

COMP9318 (19T1) ASSIGNMENT 1

DUE ON 23:59 14 APR, 2019 (SUN)

Q1. (40 marks)

Consider the following base cuboid *Sales* with *four* tuples and the aggregate function SUM:

| <i>Location</i> | <i>Time</i> | <i>Item</i> | <i>Quantity</i> |
|-----------------|-------------|-------------|-----------------|
| Sydney | 2005 | PS2 | 1400 |
| Sydney | 2006 | PS2 | 1500 |
| Sydney | 2006 | Wii | 500 |
| Melbourne | 2005 | XBox 360 | 1700 |

Location, *Time*, and *Item* are dimensions and *Quantity* is the measure. Suppose the system has built-in support for the value **ALL**.

- (1) List the tuples in the complete data cube of *R* in a tabular form with 4 attributes, i.e., *Location*, *Time*, *Item*, SUM(*Quantity*)?
- (2) Write down an equivalent SQL statement that computes the same result (i.e., the cube). You can *only* use standard SQL constructs, i.e., no **CUBE BY** clause.
- (3) Consider the following *ice-berg cube* query:

```
SELECT Location, Time, Item, SUM(Quantity)
FROM Sales
CUBE BY Location, Time, Item
HAVING COUNT(*) > 1
```

Draw the result of the query in a tabular form.

- (4) Assume that we adopt a MOLAP architecture to store the full data cube of *R*, with the following mapping functions:

$$f_{Location}(x) = \begin{cases} 1 & \text{if } x = \text{'Sydney'}, \\ 2 & \text{if } x = \text{'Melbourne'}, \\ 0 & \text{if } x = \mathbf{ALL}. \end{cases}$$

$$f_{Time}(x) = \begin{cases} 1 & \text{if } x = 2005, \\ 2 & \text{if } x = 2006, \\ 0 & \text{if } x = \mathbf{ALL}. \end{cases}$$

$$f_{Item}(x) = \begin{cases} 1 & \text{if } x = \text{'PS2'}, \\ 2 & \text{if } x = \text{'XBox 360'}, \\ 3 & \text{if } x = \text{'Wii'}, \\ 0 & \text{if } x = \mathbf{ALL}. \end{cases}$$

Draw the MOLAP cube (i.e., sparse multi-dimensional array) in a tabular form of $(ArrayIndex, Value)$. You also need to write down the function you chose to map a multi-dimensional point to a one-dimensional point.

Q2. (30 marks)

Consider binary classification where the class attribute y takes two values: 0 or 1. Let the feature vector for a test instance be a d -dimension column vector \vec{x} . A linear classifier with the model parameter \mathbf{w} (which is a d -dimension column vector) is the following function:

$$y = \begin{cases} 1 & , \text{ if } \mathbf{w}^\top \mathbf{x} > 0 \\ 0 & , \text{ otherwise.} \end{cases}$$

We make additional simplifying assumptions: \mathbf{x} is a binary vector (i.e., each dimension of \mathbf{x} take only two values: 0 or 1).

- Prove that if the feature vectors are d -dimension, then a Naïve Bayes classifier is a linear classifier in a $d + 1$ -dimension space. You need to explicitly write out the vector \mathbf{w} that the Naïve Bayes classifier learns.
- It is obvious that the Logistic Regression classifier learned on the same training dataset as the Naïve Bayes is also a linear classifier in the same $d + 1$ -dimension space. Let the parameter \mathbf{w} learned by the two classifiers be \mathbf{w}_{LR} and \mathbf{w}_{NB} , respectively. Briefly explain why learning \mathbf{w}_{NB} is much easier than learning \mathbf{w}_{LR} .

Hint 1. $\mathbf{w}_{NB} = \mathbf{w}_{LR} + \mathbf{1}$

Q3. (30 marks)

We have a sample of mixture of two chemical compound, S_1 and S_2 . The (unknown) percentages of each chemical in the sample are denoted as q_1 and q_2 (whereas $q_1 + q_2 = 1$), respectively.

We have a device that can detect the percentages of $m = 3$ different components that are contained in both chemical compounds, albeit with different percentages. We denote the components as $\{O_j\}_{j=1}^m$. We list the percentages of each components in pure S_i s in the following table:

| $p_{i,j}$ | O_1 | O_2 | O_3 |
|-----------|-------|-------|-------|
| S_1 | 0.1 | 0.2 | 0.7 |
| S_2 | 0.4 | 0.5 | 0.1 |

After measuring the three components, we obtain their percentages as $\{u_j\}_{j=1}^m$.

- (1) Write out the log likelihood function (as a function of q_i , $p_{i,j}$, and u_i).
- (2) If $u_1 = 0.3, u_2 = 0.2, u_3 = 0.5$, 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?

SUBMISSION

Please write down your answers in a file named `ass1.pdf`. You **must write down your name and student ID on the first page**.

You can submit your file by
`give cs9318 ass1 ass1.pdf`

Late Penalty. -10% per day for the first two days, and -20% for each of the following days.