Institute of Systems Science
National University of Singapore

MASTER OF TECHNOLOGY IN INTELLIGENT SYSTEMS

Graduate Certificate Online Examination
Semester I 2020/2021

Subject: *Pattern Recognition Systems*

Instructions for Paper

Date: Saturday 7 Nov 2020
Time: 9.30 a.m.
Duration: Three hours (10.00 a.m. to 1.00 p.m.)
Place: Online Examination

This is an OPEN BOOK examination. This examination paper consists of *one* Section and *two* Questions. You are to answer *ALL* questions. There are a total of *40 Marks* for this paper.

1. Read **ALL** instructions before answering any of the examination questions.
2. There will be only **ONE question paper**, which may be organized in a series of one or more sections (section A, B, C etc.) with/without appendixes. Each section contains one or more questions. You are required to answer **ALL** questions in the **SEPARATE answer booklet(s)**, according to different sections, downloaded from LumiNUS.
3. The first 30 minutes will be reading time, during which you **must not** start answering your answers.
4. Write your NUS Student ID number on the **front page** of the Answer Booklet(s) in the box provided.
5. This is an ***Open Book*** examination. If you wish, you may use reference materials to answer a question. Reference materials can be books, manuals, handouts or notes, including e-notes on PCs/laptops.
6. All answers provided should be in **digital format**. However, you can **hand draw** diagrams on paper using pens and ensure that it is readable as an image. **Insert this image into your answer booklet**. Do not send images as separate files.
7. Non-programmable calculators may be used if required.
8. **Internet access (except LumiNUS) using computers of any form e.g. laptops, tablets, smart watches etc. is not permitted during the examination**. Students who are found with suspected academic dishonesty that give them unfair advantage during assessments will be subjected to disciplinary action by the University, as laid out in NUS Code of Student Conduct.
9. State clearly any assumptions you make in answering any question where you feel the requirement is not sufficiently clear.
10. At the end of the examination:
    a) Convert the answer booklet to PDF format and compress them if necessary.
    b) Please name the PDF file with your **Student ID number** prefaced by the abbreviation of the Course and the Section e.g. **PRS_SecA_A0123456X**.
    c) Upload the answer booklet for **each Section separately**.
11. After submission of answer booklet(s), please wait for Proctor to make announcement on the closure of examination.

SECTION A

Question 1 *(Total: 20 Marks)*

Telemarketing is the direct marketing of goods or services to potential customers over the telephone, Internet, or fax. Although telemarketing is generally considered a more cost-effective marketing strategy as compared to roadshows, companies can further optimize their budget by using AI techniques to target the right consumers.

You are a machine learning engineer working in the Sintosa Bank, supporting telemarketing campaigns. You are tasked to build a pattern recognition system to predict if the bank's customers will subscribe to a term deposit. The bank aims to hit as many subscribers as possible, and at the same time find out the common characteristics of customers who are more likely to subscribe to a term deposit. You are told by the bank that they have a different budget for each telemarketing campaign and they always prefer to use models that are simpler and faster to run.

The bank has accumulated some telemarketing data of promoting term deposits over the last five years. The data set consists of 41,188 data points with 20 variables out of which 8 are numeric variables and 12 are categorical variables. Only 12% of the records are related to the customers who have subscribed to a term deposit.

The list of variables are given as follows:

V1 - customerID (categorical)

V2 - age (numeric)

V3 - job: type of job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'management', 'retired', 'self-employed', 'unemployed', 'unknown')

V4 - marital: marital status (categorical: 'divorced', 'married', 'single', 'unknown')

V5 - education (categorical: 'illiterate', 'high.school', 'university.degree', 'unknown')

V6 - default: has credit in default? (categorical: 'no', 'yes', 'unknown')

V7 - housing: has housing loan? (categorical: 'no', 'yes', 'unknown')

V8 - loan: has personal loan? (categorical: 'no', 'yes', 'unknown')

V9 - contact: contact communication type (categorical: 'cellular', 'telephone')

V10 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')

V11 - day_of_week: last contact day of the week (categorical: 'mon', 'tue', 'wed', 'thu', 'fri')

V12 - duration: last contact duration, in seconds (numeric).

V13 - campaign: number of contacts performed during this campaign and for this customer (numeric, includes last contact)

V14 - pdays: number of days that passed by after the customer was last contacted from a previous campaign (numeric)

V15 - previous: number of contacts performed before this campaign and for this customer (numeric)

V16 - poutcome: outcome of the previous marketing campaign (categorical: 'failure', 'nonexistent', 'success')

V17 - emp.var.rate: employment variation rate - quarterly indicator (numeric)

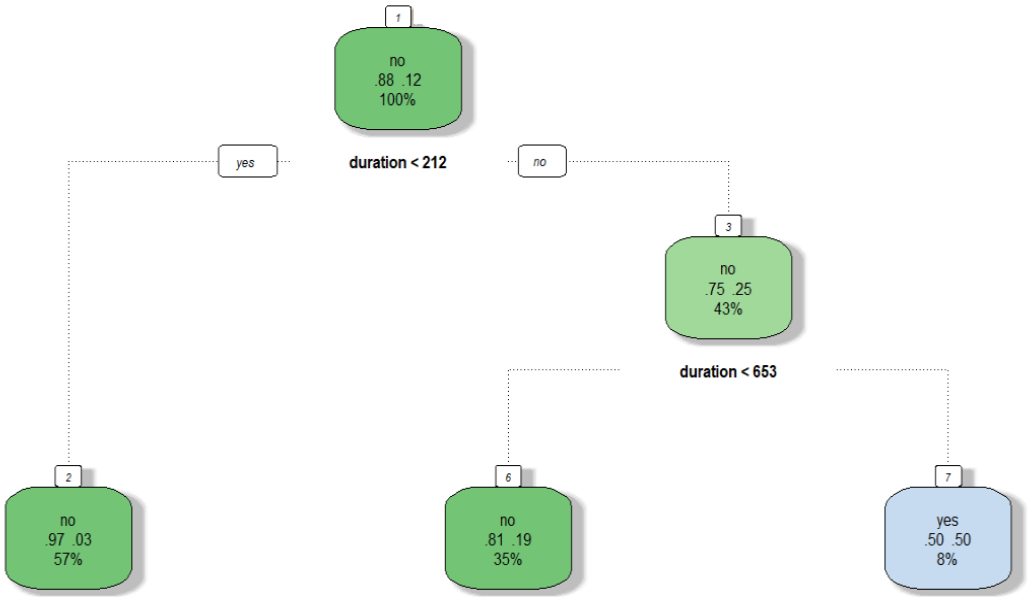
V18 - cons.price.idx: consumer price index - monthly indicator (numeric)

V19 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)

V20 - subscribed: has the customer subscribed to a term deposit? (binary: 'yes', 'no')

Answer the following questions:

a. You aim to generate a list of customers who are more likely to subscribe to term deposits and learn the common characteristics of such customers. You have built a decision tree model using V2 - V19 as the input variables and V20 as the output variable with 70% of the historical data used to train the model. The obtained model is shown below. Identify one problem with the current model and explain the main cause of this problem. What would you suggest to improve the model further?



(4 Marks)

b. You continue to build Support Vector Machine (SVM) models for the same task in part (a). You have used two different SVM kernels, i.e. linear kernel and RBF kernel. The table below shows the performance comparison of these two SVM models on the training and test sets. How would you choose which model to deploy based on the case study scenario and other business considerations you may assume?

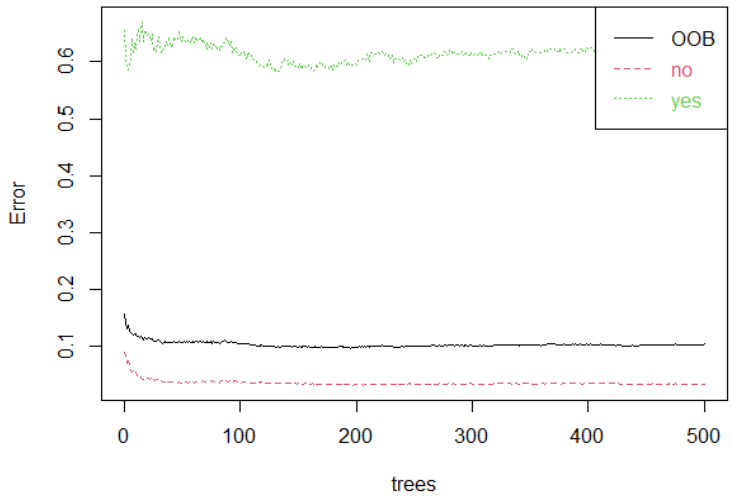
Models	**Overall Accuracy on Training Data (%)**	**Overall Accuracy on Test Data (%)**	**True Positive Rate on Test Data (%)**
Model A: SVM using linear kernel	81.5	75.4	68.2
Model B: SVM using RBF kernel	82.4	78.1	63.4

(3 Marks)

c. Due to the varying budget for each telemarketing campaign, the maximum number of customers the bank can target for each campaign is different. You are asked to automate the deployment of the model built in part (b) for multiple campaigns. How would you design the deployment scheme to meet such a requirement? Furthermore, is it possible to identify the optimal number of customers to target for each campaign? Justify your answer. State any additional information or data if required for this task.

(4 Marks)

d. You explore further using the random forest approach. A random forest model has been built using 500 trees and 6 candidate input variables for each splitting. The figure below shows the evaluation result of the random forest model on the Out-Of-Bag (OOB) data set. The three curves in the figure show the error rates on the whole set of OOB data, the OOB data belonging to the category "no", and the OOB data belonging the category "yes", with respect to the different number of trees being added into the random forest. What conclusion(s) can you draw from this evaluation result? What would you suggest to improve this random forest model?



(3 Marks)

e. You have decided to investigate the use of clustering to find out if there are natural groupings of customers who are likely to subscribe to a term deposit. Due to the large number of features, you decided to select only the following numeric features for clustering:

	count	mean	std	min	25%	50%	75%	max
age	41188.0	40.02	10.42	17.0	32.00	38.00	47.00	98.00
default	41188.0	0.10	0.30	0.0	0.00	0.00	0.00	1.00
housing	41188.0	0.54	0.49	0.0	0.00	1.00	1.00	1.00
loan	41188.0	0.16	0.36	0.0	0.00	0.00	0.00	1.00
cons.price.idx	41188.0	93.58	0.58	92.2	93.08	93.75	93.99	94.77
campaign	41188.0	2.57	2.77	1.0	1.00	2.00	3.00	56.00
pdays	41188.0	15.59	9.01	0.0	8.00	15.00	23.00	31.00
previous	41188.0	0.17	0.49	0.0	0.00	0.00	0.00	7.00
subscribed	41188.0	0.16	0.36	0.0	0.00	0.00	0.00	1.00
divorced	41188.0	0.11	0.32	0.0	0.00	0.00	0.00	1.00
married	41188.0	0.61	0.49	0.0	0.00	1.00	1.00	1.00
single	41188.0	0.28	0.45	0.0	0.00	0.00	1.00	1.00
unknown	41188.0	0.00	0.04	0.0	0.00	0.00	0.00	1.00

Take note that missing values (as defined in the case study description – eg. "unknown") have been imputed using the folllowing code:

```
missing = data.isnull().sum()

if sum(missing) > 0:
    data = data.fillna(data.mean())

data.isnull().sum()
```

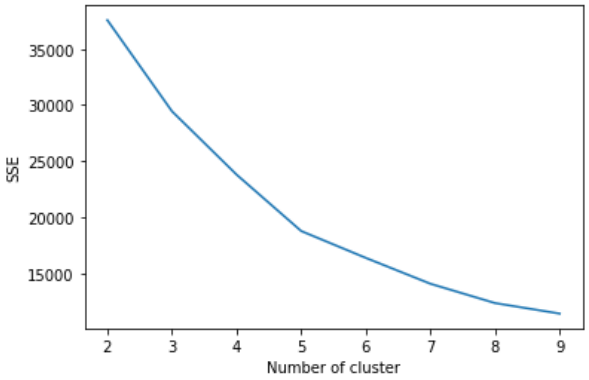
Examine the features table above and critique the feature selection and data preparation methods. Identify at least 2 issues you foresee that will impact the subsequent cluster analysis task. For each issue, suggest one alternative method that would work better.

(3 Marks)

f. You have executed the clustering and the results are as follows.

```
K = 2  Silhouette =  0.3328295289539971
K = 3  Silhouette =  0.36966016364972204
K = 4  Silhouette =  0.4269096838701964
K = 5  Silhouette =  0.4058551549982063
K = 6  Silhouette =  0.43474662947397935
K = 7  Silhouette =  0.44292946562535096
K = 8  Silhouette =  0.463006813341244
K = 9  Silhouette =  0.4691527216966012
```

Output1: Silhouette Scores



Output2: Sum of Squares Errors

Cluster	age	default	housing	loan	cons.price.idx	campaign	pdays	\
0	42.4181	0.1223	0.9760	0.0081	93.5580	2.5635	15.5495	
1	39.8467	0.0970	0.5907	1.0000	93.5675	2.6023	15.5174	
2	33.1606	0.0622	0.5390	0.0044	93.5180	2.5062	15.5556	
3	42.2233	0.1312	0.0000	0.0000	93.6471	2.6052	15.6272	
4	44.9378	0.0950	0.5260	0.0048	93.6037	2.5825	15.7991	

Cluster	previous	subscribed	divorced	married	single	unknown	count
0	0.1656	0.0081	0.0000	0.9978	0.0000	0.0022	11346
1	0.3705	1.0000	0.2080	0.6039	0.1862	0.0019	6248
2	0.2136	0.0044	0.0000	0.0000	1.0000	0.0000	9780
3	0.1447	1.0000	0.0000	0.9957	0.0000	0.0043	9876
4	0.1681	0.0048	0.9997	0.0000	0.0000	0.0003	3938

Output3: Cluster Centroids Table

After examining all the cluster results, you selected the 5-clusters solution (output3) for the Telemarketing team. Is this decision a correct one? State your reasons for agreeing and/or disagreeing.

(3 Marks)

Question 2 *(Total: 20 Marks)*

Autonomous vehicles are reliant on many various sensors such as camera, LiDAR (Light Detection and Ranging), RADAR, Ultrasonic sensors, and IMU (inertial measurement unit) sensors. Although there are currently many variations in design for autonomous vehicles (e.g. Waymo that is reliant on LiDAR, Tesla Autopilot that does not utilize LiDAR at all), it is evident that almost all designs for autonomous vehicles have included the camera sensor. The rationale for prioritising the usage of cameras over other distance sensors like LiDAR and RADAR is because of its superior ability to help autonomous vehicles to identify objects on the roads (i.e. to tell if the object is a pedestrian, car, bus, traffic light or traffic sign).

Traffic sign identification is one of the major tasks that an autonomous vehicle must perform, thus being a crucial component of autonomous vehicles. Without traffic sign identification capabilities, the autonomous vehicle is unable to adhere to traffic rules on the road (e.g. keeping within the desired speed limits, slowing down when nearing pedestrian zones, stopping at a junction) and could possibly cause accidents.

Identifying traffic signs can be categorized into two main stages: detection and classification. As part of the perception system of the autonomous vehicles, the camera sensor (together with machine vision algorithms) plays an important role in automatically detecting and classifying traffic signs (refer to examples of traffic signs in the below figure). The classification results will determine what action the autonomous vehicle should take next. For instance, if the classification result turns out to be the "STOP" sign, the vehicle will react accordingly by stopping behind the line at the "STOP" sign. This is a close representative of how we humans will react while driving.

A deep neural network method like Convolutional Neural Network (CNN) is often integrated in the machine vision algorithms for such purposes, as it has illustrated great success in performing the classification tasks; it is consistently proven that CNN has surpassed human performance even on challenging datasets.

You are an AI engineer and you are approached by an autonomous vehicle-focused company to mainly deal with the image classification problems that they are facing. They have also provided you with many images (that are extracted from the video frames taken only in Singapore) for you to train your deep learning model.

a. Your colleague has developed the first draft of a simple deep learning model (summarized in the table below) to classify the following traffic signs: "STOP", "Pedestrian Zone", "No U Turn", "50 Speed Limit", "90 Speed limit", "No Sign". The size of each input image is 256 x 256 pixels.

	Layer Type	Activation	Kernel Size	Output Shape
0	Input	-	-	(256, 256, 3)
1	Conv2D	ReLU	(12,12)	(245, 245, 16)
2	MaxPooling2D	-	(2,2)	(122, 122, 16)
3	Conv2D	ReLU	(6,6)	(117, 117, 32)
4	MaxPooling2D	-	(4,4)	(29, 29, 32)
5	Flatten	-	-	26912
6	Dense	ReLU	-	128
7	Dense	Softmax	-	6

It is important that the total parameters in the model are kept between 92,000 and 102,000 in order to achieve optimal results. Also note that the company is intending to expand their operations to various countries overseas; they highlighted 2 important points to you pertaining to this matter: (i) they do not have any more dataset to provide you, and (ii) the traffic signs in these countries differ in terms of colour combinations (the size of the signs remains the same).

Your colleague tested his model and realized that its performance is not up to expectations and that sometimes the net cannot converge. Based on the model he has shown you, identify three main issues with his model and discuss possible solutions to resolve them. Suggest an improved model by representing it in a new table with the number of parameter details (Note that you are required to construct the table from scratch in the answer sheet).

(8 marks)

b. Assuming that your developed model fails and/or encountered errors when fed with images like those below, suggest suitable solution(s) that can resolve this issue. You have to consider the limitations faced by the company as stated earlier in Question 2a.

(2 marks)

c. Your colleague Tim has developed another CNN model, which instead focuses on identifying the categories of the traffic signs (i.e. "regulatory", "warning" and "guide" which also correspond to the shape of the traffic signs; e.g. circle, triangle) based on 3D LiDAR scan data in point cloud format. Assuming that now your model is required to classify 100 various traffic signs instead of 6, discuss if his model can help you improve your model in any way. If yes, propose how you would combine these two models into a single model for training, validation and testing purposes.

(2 marks)

d. The vehicles driving profiles are composed of a continuous stream of speed parameters. They provide complementary information for the vehicle camera recognition system for optimizing urban transportation management or vehicle battery management. Figure 1 illustrates speed-time signal sequences of three vehicle driving patterns: *Fast, Medium, Slow*. You are asked to design a pattern recognition approach to predict the vehicle driving pattern on a combined speed-time signal sequences. You have considered various time-domain and frequency-domain feature extractions, and you need to apply fusion methods to combine these two kinds of features. Given two frequency-domain feature extraction proposals A and B described in Table 1, evaluate whether 'Early Fusion' or 'Late Fusion' is suitable for this fusion task.

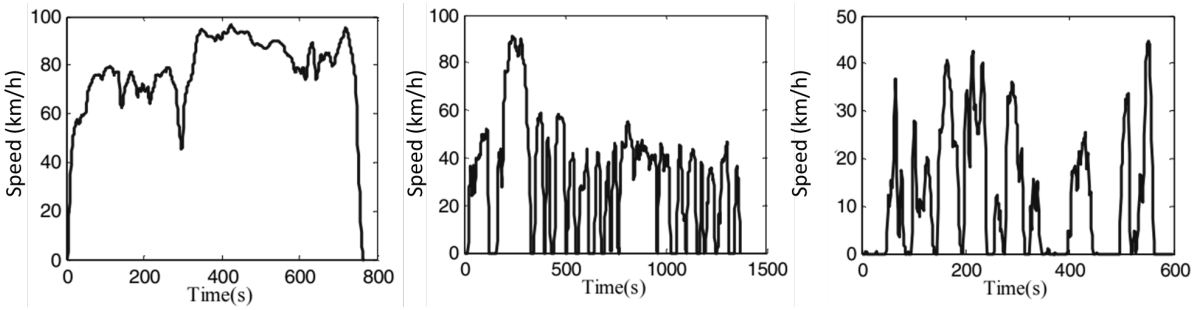


Figure 1. Three speed-time signal sequences of various vehicle driving patterns: (1) Fast (highway driving conditions, left figure); (2) Medium (city driving conditions, middle figure); (3) Slow (low speed stop-and-go traffic conditions, right figure).

Table 1: Various feature extraction proposals.

Input speed-time sequence: Two-minute sequence with sampling rate 1Hz, that is 120 data points. Time-domain feature extraction proposal: Split the input speed-time sequence to obtain non overlapping five segments, calculate three statistical measures (i.e., mean, variance, standard derivation) for each segment.	
Frequency-domain feature extraction proposal A	Frequency-domain feature extraction proposal B
Apply four-level wavelet transform on the speed-time sequence to obtain wavelet coefficients with necessary paddings, which have same dimensions with the input speed-time sequence.	First apply four-level wavelet transform on the speed-time sequence to obtain wavelet coefficients. Then further extract two statistical features (i.e., mean and variance) for each wavelet subband (approximation or detail).

(4 marks)

e. The traffic condition decisions obtained from the traffic sign classification system (Q2. a) and driving profile recognition system (Q2. d) can be further integrated together for traffic experience classification. Considering a specific travel path from a residential neighborhood to a primary school shown in Figure 2, an additional traffic sign counting system is applied using the outputs from the traffic sign classification system to make decisions: *Large* (a large number of red traffic light or STOP signs detected on the path) or *Small*. On the other hand, the driving profile recognition system can provide decisions: *Fast, Medium,* or *Slow*. You are asked to develop a multi-modal sense making approach to perform a travel experience classification: *Comfortable* or *Need to be improved,* using the training samples from various vehicles provided in Table 2. Given the new test data [*Slow, Large*], apply your developed method to make decision (i.e., *Comfortable* or *Need to be improved*). Show your calculation steps to justify your answer.

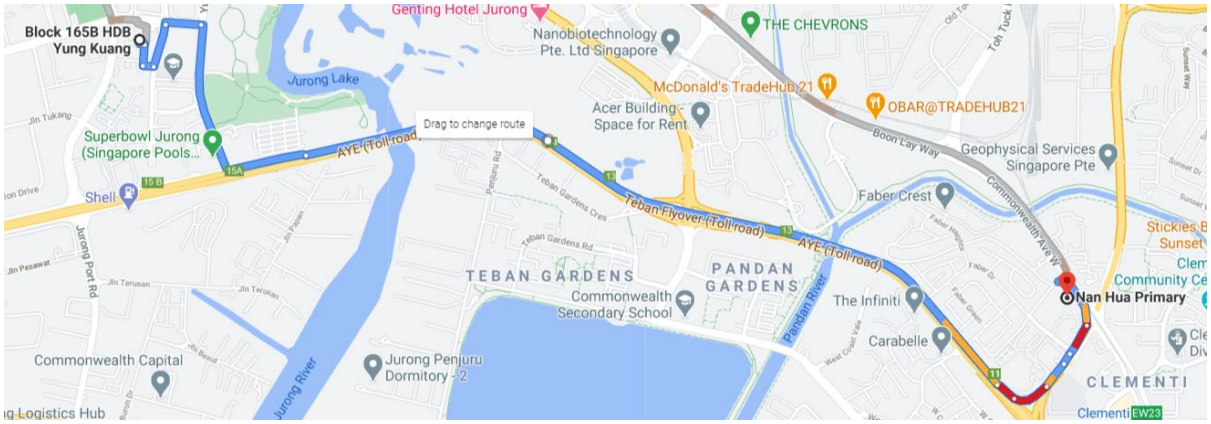


Figure 2. A travel path from a residential neighbourhood to a primary school.

Table 2. An example of training dataset collected for travel experience classification.

Training data	Decision based on driving profile recognition system	Decision based on traffic sign classification system	Travel experience classification
1	Fast	Small	Comfortable
2	Medium	Small	Comfortable
3	Slow	Small	Comfortable
4	Medium	Small	Need to be improved
5	Fast	Large	Need to be improved
6	Medium	Large	Need to be improved
7	Slow	Large	Need to be improved
8	Fast	Large	Comfortable

(4 marks)