COMP9318: Assignment 1

Wanze Liu (z5137189)

UNSW

School of Computer Science and Engineering

Sydney, Australia

11/04/2019

Question 1

1.1

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

1.2

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

```
UNION ALL
SELECT Location, NULL, Item, SUM(Quantity)
FROM Sales
GROUP BY Location, Item
UNION ALL
SELECT Location, Time, NULL, SUM(Quantity)
FROM Sales
GROUP BY Location, Time
UNION ALL
SELECT NULL, NULL, Item, SUM(Quantity)
FROM Sales
GROUP BY Item
UNION ALL
SELECT NULL, Time, NULL, SUM(Quantity)
FROM Sales
GROUP BY Time
UNION ALL
SELECT Location, NULL, NULL, SUM(Quantity)
FROM Sales
GROUP BY Location
UNION ALL
SELECT NULL, NULL, NULL, SUM(Quantity)
FROM Sales
ORDER by Location, Time, desc;
```

1.3

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

1.4

We have original value mapping :

Sydney	1		2005	1		PS2	1
Melbourne	2		2006	2		Xbox 360	2
ALL	0		ALL	0		wii	3
						ALL	0

I choose the function as

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

Thus, we transfer the table into

Location	Time	Item	Quality	Offset
0	2	1	1500	9
0	2	3	500	11
0	2	0	2000	8
0	1	1	1400	5
0	1	2	1700	6
0	1	0	3100	4
0	0	1	2900	1
0	0	3	500	3
0	0	2	1700	2
0	0	0	5100	0
2	1	2	1700	30
2	1	0	1700	28
2	0	2	1700	26
2	0	0	1700	24
1	2	1	1500	21
1	2	3	500	23
1	2	0	2000	20
1	1	2	1400	18
1	1	0	1400	16
1	0	2	2900	14
1	0	3	500	15
1	0	0	3400	12

Quality	Offset		Quality	Offset		Dense MD array
1500	9		5100	0		5100
500	11		2900	1		2900
2000	8		1700	2		1700
1400	5		500	3		500
1700	6		3100	4		3100
3100	4		1400	5		1400
2900	1		1700	6		1700
500	3		2000	8		2000
1700	2		1500	9		1500
5100	0	========>	500	11	=========>	500
1700	30		3400	12		3400
1700	28		2900	14		2900
1700	26		500	15		500
1700	24		1400	16		1400
1500	21		1400	18		1400
500	23		2000	20		2000
2000	20		1500	21		1500
1400	18		500	23		500
1400	16		1700	24		1700
2900	14		1700	26		1700
500	15		1700	28		1700
3400	12		1700	30		1700

Question 2

2.1

Based on the Bayes rule , the classifer NB can been wirtten as follow

$$NB(x) = \begin{cases} 1 & , \dfrac{P(y=1|x)}{P(y=0|x)} \geq 1 \\ 0 & , \dfrac{P(y=1|x)}{P(y=0|x)} < 1 \end{cases}$$

now, We can determine the value of $\frac{P(y=1|x)}{P(y=0|x)}$,according to the formula below

$$P(y=1|x) = \frac{P(x|y=1)P(y=1)}{P(x)}$$

$$P(y=0|x) = \frac{P(x|y=0)P(y=0)}{P(x)}$$

then ,we can get

$$\frac{P(y=1|x)}{P(y=0|x)} = \frac{P(x|y=1)P(y=1)}{P(x|y=0)P(y=0)}$$

$$= \frac{P(y=1)\prod_{i=1}^{m} P(x_i|y=1)}{P(y=0)\prod_{i=1}^{n} P(x_i|y=0)}$$

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

We denote $p = P(y=1)$,then $1-p = P(y=0)$
$a_i = P(x=1|y=1)$,then $1-a_i = P(x=0|y=1)$

$$\text{So } ,P(x_i|y=1) = a_i^{x_i}(1-a_i)^{1-x_i}$$

$b_i = P(x=1|y=0)$,then $1-b_i = P(x=0|y=0)$

$$\text{And also } ,P(x_i|y=1) = b_i^{x_i}(1-b_i)^{1-x_i}$$

Then , we can get the formula

$$\frac{P(x|y=1)P(y=1)}{P(x|y=0)P(y=1)} = \frac{p}{1-p}\prod_{i=1}^{n}\frac{a_i^{x_i}(1-a_i)^{1-x_i}}{b_i^{x_i}(1-b_i)^{1-x_i}}$$

Then , we apply log caculation on both side of the formula and based on the hind provided , we can get

$$log\frac{P(x|y=1)P(y=1)}{P(x|y=0)P(y=1)} = log(\frac{p}{1-p}\prod_{i=1}^{n}\frac{a_i^{x_i}(1-a_i)^{1-x_i}}{b_i^{x_i}(1-b_i)^{1-x_i}})$$

$$= log\frac{p}{1-p} + \sum_{i=1}^{n} log\frac{a_i^{x_i}(1-a_i)^{1-x_i}}{b_i^{x_i}(1-b_i)^{1-x_i}}$$

$$= log\frac{p}{1-p} + \sum_{i=1}^{n} log\frac{a_i^{x_i}(1-a_i)^{-x_i}(1-a_i)}{b_i^{x_i}(1-b_i)^{-x_i}(1-b_i)}$$

$$= log\frac{p}{1-p} + \sum_{i=1}^{n} log\frac{1-a_i}{1-b_i} + \sum_{i=1}^{n} x_i log\frac{(1-b_i)a_i}{(1-a_i)b_i}$$

As we can know $log\frac{p}{1-p}$ and $\sum_{i=1}^{n} log\frac{1-a_i}{1-b_i}$ are constant number

So , we can get

$$b = log\frac{p}{1-p} + \sum_{i=1}^{n} log\frac{1-a_i}{1-b_i}$$

and

$$w_i = \frac{(1-b_i)a_i}{(1-a_i)b_i}$$

So ,the furmula we deduce below

$$b + \sum_{i=1}^{n} w_i x_i$$

which is the liner classifier

2.2

It is manily because naive Bayes classifier is simple to do perdicitions by appling the trained value directly, and all of dataset learned independently(caculate the value of $P(x)$, $P(y)$, $P(x|y)$, $P(y|x)$) , however , Logistic Regression classifier is more sophisticated than naive Bayes ,as it need to full search in data and more training process like Gradient Ascent or Dscent need to be applied to control the accuracy of convergence, and also dataset need to learn jointly, moreover , the data complexity requirement for learning w_{LR} is $O(n)$, while it is $O(log\ n)$ for learning w_{NB} whcih is smaller than w_{LR}.

question 3

3.1

0