COMP9318 (18S1) ASSIGNMENT 1

DUE ON 23:59 23 MAY, 2018 (WED)

Q1. (*40 marks*)

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

Location	*Time*	*Item*	*Quantity*
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., $Location, Time, Item, \mathrm{SUM}(Quantity)$?

(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:

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

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

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

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-dimensioinal 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 $\mathbf{w}_{\text{LR}}$ and $\mathbf{w}_{\text{NB}}$, respectively. Briefly explain why learning $\mathbf{w}_{\text{NB}}$ is much easier than learning $\mathbf{w}_{\text{LR}}$.

Hint 1. $\log \prod_i x_i = \sum_i \log x_i$

Q3. (*30 marks*)

Consider a dataset consisting of n training data $\mathbf{x}_i$ and the corresponding class label $y_i \in \{0, 1\}$.

(1) Consider the standard logistic regression model:

$$P[y = 1 \mid \mathbf{x}] = \sigma(\mathbf{w}^\top \mathbf{x})$$

where σ is the sigmoid function.

The learning of the model parameter is to find $\mathbf{w}^*$ that minimizes some function of $\mathbf{w}$, commonly known as the *loss function*.

Prove that the loss function for logistic regression is:

$$\ell(\mathbf{w}) = \sum_{i=1}^{n} \Big(-y_i \mathbf{w}^\top \mathbf{x}_i + \ln(1 + \exp(\mathbf{w}^\top \mathbf{x}_i)) \Big)$$

(2) Consider a variant of the logistic regression model:

$$P[y = 1 \mid \mathbf{x}] = f(\mathbf{w}^\top \mathbf{x})$$

where $f : \Re \to [0, 1]$ is a squashing function that maps a real value to a value between 0 and 1.

Write out its loss function.

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.