SELECT “ALL”, Time, “ALL”, SUM(Quantity)
FROM SALES
GROUP BY Time
UNION
SELECT Location, “ALL”, “ALL”, SUM(Quantity)
FROM SALES
GROUP BY Location
UNION
SELECT “ALL”, “ALL”, “ALL”, SUM(Quantity)
FROM SALES
```

(3) Answer:

| ***Location*** | ***Time*** | ***Item*** | ***SUM(Quantity)*** |
|---|---|---|---|
| Sydney | 2006 | ALL | 2000 |
| Sydney | ALL | PS2 | 2900 |
| Sydney | ALL | ALL | 3400 |
| ALL | 2005 | ALL | 3100 |
| ALL | 2006 | ALL | 2000 |
| ALL | ALL | PS2 | 2900 |
| ALL | ALL | ALL | 5100 |

(4) Answer:

Step1:

| ***Location*** | ***Time*** | ***Item*** | ***SUM(Quantity)*** |
|---|---|---|---|
| 1 | 1 | 1 | 1400 |
| 1 | 2 | 1 | 1500 |
| 1 | 2 | 3 | 500 |
| 2 | 1 | 2 | 1700 |
| 1 | 1 | 0 | 1400 |
| 1 | 2 | 0 | 2000 |
| 1 | 0 | 1 | 2900 |
| 1 | 0 | 3 | 500 |
| 1 | 0 | 0 | 3400 |
| 2 | 1 | 0 | 1700 |
| 2 | 0 | 2 | 1700 |
| 2 | 0 | 0 | 1700 |
| 0 | 1 | 1 | 1400 |
| 0 | 1 | 2 | 1700 |
| 0 | 1 | 0 | 3100 |

| | | | |
|---|---|---|---|
| 0 | 2 | 1 | 1500 |
| 0 | 2 | 3 | 500 |
| 0 | 2 | 0 | 2000 |
| 0 | 0 | 1 | 2900 |
| 0 | 0 | 2 | 1700 |
| 0 | 0 | 3 | 500 |
| 0 | 0 | 0 | 5100 |

Step2:

| ***Offset*** | ***Quantity*** |
|---|---|
| 17 | 1400 |
| 21 | 1500 |
| 23 | 500 |
| 30 | 1700 |
| 16 | 1400 |
| 20 | 2000 |
| 13 | 2900 |
| 15 | 500 |
| 12 | 3400 |
| 28 | 1700 |
| 26 | 1700 |
| 24 | 1700 |
| 5 | 1400 |
| 6 | 1700 |
| 4 | 3100 |
| 9 | 1500 |
| 11 | 500 |
| 8 | 2000 |
| 1 | 2900 |
| 2 | 1700 |
| 3 | 500 |
| 0 | 5100 |

Reorder:

| ***Offset*** | ***Quantity*** |
|---|---|
| 0 | 5100 |
| 1 | 2900 |
| 2 | 1700 |
| 3 | 500 |
| 4 | 3100 |
| 5 | 1400 |
| 6 | 1700 |
| 8 | 2000 |
| 9 | 1500 |
| 11 | 500 |
| 12 | 3400 |
| 13 | 2900 |

| | |
|---|---|
| 15 | 500 |
| 16 | 1400 |
| 17 | 1400 |
| 20 | 2000 |
| 21 | 1500 |
| 23 | 500 |
| 24 | 1700 |
| 26 | 1700 |
| 28 | 1700 |
| 30 | 1700 |

MD ARRAY:

| ***MD ARRAY*** |
|---|
| 5100 |
| 2900 |
| 1700 |
| 500 |
| 3100 |
| 1400 |
| 1700 |
| 2000 |
| 1500 |
| 500 |
| 3400 |
| 2900 |
| 500 |
| 1400 |
| 1400 |
| 2000 |
| 1500 |
| 500 |
| 1700 |
| 1700 |
| 1700 |
| 1700 |

**Function:**

$\mathrm{f(location, time, item)} = (3 * \mathrm{location} + \mathrm{time}) * 4 + \mathrm{item}$

## Q2:

(1) According to the naïve bayes classifier:

$$P(y|x) = \ argmax_{x\epsilon\{C_j\}} \prod_{i=1}^{n} P(k_i|C_0) \cdot P(C_0)$$

Since the feature of the value is binary which is 0 and 1, if the value is 0:

$$f(x=0) = \ argmax_{x\epsilon\{C_j\}} \prod_{i=1}^{n} P(k_i|C_0)\cdot P(C_0)$$

and if the value is 1:

$$f(x=1) = \ argmax_{x\epsilon\{C_j\}} \prod_{i=1}^{n} P(k_i|C_1)\cdot P(C_1)$$

if f(x=0)-f(x=1)>0, then f(x) is classified to 1, otherwise classified to 0

we can change it to d dimension:

$$f(y=1) = \ P(y=1)\prod_{i=1}^{n} P(x_i=1|y=1)^{x_i}\cdot P(x_i=0|y=1)^{1-x_i}$$

$$f(y=0) = \ P(y=0)\prod_{i=1}^{n} P(x_i=1|y=0)^{x_i}\cdot P(x_i=0|y=0)^{1-x_i}$$

use log and then:

$$f(y) = \log\big(P(y)\big) + \sum_{i=0}^{d} P(x_i=1|y)^{x_i}\cdot P(x_i=0|y)^{1-x_i}$$

since $k_i \ \epsilon \ \{0,1\}, k_1 = 1 - \ k_0$, so can change the function to be:

$$f(y) = \ \log\frac{P(y=0)}{P(y=1)} + \sum_{i=0}^{d} \log\frac{P(x_i=0|y=0)}{P(x_i=0|y=1)} + x_i \sum_{i=0}^{d} \log\frac{P(x_i=0|y=0)P(x_i=1|y=0)}{P(x_i=0|y=1)P(x_i=1|y=1)}$$

we can let:

$$x_0 = \ \log\frac{P(y=0)}{P(y=1)} + \sum_{i=0}^{d} \log\frac{P(x_i=0|y=0)}{P(x_i=0|y=1)}$$

$$w_i = \ \log\frac{P(x_i=0|y=0)P(x_i=1|y=0)}{P(x_i=0|y=1)P(x_i=1|y=1)}$$

so $f(y) = \ x_0 + \sum_{i=0}^{d} w_i x_i$

So it satisfied the linear classifier, and the x range from [1, $x_d$] which is d+1 dimension of linear classifier.

(2) $w_{NB}$ is easier to get than $w_{LG}$ that is because NB model can learned data by conditional independence data from training data set, like P(y), P(x|y)

and so on, it is not that difficult to be calculated, while LR classifier do not have such preprocess data calculation, it needs to search whole linear space of possible models. So NB models hold the necessary parameters on the same time which help it to easy calculate the solution.

## Q3:

(1) Answer:

According to the questions, we can conclude that it is a binomial distribution, so the log likely function should be like this:

According to the log likely function:

$$\mathrm{l}(\mathrm{y}|\theta) = \log\ \mathrm{P}(\mathrm{y}|\theta)$$

$$\log\mathrm{P}(p_o|p_m) = \log\left(\binom{n}{p_0 n}\mathrm{p}_m^{p_o n}\cdot(1-p_m)^{(1-p_o)n}\right)$$

so

$$\log\mathrm{P}\left(u_j\middle|\mathrm{q}_1, q_2\right) = \log\left(p_{1,j}q_1 + p_{2,j}q_2\right)^{u_j}$$

So total log likely function is:

$$\log\mathrm{L} = \log\mathrm{P}\left(u_j\middle|\mathrm{q}_i\right) = \log\prod_{u=1}^{3}\sum_{i=1,j=1}^{2,3}\left(p_{i,j}q_i\right)^{u_j}$$

$$\log\mathrm{L} = \log\mathrm{P}\left(u_j\middle|\mathrm{q}_i\right) = \log\prod_{u=1}^{3}\sum_{i=1,j=1}^{2,3}\left(p_{i,j}q_i\right)^{u_j}$$

$$\log\mathrm{L} = \log\mathrm{P}\left(u_j\middle|\mathrm{q}_1, q_2\right) = \log\left(\left(0.1q_1 + 0.4q_2\right)^{u_1}\times\left(0.2q_1 + 0.5q_2\right)^{u_2}\times\left(0.7q_1 + 0.1q_2\right)^{u_3}\right)$$

$$= u_1\log\left(0.1q_1 + 0.4q_2\right) + u_2\log\left(0.2q_1 + 0.5q_2\right) + u_3\log\left(0.7q_1 + 0.1q_2\right)$$

If $q_1$ is only parameter ($q_2 = 1 - q_1$):

$$(logL)' = \frac{0.1u_1}{0.1q_1 + 0.4q_2} + \frac{0.2u_2}{0.2q_1 + 0.5q_2} + \frac{0.7u_3}{0.7q_1 + 0.1q_2}$$

$$(logL)' = \frac{-0.3u_1}{0.4 - 0.3q_1} + \frac{-0.3u_2}{0.5 - 0.3q_1} + \frac{0.6u_3}{0.1 + 0.6q_1}$$

(2)

Let $(logL)' = 0, u_1 = 0.3, u_2 = 0.2, u_3 = 0.5$, and $q_1$ is only parameter:

After calculation, $(logL)' = 540q^2 - 1179q + 531 = 0$

So $q \approx 0.6351$

so $q_1 \approx 0.635, q_2 \approx 0.365$

The expected of each component is:

u1=0.1 * q1 + 0.4 * q2 = 0.209

u2=0.2 * q1 + 0.5 * q2 = 0.30

u3=0.7 * q1 + 0.1 * q2 = 0.480