THE UNIVERSITY OF AUCKLAND

SEMESTER ONE 2020
Campus: City

COMPUTER SCIENCE

Advanced Machine Learning

NOTE:

- This Final Assessment is out of **100** marks.
- Attempt **ALL** questions.
- You will need to put your answers into the **Answer Booklet** and save it as a **pdf document**. Upload the file on Canvas, like you do for the assignments. You are allowed up to a **MAXIMUM** of 5 submission attempts.
- This is an **Open Book** assessment. You may refer to and cite any written/printed material, including online sources.
- You need to demonstrate your **understanding of the subject matter** and the ability to **construct a well described solution or organised arguments** to answer the question(s). Quotations (if used) should be used rarely and selectively.
- You should include proper **referencing** of any material you have used (including author and year of publication). It is important that you do not just provide a list of quotations. Quotations should be used to support your own argument not replace it.
- When a question requests you to explain your answer, that means you need to justify how you came up with the solution or why you made a certain choice. Be concise and as clear as possible.
- You may choose to use diagram(s) to aid in your discussion. If you choose to do so, you may embed photo(s) of hand drawn diagram(s) into the answer booklet. **It is your responsibility to ensure that the diagrams are clear, legible, and have proper resolution.**
- Please use standard text processing tools to type the answers and avoid hand-written answers where possible. Use images only when it is required.
- This Final Assessment has been designed so that a well-prepared student could complete it within 2 hours.

- If you wish to raise concerns during the Final Assessment, please call the Contact Centre for advice: Auckland: 09 373 7513, Outside Auckland: 0800 61 62 63, International: +64 9 373 7513
- It is your responsibility to ensure your assessment is successfully submitted on time. Please don’t leave it to the last minute to submit your assessment.
- For any Canvas issues, please use 24/7 help on Canvas by chat or phone.
- If any corrections are made during the 24 hours, you will be notified by a Canvas Announcement. Please ensure your notifications are turned on during this period.

Academic honesty declaration

By completing this assessment, I agree to the following declaration:

I understand the University expects all students to complete coursework with integrity and honesty. I promise to complete all online assessments with the same academic integrity standards and values. Any identified form of poor academic practice or academic misconduct will be followed up and may result in disciplinary action.

As a member of the University's student body, I will complete this assessment in a fair, honest, responsible and trustworthy manner. This means that:

- *I declare that this assessment is my own work.*
- *I will not seek out any unauthorised help in completing this assessment.*
- *I am aware the University of Auckland will use plagiarism detection tools to check my content.*
- *I will not discuss the content of the assessment with anyone else in any form, including, Canvas, Piazza, Facebook, Twitter or any other social media or online platform within the assessment period.*
- *I will not reproduce the content of this assessment anywhere in any form at any time.*
- *I declare that I generated the calculations and data in this assessment independently, using only the tools and resources defined for use in this assessment.*
- *I will not share or distribute any tools or resources I developed for completing this assessment.*

Answer ALL questions.

PART A: Association Rule Mining, Clustering, Data Stream Mining, Anomaly Detection

Question 1 **15 marks**

(a) If you were given two of the following rules, which of the rules would you consider more interesting and why? Please discuss your answer.

Rule 1: $X1 \rightarrow Y1$, rule support of 100% and rule confidence of 100%.

Rule 2: $X2 \rightarrow Y2$, rule support of 30% and rule confidence of 95%.

(7 marks)

The text-based answer (excluding diagrams, equations and calculations) should be 100 words or less.

(b) Consider a transactional database where **A, B, C, D, E, F, G, H** are items. Using Apriori, find all frequent itemsets when we set the minimum support to 59%. Please indicate the candidate itemsets and frequent itemsets separately. You MUST show your working, similar to what is shown in the "Pattern Mining: Association Analysis I" lecture in Week 5. You may choose to use diagram(s) to aid in your discussion.

TID	Items
1	A, B, C, D, E, G
2	A, B, C, D, E, G
3	A, B, C, D, E, G
4	A, B, C, D, E, G
5	A, B, C, E, G
6	A, B, C
7	A, B, C, D, F, H
8	A, B, C, D, E, G
9	A, B, C, D, G
10	A, B, C, D, E, F, H

(5 marks)

The text-based answer (excluding diagrams, equations and calculations) should be 350 words or less.

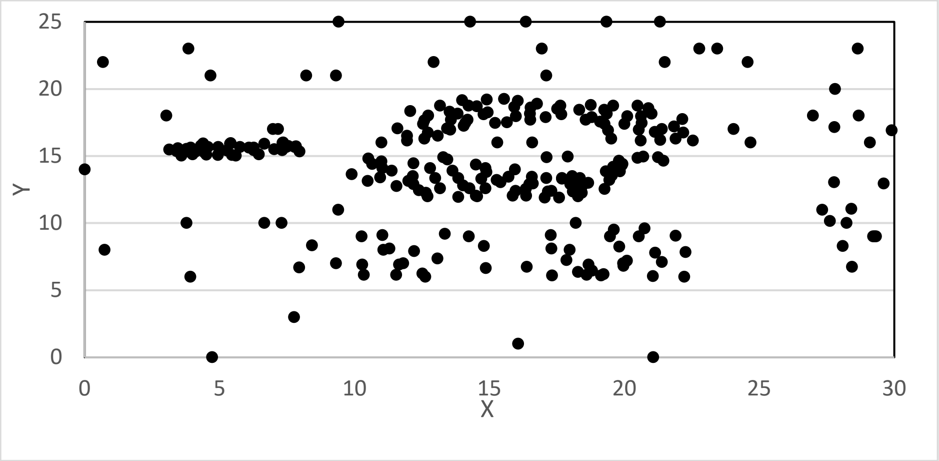
(c) Create an example dataset with 10 transactions that intuitively will have better computational (time and/or memory) performance when using the FP-Growth algorithm as compared with the Apriori algorithm. Show the dataset and explain your answer.

(3 marks)

The text-based answer (excluding diagrams, equations and calculations) should be 100 words or less.

Question 2 **12 marks**

You are given the following dataset visualised in a two-dimensional plot.



Discuss the clustering process step-by-step and its output(s), when you cluster the dataset (above) using each of the follow algorithms separately (you must show the use of both for full marks):

(a) K-means, and

(b) DBSCAN.

Based on your informed judgement, please discuss the appropriate parameter settings for these tasks as well. You may choose to use diagram(s) to aid in your discussion.

(12 marks)

The text-based answer (excluding diagrams, equations and calculations) should be 300 words or less.

Question 3 **10 marks**

(a) You are tasked to design a predictive system for a data stream. The system needs to fulfil the following properties:

i) the ability to adapt to changes in the data stream,

ii) the ability to signal drift explicitly,

iii) the ability to train a new model that is computationally inexpensive.

You are required to design a system that can fulfil the criteria above. Please discuss the change management strategy, the choices of the concept drift detector and classifier, the interactions between the components, and any other salient points. You may choose to use diagram(s) to aid in your discussion.

(6 marks)

The text-based answer (excluding diagrams, equations and calculations) should be 300 words or less.

(b) The ADWIN drift detector uses an exponential histogram. If we replaced the exponential histogram with an equal width histogram, what impact does that have on the performance of the drift detector? Discuss you answer.

(4 marks)

The text-based answer (excluding diagrams, equations and calculations) should be 100 words or less.

Question 4 **13 marks**

(a) Supervised anomaly detection can outperform unsupervised anomaly detection in terms of higher detection rate and lower false alarm rate, when training data is available. Discuss a scenario where this is not the case. Explain how an unsupervised technique would be more useful for the specific case.

(8 marks)

The text-based answer (excluding diagrams, equations and calculations) should be 150 words or less.

(b) You are given the following list of 2D data points.

```
[[1,1],[1,2],[2,2],[2,1],[3,3],[2,5],[2,3]]
```

If you had to select one point to be anomalous, which would you pick? Explain your answer. Please link your explanation to an anomaly detection technique.

(5 marks)

The text-based answer (excluding diagrams, equations and calculations) should be 100 words or less.

PART B: Preprocessing, Instance-based Learning, SVM, Bayes Learning, Reinforcement Learning

Question 5 **15 marks**

You are given the following data set:

	F_1	F_2	F_3	Target
Train	0	100	High	A
	0	50	High	A
	1	0	Low	B
	1	10	Low	B
Test	0	10	High	B
	1	5	Low	B
	1	50	High	A

Train a 1-Nearest Neighbour classifier on the training set and apply it on the test set. Use the following distance metric as a measure of closeness in the input space.

$$D(\vec{x}, \vec{y}) = \sum_{i=1}^{N} |x_i - y_i|$$

Compare its performance on the training set and the test set in terms of accuracy and discuss if you trained a good classifier. Show the modelling process step-by-step and explain your answer.

(15 marks)

The text-based answer (excluding diagrams, equations and calculations) should be 200 words or less.

Question 6 **12 marks**

You are given a small training data set with two input features F_1 and F_2.

ID	F_1	F_2	Target
A	1	3	positive
B	2	3	positive
C	1	2	positive
D	1	4	positive
E	3	2	negative
F	4	2	negative
G	3	1	negative
H	4	3	negative

(a) If you train a linear hard margin support vector machine on this data set, what will the decision boundary look like? Draw and properly annotate both the decision boundary and the margins in the input space. What are the support vectors for this data set? How many examples will be miss-classified on the training set with this hard margin classifier? Will this model perform well on unseen data?

(b) Now let us say you train the same method but on different sub-samples (that will happen, for example, if you used cross-validation). How much will the decision boundary change for different sub-samples? Give two sub-samples that will change the decision boundary and another two sub-samples that will not change the decision boundary.

Explain your answers.

(12 marks)

The text-based answer (excluding diagrams, equations and calculations) should be 200 words or less.

Question 7 **10 marks**

The continuation of a project of a biotech company depends on the status of the last experiment. If it gives positive results, the project will be continued, if not, it will be stopped. Additionally, the company needs more money from the bank to continue the project. In the past, the company asked four times for a loan and the bank approved it in three cases. It is probable, that the lab worker, who is working on the experiment, is not concentrating enough and hence changes the result of the experiment. This happens once in 10 experiments. We know that the lab equipment used to perform the experiment experiences technical failure in 1% of the performed experiments. We also know that occasionally the head of the lab misinterprets the experimental results, i.e. in one out of eight experiments. Design a Bayesian network from the variables L (lab worker is not concentrating at work), E (experimental result is positive), Z (lab equipment is broken), P (project is continued), A (head of lab misinterprets the results), and B (bank gives money).

(a) Draw the network and comment why you designed it like that. What the network represents? Which variables are conditionally independent? Explain your answers.

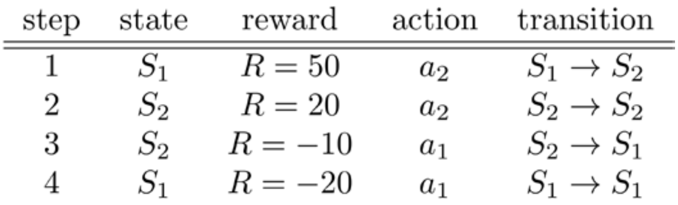
(b) Write the conditional probability table for all nodes, and fill them in based on the information provided in the text. How many probability values are missing to fully define the network? Write the expression for the joint probability distribution of L, E, Z, P, A, and B in its factored form. Show your working and explain your decisions.

(10 marks)

The text-based answer (excluding diagrams, equations and calculations) should be 200 words or less.

Question 8 **13 marks**

You have agent that can interact with a very simple environment, described by two states S_1 and S_2. In each state the agent can take two deterministic actions a_1 and a_2. In order to achieve its goals the agent uses Q-learning algorithm. Given that the discount factor is 0.1, and the first four steps are carried out as follows:



step	state	reward	action	transition
1	S_1	$R = 50$	a_2	$S_1 \rightarrow S_2$
2	S_2	$R = 20$	a_2	$S_2 \rightarrow S_2$
3	S_2	$R = -10$	a_1	$S_2 \rightarrow S_1$
4	S_1	$R = -20$	a_1	$S_1 \rightarrow S_1$

return the state of the Q-table after each step, and π and V after these four steps. Show your working step-by-step, i.e. how you calculated each value in Q, V, and π. Which state is more desired for the agent? Is π= π* in this case? Explain your answer.

(13 marks)

The text-based answer (excluding diagrams, equations and calculations) should be 200 words or less.