COMP9318 (21T1) ASSIGNMENT 1

DUE ON 20:59 16 APR, 2021 (FRI)

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 c tributes, i.e., *Location*,

(2) Write down an equi cube). You can **E 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:

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

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

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

If we want to draw the MOLAP cube (i.e., sparse multi-dimensional array) in a tabular form of $(ArrayIndex, Value)$, then which of the following function is feasible? Why? You also need to draw the MOLAP cube.

- $f(x) = 9 \cdot f_{Location}(x) + 3 \cdot f_{Time}(x) + f_{Item}(x)$
- $f(x) = 16 \cdot f_{Location}(x) + 4 \cdot f_{Time}(x) + f_{Item}(x)$

Q2. (*30 marks*)

Consider the following training examples which are used to construct a decision tree to help predict whether a patient is likely to have a lung cancer.

Patient ID	Gender	Smokes?	Chest pain?	Cough?	Lung Cancer
1	Female	Yes	Yes	Yes	Yes
2	Male	Yes	No	Yes	Yes
3	Male	No	No	No	Yes
					No
					Yes
					No

(1) Use Gini index to construct a decision tree that pred to have a lung cancer. You need to show every step of t

(2) Translate your decision tree into decision rules.

Q3. (*30 marks*)

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

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

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).

(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.

(2) 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. $\log \prod_i x_i = \sum_i \log x_i$

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. 0 mark if not submit on time (i.e., firm deadline).