COMP9318 Assignment

Z5142340 Haiyu LYU

Q1:

(1).

	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	Melbourne	2005	ALL	1700
8	Sydney	ALL	PS2	1900
9	Sydney	ALL	Wii	500
10	Melbourne	ALL	XBox 360	1700
11	ALL	2005	PS2	1400
12	ALL	2005	XBox 360	1700
13	ALL	2006	PS2	1500
14	ALL	2006	Wii	500
15	Sydney	ALL	ALL	3400
16	Melbourne	ALL	ALL	1700
17	ALL	2005	ALL	3100
18	ALL	2006	ALL	2000
19	ALL	ALL	PS2	2900
20	ALL	ALL	Wii	500
21	ALL	ALL	XBox 360	1700
22	ALL	ALL	ALL	5100

(2).

```
SELECT Location, Time, Item, SUM(Quantity)
FROM R
GROUP BY Location, Time, Item
```

UNION ALL

```
SELECT Location, Time, ALL, SUM(Quantity)
FROM R
GROUP BY Location, Time
```

UNION ALL

```
SELECT Location, ALL, Item, SUM(Quantity)
FROM R
GROUP BY Location, Item
```

UNION ALL

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

(3).

	Location	Time	Item	SUM(Quantity)
1	Sydney	ALL	PS2	2900
2	Sydney	2006	ALL	2000
3	Sydney	ALL	ALL	3400
4	ALL	ALL	PS2	2900
5	ALL	2005	ALL	3100
6	ALL	2006	ALL	2000
7	ALL	ALL	ALL	5100

(4).

$$f(Location, Time, Item) = (Location * 3 + Time) * 4 + Item$$
$$= 12 * Location + 4 * Time + Item$$

Location	Time	Item	SUM(Quantity)	Offset
1	1	1	1400	17
1	2	1	1500	21
1	2	3	500	23
2	1	2	1700	30
1	1	0	1400	16
1	2	0	2000	20
2	1	0	1700	28
1	0	1	1900	13
1	0	3	500	15
2	0	2	1700	26
0	1	1	1400	5

0	1	2	1700	6
0	2	1	1500	9
0	2	3	500	11
1	0	0	3400	12
2	0	0	1700	24
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

MOLAP cube in tabular form:

ArrayIndex	Value
17	1400
21	1500
23	500
30	1700
16	1400
20	2000
28	1700
13	1900
15	500
26	1700
5	1400
6	1700
9	1500
11	500
12	3400
24	1700
4	3100
8	2000
1	2900
3	500
2	1700
0	5100

Q2:

(1).

We have d-dimension column vector $\vec{x} = [x_1, x_2 \cdots x_d]$, and each x_i takes only two values (0 or 1).

We use odds and Bayes' Rule:

$$\mathrm{O}(\mathrm{Y}|\vec{x}) = \frac{p(y=1|\vec{x})}{p(y=0|\vec{x})} = \frac{p(y=1)}{p(y=0)} \cdot \frac{p(\vec{x}|y=1)}{p(\vec{x}|y=0)}$$

Using Independence Assumption:

$$\mathrm{O}(\mathrm{Y}|\vec{x}) = \frac{p(y=1|\vec{x})}{p(y=0|\vec{x})} = \frac{p(y=1)}{p(y=0)} \cdot \prod_{i=1}^{d} \frac{p(\vec{x_i}|y=1)}{p(\vec{x_i}|y=0)}$$

Since $\vec{x_i}$ takes either 0 or 1:

We assume that:

$$p = p(y=1)$$
$$q_i = p(x_i = 1|y=1)$$
$$r_i = p(x_i = 1|y=0)$$

Then

$$\mathrm{O}(\mathrm{Y}|\vec{x}) = \frac{p(y=1)}{p(y=0)} \cdot \prod_{i=1}^{d} \frac{p(\vec{x_i}|y=1)}{p(\vec{x_i}|y=0)}$$

$$= \frac{p}{1-p} \cdot \prod_{i=1}^{d} \frac{p(\vec{x_i}=1|y=1)}{p(\vec{x_i}=1|y=0)} \cdot \prod_{i=1}^{d} \frac{p(\vec{x_i}=0|y=1)}{p(\vec{x_i}=0|y=0)}$$

$$= \frac{p}{1-p} \cdot \prod_{i=1}^{d} \frac{q_i}{r_i} \cdot \prod_{i=1}^{d} \frac{1-q_i}{1-r_i}$$

$$= \frac{p}{1-p} \cdot \prod_{i=1}^{d} \frac{1-q_i}{1-r_i} \cdot \prod_{i=1}^{d} \frac{q_i(1-r_i)}{r_i(1-q_{i)}} \cdot \vec{x_i}$$

$$\log \mathrm{O}(\mathrm{Y}|\vec{x}) = \log\left(\frac{p}{1-p} \cdot \prod_{i=1}^{d} \frac{1-q_i}{1-r_i}\right) + \log\left(\prod_{i=1}^{d} \frac{q_i(1-r_i)}{r_i(1-q_{i)}} \cdot \vec{x_i}\right)$$

$$= \boxed{\log\frac{p}{1-p} + \sum_{i=1}^{d} \log\frac{1-q_i}{1-r_i}} + \boxed{\sum_{i=1}^{d} \log\frac{q_i(1-r_i)}{r_i(1-q_{i)}} \cdot \vec{x_i}}$$

We can see that the first part is a constant. So $w'_0 = \log\frac{p}{1-p} + \sum_{i=1}^{d} \log\frac{1-q_i}{1-r_i}$

For each $w_i (0 < i \le d)$, we have that $w'_i = \log\frac{q_i(1-r_i)}{r_i(1-q_{i)}}$

Therefore, the vector **w** that the Naïve Bayes classifier learns:

$$\boldsymbol{w}^T = [w'_0, w'_1, w'_2 \cdots w'_d]$$

(2).

The difference between the two is that the way they seek weight is different. Learning $\mathbf{w}_{NB}$ use independent assumption. Because the conditions are independent, the Bayes' approach does not require gradient descent. $\mathbf{w}_{NB}$ will be identified by calculating odds ratio of each feature. However, learning $\mathbf{w}_{LR}$ needs to calculate the coupling information between each feature by gradient descent to get the weight. Therefore, learning $\mathbf{w}_{NB}$ is much easier than learning $\mathbf{w}_{LR}$.

Q3:

(1).

$u_1 = q_1p_{11} + q_2p_{21}$

$u_2 = q_1p_{12} + q_2p_{22}$

$u_3 = q_1p_{13} + q_2p_{23}$

Since the sample is composed of three components, each of which is u_j.

$u_1 + u_2 + u_3 = 1$

$q_2 = 1 - q_1$

$$\mathrm{P}(\mathrm{U}|q_1) = \mathrm{P}(u_1, u_2, u_3|q_1) = \prod_{j=1}^{3} P(u_j|q_1)$$

$$= (q_1p_{11} + (1-q_1)p_{21})^{u_1}(q_1p_{12} + (1-q_1)p_{22})^{u_2}(q_1p_{13} + (1-q_1)p_{23})^{u_3}$$

Log likelihood function:

$$\ell(u_1, u_2, u_3|\theta) = \sum_{j=1}^{3} \log P(u_j|\theta)$$

$$= u_1(q_1p_{11} + q_2p_{21}) + u_2(q_1p_{12} + q_2p_{22}) + u_3(q_1p_{13} + q_2p_{23})$$

$$= \sum_{j=1}^{3} u_j \log(\sum_{i=1}^{2} q_ip_{ij})$$

(2).

If $u_1 = 0.3, u_2 = 0.2, u_3 = 0.5$, we have:

$$\ell(q_1) = 0.3 * \log(q_1 * 0.1 + (1-q_1) * 0.4) + 0.2 * \log(q_1 * 0.2 + (1-q_1) * 0.5) + 0.5 * \log(q_1 * 0.7 + (1-q_1) * 0.1)$$

Finding the maximum:

$$\frac{d\ell(q_1)}{dq_1} = -\frac{0.03909}{-0.3q_1 + 0.4} - \frac{0.02606}{-0.3q_1 + 0.5} + \frac{0.13029}{0.6q_1 + 0.1}$$

$$= \frac{0.02345{q_1}^2 - 0.05120q_1 + 0.02306}{(0.6q_1 + 0.1)(-0.3q_1 + 0.4)(-0.3q_1 + 0.5)}$$

Let $0.02345{q_1}^2 - 0.05120q_1 + 0.02306 = 0$

$$q_1 = \frac{0.0512 - \sqrt{0.000458412}}{0.0469} = 0.635169336 \approx 0.635$$

$$q_2 = 1 - q_1 = 0.365$$

Since we got the value of q_1 and q_2, we can easily find each u_i:

$$u_1 = q_1p_{11} + q_2p_{21} = 0.2095$$

$$u_2 = q_1p_{12} + q_2p_{22} = 0.3095$$

$$u_3 = q_1 p_{13} + q_2 p_{23} = 0.4810$$

We verify the correctness:

$$u_1 + u_2 + u_3 = 0.2095 + 0.3095 + 0.4810 = 1$$