UNSW
COMP9318
Data Warehousing and Data Mining 2019 T1

Written Assignment 1

Revision 1.0
Student: Lingxu Meng
zID: z5147810
Updated: 06th Apr 2019

Question 1

(1) List the tuples in the complete data cube of R in a tabular form with 4 attributes, i.e., Location, T ime, Item, SUM(Quantity)?

(2) Write down an equivalent SQL statement that computes the same result (i.e., the cube)

(3) Draw the result of the query in a tabular form.

(4) Draw the MOLAP cube (i.e., sparse multi-dimensional array) in a tabular form of (ArrayIndex, Value).

Question 2

(1) Prove that if the feature vectors are d-dimension, then a Naive Bayes classifier is a linear classifier in a d + 1-dimension space.

(2) Why learning wNB is much easier than learning wLR.

Question 3

(1) Write out the log likelihood function (as a function of qi , pi, j , and ui).

(2) What are the MLE of q1 and q2? What are the expected percentage of each component under a model with the MLE parameters?

Question 1

(1) List the tuples in the complete data cube of R in a tabular form with 4 attributes, i.e., Location, T ime, Item, SUM(Quantity)?

Location	Time	Item	SUM(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
Sydney	ALL	PS2	2900
Sydney	ALL	Wii	500
Sydney	ALL	ALL	3400
Melbourne	2005	ALL	1700
Melbourne	ALL	XBox 360	1700

Melbourne	ALL	ALL	1700
ALL	2005	PS2	1400
ALL	2005	XBox 360	1700
ALL	2006	PS2	1500
ALL	2006	Wii	500
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

(22 Rows...)

(2) Write down an equivalent SQL statement that computes the same result (i.e., the cube)

```
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) Draw the result of the query in a tabular form.

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) Draw the MOLAP cube (i.e., sparse multi-dimensional array) in a tabular form of (ArrayIndex, Value).

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

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	2	1	1500
0	2	3	500
0	1	0	3100
0	2	0	2000
0	0	1	2900
0	0	3	500
0	0	2	1700
0	0	0	5100

Convert to one-dimension:

Offset	**SUM(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
9	1500
11	500
4	3100
8	2000
1	2900
3	500
2	1700
0	5100

Question 2

(1) Prove that if the feature vectors are d-dimension, then a Naive Bayes classifier is a linear classifier in a d + 1-dimension space.

From Naive Bayes Classifier definition, plus **X** is a binary vector, we have:

When y = 1:

$$f(y=1) = \prod_{k=0}^{n} P(x_k|y=1)P(y=1) \tag{1}$$

When y = 0:

$$f(y=0) = \prod_{k=0}^{n} P(x_k|y=0)P(y=0) \tag{2}$$

We assume that (1) is larger than (2), than we have

$$\frac{P(y=1)}{P(y=0)} \cdot \prod_{k=0}^{n} \frac{P(x_k|y=1)}{P(x_k|y=0)} \geq 1 \tag{3}$$

Now, let $P(y=1) = p$, then obviously, $P(y=0) = 1-p$. And we let $P(x_k = 1|y=1|) = a_k$ and $P(x_k = 1|y=0|) = b_k$, according to $L(x) = p^y(1-p)^{1-y}$, (3) could be transmitted into:

$$\frac{p}{1-p} \cdot \prod_{k=0}^{n} \frac{a_k^{x_k}(1-a_k)^{(1-x_k)}}{b_k^{x_k}(1-b_k)^{(1-x_k)}} \geq 1$$

Collect constants together, the equation becomes:

$$(\frac{p}{1-p} \cdot \prod_{k=0}^{n} \frac{1-a_k}{1-b_k}) \cdot \prod_{k=0}^{n} (\frac{a_k}{b_k} \cdot \frac{1-b_k}{1-a_k})^{x_k} \geq 1 \tag{4}$$

Then use log to simplify (4) and according to $log \prod_i x_i = \sum_i log x_i$, we should get:

$$log(\frac{p}{1-p} \cdot \prod_{k=0}^{n} \frac{1-a_k}{1-b_k}) + \sum_{k=0}^{n} log(\frac{a_k}{b_k} \cdot \frac{1-b_k}{1-a_k}) x_k \geq 0 \tag{5}$$

The first item in (5) should be a constant because it doesn't have any x_k items. And we let $log(\frac{a_k}{b_k} \cdot \frac{1-b_k}{1-a_k})$ be w_k. So the final simplified equation should be:

$$c + \sum_{k=0}^{n} w_k x_k \geq 0$$

In this case, the predicting label is 1. $w^T = [1, x]$, it means that the classifier is linear.

(2) Why learning w_{NB} is much easier than learning w_{LR}.

Because in NB, we can learn from $P(y)$ and $P(x|c)$, which are condition independent, and it's very straightforward to calculate and build the training model from frequency. However, in LR, it requires a full research over the linear space of possible models, while also has a larger data requirement, which are $O(logn)$ and $O(n)$ respectively. So, NB is much easier.

Question 3

(1) Write out the log likelihood function (as a function of q_i , $p_{i,j}$, and u_i).

The likelihood function is $L(\mathbf{y}|\boldsymbol{\theta}) = P(\mathbf{y}|\boldsymbol{\theta})$ and we want to get $\widehat{\theta} \in \{\arg\max_{\theta\in\Theta} \mathcal{L}(\theta; x)\}$.

In this case, assume we are sampling from the mixture, and we can set each chemical in the sample are denoted as q_1 and q_2

From the construction of the compounds and mixture, we can know that $q_1 + q_2 = 1$, which is also $q_2 = 1 - q_1$. And $\sum_{i=1}^{m=3} u_i = 1$, two compounds each has this relationship: $\sum_{i=1}^{m=3} p_i = 1$, $\sum_{i=1}^{m=3} p_j = 1$

So, u_i in the whole mixture should be the add result of each two compounds, we get :
$u_i = p_i * qi + p_j * (1 - q_i) = p_j + q_i|(p_i - p_j|)$

Here, u_i is the observation, and q_i should be the proportion:

$$L(u_i; q_i) = \frac{p_{i,j}}{q_i(1-q_i)} p_{i,j}^{q_i} (p_{i,j})^{1-q_i}$$

$$logL(u_j; q_i) = log(\frac{p_{i,j}}{q_i(1-q_i)} p_{i,j}^{q_i} (p_{i,j})^{1-q_i}) = log\frac{p_{i,j}}{q_i(1-q_i)} + q_i log p_{i,j} + (1-q_i) log(p_{i,j})$$

(i = 1, 2; j = 1, 2, 3)

(2) What are the MLE of q_1 and q_2 ? What are the expected percentage of each component under a model with the MLE parameters?

From the function that we got from the first question, we can obtain the analytical solution by requiring:

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}_i} = \frac{q_i}{p_{i,j}} - \frac{1 - q_i}{p_{i,j}} = 0$$

If $u_1 = 0.3$, $u_2 = 0.2$, $u_3 = 0.5$, we can calculate that

When we use u_1, and we can use the function that we got from the first question

$$u_i = p_{1(2)} * q_i + p_{2(1)} * (1 - q_i) = p_{1(2)} + q_i |p_{2(1)} - p_{1(2)}|.$$

(Sorry, I really don't know how to accurately represent the difference of p...)

$$\frac{q_i}{p_{1,j}} = \frac{1 - q_i}{p_{2,j}}$$

If we know $p_{1,j}$, then $p_{2,j} = u_i - q_i |p_{1,j} - p_{2,j}|$

So, when we use p_1 for the calculation, we need to find a q that larger than 0 and smaller than 1. This $q_1 = 0.57$, and $q_2 = 0.43$
And we use p_2, so $q_2 = 0.36$, and $q_1 = 0.64$.
For u_2 when we use p_1 This $q_1 = 0.33$, $q_2 = 0.67$

And we use p_2, $q_2 = 0.11$, $q_1 = 0.89$.

For u_3, when we use p_1, This $q_1 = 0.31$, $q_2 = 0.69$.
And we use p_2, $q_2 = 0.59$, $q_1 = 0.41$.

From the possibilities of q above and according to percentages of each components in pure $S_i s$, we can easily get the expected percentage of each component under this MLE parameters which is most close to the result is:

$u_1 = 0.208$, $u_2 = 0.308$, $u_3 = 0.434$, at this time, $p_1 = 0.64$, $p_2 = 0.36$.