Name: __________________________ , __________________________

(Family name) (Given name)

Student ID: __________________________

THE UNIVERSITY OF NEW SOUTH WALES
Final Exam

COMP9318
Data Warehousing and Data Mining

TERM T1, 2021

- Time allowed: **10 minutes** reading time + **2 hours** + **20 minutes** submitting time
  - Exception: students with extra exam time approved by **Equitable Learning Services (ELS)** can make submissions after 11:30, 7 May 2021 within their **approved extra time**.
- Total number of questions: **6**.
- Total number of marks: **100**
- Total number of pages: **6 excluding this cover page**
- This is an open-book exam. You are allowed to use textbook(s), lecture notes and other study materials. However, you are **not** allowed to (1) communicate with anyone else or (2) use the Internet during the exam.
- Items allowed: UNSW approved calculators.
- You can answer the questions in any order.
- Start each question on a **new page**.
- Write your name and student id on each page.
- Answers must be written in ink on A4 papers and scanned into a PDF file. Alternatively, you can use any software to directly generate the answers in a PDF file.

SECTION A: Data Warehousing

Question 1 (16 marks)

Consider a fact table with three dimensions A, B, and C, and one measure M. The only hierarchies on the dimensions are from the dimension values to **ALL**.

(a) We consider the lattice of the cuboids. How many *cuboids* are there in the complete data cube? You need to justify your answer.

(b) What's the minimum and maximum possible number of *tuples* in the complete data cube, given that the fact table contains 3 tuples. You need to justify your answer.

(c) Given the below fact table. When using the BUC algorithm *with single-tuple optimization* as described in Lab-2 to generate the compelete data cube, how many times the single-tuple optimization trick will be applied? You need to justify your answer by showing the steps.

A	B	C	M
1	1	2	10
1	2	1	20
2	1	1	30

SECTION B: Cluster Analysis

Question 2 (16 marks)

Given the following 2-dimensional data points, you need to simulate the k-means algorithm.

ID	x_1	x_2
1	1.90	0.97
2	5.98	2.68
3	2.68	1.18
4	3.14	4.24
5	1.54	1.80
6	3.82	4.50
7	5.74	3.84
8	2.46	1.86
9	3.17	4.96
10	5.44	3.18

Suppose the coordinates of data points 1, 6 and 10 are used as the initial cluster centers for clusters A, B and C, respectively. Please simulate the k-means algorithm with Euclidean distance for 1 iteration.

(a) What are the cluster assignments after 1 iteration? You need to justify your answer.

(b) What are the new cluster centers for clusters A, B and C, respectively? You need to justify your answer.

(c) In the k-means algorithm, when using Manhattan distance (i.e., ℓ_1 distance, where $dist(x, y) = \sum_{i=1}^{d} |x_i - y_i|$) instead of Euclidean distance (i.e., ℓ_2 distance) to measure the distance between two objects (e.g., points), should we change the way of choosing centroids? If not, you need to explain why. If yes, you need to give the correct way of choosing centroids and justify your answer.

Question 3 (16 marks)

Consider the **distance** matrix in the table below to perform hierarchical clustering.

	p_1	p_2	p_3	p_4	p_5
p_1	0.00	0.10	0.41	0.55	0.35
p_2	0.10	0.00	0.64	0.47	0.98
p_3	0.41	0.64	0.00	0.44	0.85
p_4	0.55	0.47	0.44	0.00	0.76
p_5	0.35	0.98	0.85	0.76	0.00

(a) Show the steps and final result of running the *group average* hierarchical clustering algorithm. Show your final results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged.

(b) Consider the following top-down algorithm

1. build a complete weighted undirected graph $G = (P, E)$ where P is the set of all the objects, and E is the set of edges between any two objects in P. We assign the corresponding distance between p_i, p_j as the weight to the edge $< p_1, p_2 >$. E.g., edge weight of $< p_1, p_2 >$ is 0.10.
2. Generate the Minimum Spanning Tree of the graph G, denoted as $G' = (P, E')$.
3. Remove the edge e from G', where e has the largest weight in E'. Then we have two sub-trees, which correspond to the two clusters in the top-level.
4. Repeat step 3 recursively on the generated sub-trees to further split the corresponding clusters, until there are no more than 2 nodes in one tree.

What is the equivalent bottom-up hierarchical clustering algorithm to the above one? You need to justify your answer.

SECTION C: Classification

Question 4 (20 marks)

Consider the following training dataset.

id	a_1	a_2	a_3	*class*
1	T	T	1.0	Y
2	T	T	6.0	Y
3	T	F	5.0	N
4	F	F	4.0	Y
5	F	T	7.0	N
6	F	T	3.0	N
7	F	F	8.0	N
8	T	F	7.0	Y
9	F	T	5.0	N

(a) Assume a_3 values have to be discretised to a'_3 as follows:

a_3	a'_3
$0.0 \leq a_3 < 3.0$	L
$3.0 \leq a_3 < 6.0$	M
$6.0 \leq a_3 < 9.0$	H

Show the dataset after applying the above transformation. For the rest of the questions, we will use the transformed dataset (i.e., consisting of attributes a_1, a_2 and a'_3).

(b) What is the **first** splitting condition used by the CART decision tree induction algorithm (i.e., the algorithm that learns a binary decision tree using Gini index as the splitting criterion)? You need to justify your answer.

(c) Show the steps and the classification result a new tuple $(10, \mathrm{T}, \mathrm{F}, \mathrm{X})$ as predicted by a Naive Bayes classifier built on the above training dataset. You should use **add-0.5 smoothing** (i.e., similar to add-1 smoothing, but add 0.5 instead of 1 to the numerator. you will also need to work out the correct B to add to the denominator) when estimating the conditional probabilities. Note that "X" is a value that has never been encountered in the training dataset.

Question 5 (16 marks)

Suppose we are training a linear SVM (i.e., no kernel) with only have 6 training examples in 2-dimensional Euclidean space, which are:

- positive examples: $x_1 = (4, 1)$, $x_2 = (5, 2)$, and $x_3 = (4, 3)$,
- negative examples: $x_4 = (1, 0)$, $x_5 = (2, 1)$, and $x_6 = (h, 2)$.

Where $0 \leq h \leq 6$ is a parameter.

(a) Give the range of h where the training examples are linearly separable. You need to justify your answer.

(b) Let $h = 1$, what is the functional margin of the training dataset? You need to justify your answer.

(c) Give the range of h where the decision boundary will remain the same as when $h = 1$. You need to justify your answer.

(d) We call a training example "critical" if removing it from the training set and re-training SVM will get a different decision boundary to the one that is trained on the full training set. Let $h = 1$, list all the "critical" training examples. You need to justify your answer.

SECTION D: Association Rule Mining

Question 6 (16 marks)

Let $sup(A)$ be the support of A, and $conf(A \to B)$ be the confidence of the rule $A \to B$.

(a) Given itemset S, let $S' \subseteq S$ and $S' \neq \emptyset$. Prove that $sup(S') \geq sup(S)$.

(b) Given frequent itemset A and $S' \subseteq S \subseteq A$. Prove that $conf(S \to A \setminus S) \geq conf(S' \to A \setminus S')$.

(c) We have discussed a method to generate association rules from frequent itemsets (e.g., page 17 of the slides). Propose a new method that is more efficient. You can use the above two statements no matter whether you have proved them or not. You also need to explain why it is more efficient than the original one.

END OF EXAM PAPER