UNIVERSITY
OF WOLLONGONG
AUSTRALIA

School of Computing and Information Technology

Student to complete:

Family name	
Other names	
Student number	

INFO911
Data Mining and Knowledge Discovery
South Western Sydney and Wollongong

Examination Paper
Autumn 2022

Exam duration	3 hours
Weighting	50%
Items permitted by examiner	OPEN BOOK - any reference materials permitted
Aids supplied	Nil

Directions to students

- 7 questions to be answered.
- This exam contributes to 50% of the total evaluation of the subject.
- Answer each question on a separate page clearly
- Convert the answers into one pdf file
- Submit the pdf file via Moodle

Question 1 **(6 marks in total)**

(A) In data mining, tasks can be categorised as predictive tasks or descriptive tasks. Describe their differences and name one algorithm for each of the two kinds of tasks. **(1 mark)**

(B) In data mining algorithms, a sample is often interpreted as a point in a multi-dimensional space. Explain how this interpretation is made and what the space is. **(1 mark)**

(C) What are outliers in data mining? And when shall outliers be removed or kept? **(1 mark)**

(D) Name two ways that can be used to deal with missing values and discuss their strengths and weaknesses. **(1 mark)**

(E) Dimensionality reduction and feature subset selection are two techniques of data pre-processing. Discuss their differences. **(1 mark)**

(F) List three important factors that drive the development of mining of big data. **(1 mark)**

Question 2 **(6 marks in total)**

(A) Clustering is a commonly used technique in data mining. Explain the purpose of conducting clustering and provide an example application of clustering. **(2 marks)**

(B) Suppose five clusters C1, C2, C3, C4 and C5 are formed during the process of hierarchical clustering. **(2 marks)**

a. Describe the corresponding proximity matrix in terms of its size and entries.

b. Assuming that clusters C3 and C4 are now merged into a new cluster, list the entries in the above proximity matrix that must be updated in order to continue performing hierarchical clustering.

(C) Suppose you are going to use self-organizing map (SOM) to map 1000 samples into a 2D grid of size of 10 x 10 neurons. Each sample has 20 attributes. Answer the following questions: **(2 marks)**

a. How many codebook vectors will be learned in this process?

b. What is the dimensionality of each codebook vector in this case?

Question 3 **(10 marks in total)**

(A) The performance of k-nearest neighbour (k-NN) classifier significantly depends on the value of k. Suppose you are given a set of M samples and the class label of each sample is also provided to you. Describe the procedure that you will use to select the best k value such that the k-NN classifier can expect to achieve the highest classification accuracy on a future test set. **(2 marks)**

(B) The key step to train a multi-layer perceptron (MLP) network is to adjust the weight of connections. Answer the following questions related to this step: **(2 marks)**

a. Describe a commonly used cost function in MLP for adjusting weights and explain its meaning.
b. Name the commonly used algorithm for MLP to adjust weights and list its key steps.

(C) Suppose you are going to train an MLP network with the five properties shown below. Calculate the total number of weights (i.e., weight parameters) that will be adjusted during the training process. Show and explain how you derive your answer. Note that you may not need to use all the properties provided. **(2 marks)**

a. The training set consists of N samples.
b. The dimensionality of each sample is D1.
c. The dimensionality of each target value is D2.
d. The MLP is fully connected and it has two hidden layers with the number of hidden neurons of L1 and L2, respectively.
e. The MLP network will be trained for T iterations.

(D) A student designs a binary (i.e., positive class vs. negative class) classifier and evaluates it on a test set consisting of 10 samples. The result is recorded in the following table. **(4 marks)**

Sample ID	Predicted probability for positive class	True label of this sample
1	0.19	Negative class
2	0.75	Positive class
3	0.99	Positive class
4	0.07	Negative class
5	0.52	Positive class
6	0.11	Negative class
7	0.06	Negative class
8	0.61	Positive class
9	0.67	Negative class
10	0.34	Positive class

The first column of this table is the sample identity number; the second column records the prediction made by the classifier, which is the probability that this sample belongs to the positive class. The last column is the true class label of the corresponding sample.

a. Construct a Receiver Operating Characteristic (ROC) curve based on the result in the table. Provide the key steps of this process;
b. Compute the Area Under the ROC curve (AUC).

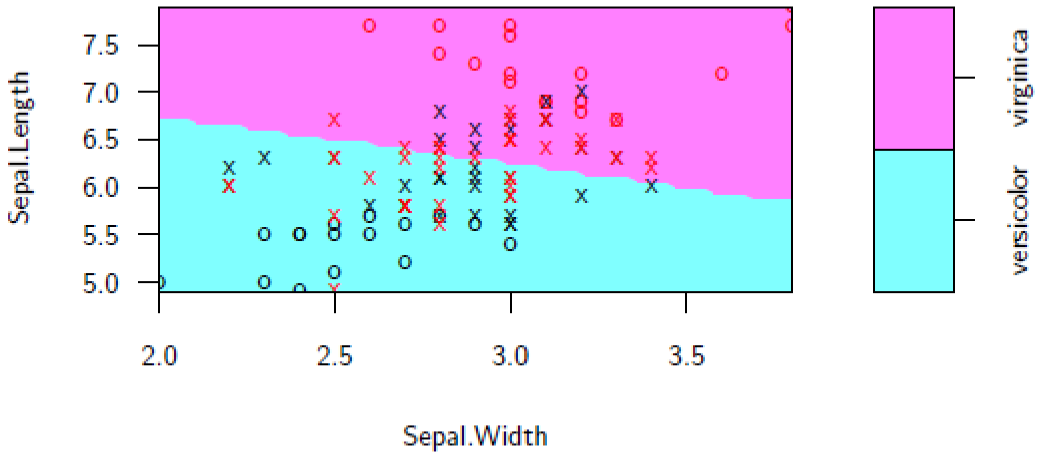
Question 4 **(3 marks in total)**

(A) Describe the main steps of the Apriori algorithm for mining association rules. Explain how the algorithm generates the candidate itemsets and how the algorithm prunes the candidate itemsets. **(1 mark)**

(B) Let L = {A,B,C,D} denote a frequent itemset. Use it to mathematically show that “confidence of rules generated from the same itemset has an anti-monotone property.” **(2 marks)**

Question 5 **(12 marks in total)**

(A) In the context of a classification task using support vector machine, why are some points represented by o and others by x? Explain clearly the important differences between them. **(2 marks)**



(B) Explain clearly the differences between soft margin classifier and maximal margin classifier. **(2 marks)**

(C) Explain the relationships between the slack variables, $\epsilon_i, i = 1, \ldots, n$, where n is the number of observations in the training dataset and the tuning parameter C in the soft margin classifier approach. **(2 marks)**

(D) Explain how the support vector machine approach can be used to classify a new observation y^* in the test set based on its three predictors x_1^*, x_2^*, and x_3^*. **(2 marks)**

(E) Explain the overfitting problem. **(2 marks)**

(F) State one obvious reason why linear regression is not appropriate for binary/categorical response variable. **(2 marks)**

Question 6 **(9 marks in total)**

(A) Compute the mean square error and mean absolute error for the following observations. Note that y is the observed values and $\hat{y}$ is the predicted values. **(2 marks)**

Observed values (y)	Predicted values ($\hat{y}$)
3220	3225.434
2745	2747.500
3320	3314.206

(B) Explain what the optimal Bayes classifier is and state clearly one important difference between optimal Bayes and naïve Bayes classifiers. **(3 marks)**

(C) This question is related to a classification task using a decision tree. Suppose that in the root node, you have 10 observations belong to class 1, 10 observations belong to class 2, and 10 observations belong to class 3. Then, following the first split, you have:

- Left Node: 4 observations with class 1, 3 observations with class 2, and 4 observations with class 3
- Right Node: 6 observations with class 1 and 7 observations with class 2, and 6 observations with class 3

Compute the entropy gain or loss. **(4 marks)**

Question 7 **(4 marks in total)**

There is a belief that sleep is important because of its impact on wages through the labour-market productivity. The dataset contains 706 individuals and is a subset of the data used by Biddle and Hamermesh (1990), "Sleep and the Allocation of Time," *Journal of Political Economy*, 98, 1, pp 922-943. The list of variables is:

sleep	minutes sleep at night per week
totwrk	minutes worked per week
age	age in years
male	=1 if male
educ	years of schooling
kid	=1 if there present children under 3 years of age
inlf	=1 if in labour force

Consider the following model that relates number of minutes sleeping on time spent working and other factors that may be affecting sleep.

$$sleep_i = \beta_0 + \beta_1 totwrk_i + \beta_2 age_i + \beta_3 male_i \;\; + \beta_4 educ_i + \beta_5 kid_i + \beta_6 inlf_i + e_i$$

When running the commands lm and summary in R you obtain the following results

```
Call:
lm(formula = sleep ~ (totwrk + age + factor(male) + educ + factor(kid) +
           factor(inlf)), data = sleep.data)

Residuals:
    Min      1Q   Median      3Q     Max
-2345.13 -237.92    5.47  265.08 1344.25

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3636.1751  119.0833  30.535  <2e-16 ***
totwrk         -0.1659    0.0181  -9.164  <2e-16 ***
age             2.0211    1.5246   1.326  0.1854
factor(male)1  87.9292   34.8295   2.525  0.0118 *
educ          -11.7273    5.8836  -1.993  0.0466 *
factor(kid)1    4.4487   50.1296   0.089  0.9293
factor(inlf)1    4.5419    36.9562   0.123   0.9022
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 418.3 on 699 degrees of freedom
Multiple R-squared:  0.1216,	Adjusted R-squared:  0.1141
F-statistic: 16.13 on 6 and 699 DF,  p-value: < 2.2e-16
```

(A) Is there evidence that there are some differences in the minutes sleep at night per week between men and women? Justify your results using significance levels of 0.05. **(1 mark)**

(B) Based on the significance of the coefficients at 5% significance level, are there any variables that you might exclude from the equation? Why? **(1 mark)**

(C) If a person works four more hours per week, by how many minutes is sleep predicted to fall? Show your work. **(2 marks)**

--End of this document--