Assignment1

COMP9318

Cong Cong

Z3414050

Q1.

1):

Location	Time	Item	Quantity
Sydney	2005	PS2	1400
Sydney	2006	PS2	1500
Sydney	2006	Wii	500
Melbourne	2005	XBox 360	1700
Sydney	2005	ALL	1400
Sydney	2006	ALL	2000
Melbourne	2005	ALL	1700
Sydney	ALL	PS2	2900
Sydney	ALL	Wii	500
Melbourne	ALL	XBox 360	1700
ALL	2005	PS2	1400
ALL	2006	PS2	1500
ALL	2006	Wii	500
ALL	2005	XBox 360	1700
Sydney	ALL	ALL	3400
Melbourne	ALL	ALL	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

2):

```
SELECT Location, Time, Item, SUM(Quantity)
FROM Sales
GROUP BY Location, Time, Item
UNION ALL
SELECT Location, Time, ALL, SUM(Quantity)
FROM Sales
GROUP BY Location, Time
UNION ALL
SELECT Location, ALL, Item, SUM(Quantity)
FROM Sales
GROUP BY Location, Item
UNION ALL
SELECT ALL, Time, Item, SUM(Quantity)
FROM Sales
GROUP BY Time, Item
UNION ALL
SELECT Location, ALL, ALL, SUM(Quantity)
FROM Sales
GROUP BY Location
UNION ALL
SELECT ALL, Time, ALL, SUM(Quantity)
FROM Sales
GROUP BY Time
UNION ALL
SELECT ALL, ALL, Item, SUM(Quantity)
FROM Sales
GROUP BY Item
UNION ALL
SELECT ALL, ALL, ALL, SUM(Quantity)
FROM Sales
```

3):

Location	Time	Item	Quantity
Sydney	ALL	ALL	3400
Sydney	2006	ALL	2000
Sydney	ALL	PS2	2900
ALL	2005	ALL	3100
ALL	2006	ALL	2000
ALL	ALL	PS2	2900
ALL	ALL	ALL	5100

4):

In order to find an injective mapping function:

$$offset = f(l,t,i) = a*l + t*t + c*i$$

I tried $a = 10, b = 4\ and\ c = 1$ but it doesn't provide an one to one mapping, then I take $a = 12, b = 4\ and\ c = 1$ and this gives me an one to one mapping.

$$offset = f(l,t,i) = 12*l + 4*t + i$$

Location	Time	Item	offset
1	1	1	17
1	2	1	21
1	2	3	23
2	1	2	30
1	1	0	16
1	2	0	20
2	1	0	28
1	0	1	13
1	0	3	15
2	0	2	26
0	1	1	5
0	2	1	9
0	2	3	11
0	1	2	6
1	0	0	12
2	0	0	24
0	1	0	4
0	2	0	8
0	0	1	1
0	0	3	3
0	0	2	2
0	0	0	0

And the sparse multi-dimensional array is (sorted by offset):

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

Q2

1):

I used $\log odds = \log(\frac{p(y=1|x)}{p(y=0|x)})$, so if $\log odds$ > 1, I can classify input features as $P(y=1)$ class, otherwise, the input features can be classified as $P(y=0)$ class.

Next:

$$\log odds = \log\left(\frac{p(y=1|x)}{p(y=0|x)}\right)$$

And based on Bayesian Theorem:

$$\log odds = \log\left(\frac{p(y=1) * \prod_1^d P(x_i|y=1)}{p(y=0) * \prod_1^d P(x_i|y=0)}\right)$$

$$\log odds = \log\left(P(y=1) * \prod_1^d P(x_i|y=1)\right) - \log\left(P(y=0) * \prod_1^d P(x_i|y=0)\right)$$

$$\log odds = \log\big(P(y=1)\big) + \sum_i^d \log(P(x_i|y=1))$$

$$-\log(P(y=0)) - \sum_i^d \log(P(x_i|y=0)$$

The above equation equals to:

$$\log odds = \sum_i^d \log(\frac{P(x_i|y=1)}{P(x_i|y=0)}) + \log(\frac{P(y=1)}{P(y=0)}) \ [1]$$

Now, we can use **Bernoulli Naïve Bayes** to further simplify $\sum_i^d \log(\frac{P(x_i|y=1)}{P(x_i|y=0)})$:

$$\log\left(\frac{P(x_i|y=1)}{P(x_i|y=0)}\right) = \log\left(\frac{\sum_1^d P_1^{x_i}(1-P_1)^{1-x_i}}{\sum_1^d P_0^{x_i}(1-P_0)^{1-x_i}}\right)$$

$$= \log\left(\sum_1^d P_1^{x_i}(1-P_1)^{1-x_i}\right) - \log\left(\sum_1^d P_0^{x_i}(1-P_0)^{1-x_i}\right)$$

$$= x_i\sum_1^d \log(P_1) + (1-x_i)\sum_1^d \log(1-P_1)$$

$$-x_i\sum_1^d \log(P_0) - (1-x_i)\sum_1^d \log(1-P_0)$$

And the above equation can be simplified to:

$$= x_i \sum_{1}^{d} \log(\frac{P_1(1-P_0)}{P_0(1-P_1)}) + \sum_{1}^{d} \log\left(\frac{1-P_1}{1-P_0}\right) \ [2]$$

I assume:

$$\alpha = \sum_{1}^{d} \log\left(\frac{P_1(1-P_0)}{P_0(1-P_1)}\right), \beta = \sum_{1}^{d} \log\left(\frac{1-P_1}{1-P_0}\right)$$

And equation is:

$$x_i * \alpha + \beta$$

Now, we can substitute equation [2] back to equation [1], and we obtain:

$$\log odds = x_i \sum_{1}^{d} \alpha + \beta + \log(\frac{P(y=1)}{P(y=0)})$$

If I assume:

$$\gamma = \beta + \log(\frac{P(y=1)}{P(y=0)})$$

Finally, the equation can be simplified to:

$$\log odds = x_i \sum_{1}^{d} \alpha + \gamma$$

This is obviously a linear classifier in d+1 dimension space, with the vector

$$\omega = [\,\gamma, \alpha_1, \alpha_2 \ldots, \alpha_d]$$

2):

From part 1), I know that

$$\omega_{NB} = [\,\gamma, \alpha_1, \alpha_2 \ldots, \alpha_d]$$

Every element in ω_{NB} can be obtained directly

However, if I use Logistic Regression:

$$P(y=1|x_i) = \frac{1}{1+e^{-\omega^t x}}$$

Here, I need to find ω^t which maximize the likelihood:

$$l(w) = \prod_{i=1}^{d} P(y_i=1|x_i)^{y_i} \left(1-P(y_i=1|x_i)\right)^{1-y_i}$$

And Log-likelihood is:

$$\log\left(l(w)\right) = \sum_{i=1}^{d} y_i \log(\ P(y_i=1|x_i)) + (1-y_i) log\left(1-P(y_i=1|x_i)\right)$$

So, if we take the derivative of $\log\left(l(w)\right)$:

$$\frac{\log\left(l(w)\right)}{dw} = \sum_{i=1}^{d} (x_i y_i - x_i P(y_i=1|x_i))$$

As, $P(y_i = 1|x_i)$ is a function of w, our aim is to try to make our estimation $x_i P(y_i = 1|x_i)$ as close to the observed data $x_i y_i$ as possible.
To achieve that goal, there are several ways to do, one of them is to use **Gradient Ascent,** but obviously this method is more complicated than calculating ω_{NB} directly.

Q3:
1):
q_1, q_2 is the percentages sample S_1, S_2 in the mixture, and $p_{i,j}$ is the percentage of Object O_j in the sample q_i, now after the measurements, we are given the percentage of Object O_i in the whole mixture and noted as u_j.

The likelihood function can be written as:

$$P(u_i|p_{i,j}, q_i) = (p_{1,1}q_1 + p_{2,1}q_2)^{u_1}(p_{1,2}q_1 + p_{2,2}q_2)^{u_2}\ (p_{1,3}q_1 + p_{2,3}q_2)^{u_3}$$

$$= \prod_{j=1}^{3}(\sum_{i=1}^{2}(p_{i,j}\ \ q_i)^{u_i})$$

Take the log of the above equation:

$$\log\Big(\ P(u_i|p_{i,j}, q_i)\Big) = \log(\prod_{j=1}^{3}(\sum_{i=1}^{2}(p_{i,j}\ \ q_i)^{u_i})\ )$$

$$= \sum_{j=1}^{3} u_i \log(\sum_{i=1}^{2} p_{i,j}\ \ q_i)$$

And the above equation is the log likelihood function.
2):
To make simplification easier, it is better to use ln instead of using log, now the values of u_i are given and the values of $p_{i,j}$ are provided in the table, thus the only unknown variable in the above equation is q_i, but I know $q_2 = 1 - q_1$ and I can use q_1 to represent q_2. First substituting the known values into the equation:

$$\log\Big(\ P(u_i|p_{i,j}, q_i)\Big) = 0.3\ln(0.4 - 0.3q_1) + 0.2\ln(0.5 - 0.3q_1)$$
$$+0.5\ln(0.1 + 0.6q_1)$$

To find the MLE, I first take the derivative of the above equation and set it equals to zero, what I get is:

$$\frac{-0.09}{0.4 - 0.3q_1} + \frac{-0.06}{0.5 - 0.3q_1} + \frac{0.3}{0.1 + 0.6q_1} = 0$$

Solve the above equation, I got $q_1 = 0.635$ and $q_2 = 0.365$

And the expected percentage of each component is:

$$O_1 = p_{1,1}q_1 + p_{2,1}q_2 = 0.1 * 0.635 + 0.4 * 0.365 = 0.2095$$
$$O_2 = p_{1,2}q_1 + p_{2,2}q_2 = 0.2 * 0.635 + 0.5 * 0.365 = 0.3095$$
$$O_3 = p_{1,3}q_1 + p_{2,3}q_2 = 0.7 * 0.635 + 0.1 * 0.365 = 0.481$$