LNMIIT
The LNM Institute of Information Technology

The LNM Institute of Information Technology

Department of Computer Science and Engineering

Introduction to Data Science (CSE 327)

End Term

Ujjawal Kumar
17UCS171.

Max. Time: 180 minutes **Date:** 02/12/2019 **Max. Marks:** 80

Instructions: The question paper consists of two parts (PART A and PART B) and each part must be answered on a separate answer sheet. Each question should be answered in a new page. Answer the questions in the same order as it appears in the question paper.

PART A (20 Marks)

Q1. (2 Marks) Consider the statement: "The pen is mightier than the sword". Find if there are any relations present in the statement like Homonymy, Polysemy or Metonymy? Explain with your reasoning in one sentence.

Q2. (4 Marks) Find the relationship between following pairs (i.e., whether they are Hyponymy, Hypernymy, Meronymy, Antonymy or Synonymy). Give one-line reasoning for your answers.

a. Gear and Car
b. Automobile and Car
c. Car and Vehicle
d. Finger and Hand

Q3. (6 Marks) For obese patients the Blood glucose levels have a mean of 100 with a standard deviation of 15. A medical researcher has an intuition that a diet high in raw cornstarch will have a positive or negative effect on blood glucose levels. A sample of 35 patients who have tried the raw cornstarch diet have a mean glucose level of 140. Test the hypothesis that the raw cornstarch had an effect (Note: You don't need z-table for this!).

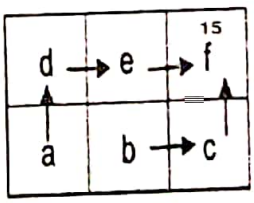
Q4. (4+4 = 8 Marks) Given are the input policy for the movement in the grid and six episodes (see the Picture).

a. Find the T and R function from the given episodes and
b. Find the utility values of the states by Model-free approach (Direct evaluation).

b, East, e, -1
e, East, b, -1
b, East, c, -1
c, North, f, -1
f, Exit, 15

a, North, d, -1
d, East, e, -1
e, East, f, -1
f, Exit, 15

a, North, d, -1
d, East, e, -1
e, East, f, -1
f, Exit, 15



a, North, d, -1
d, East, e, -1
e, East, b -1
b, East, c, -1
c, North, f, -1
f, Exit, 15

e, East, b, -1
b, East, c, -1
c, North, f, -1
f, Exit, 15

d, East, e, -1
e, East, d, -1
d, East, e, -1
e, East, f, -1
f, Exit, 15

PART B (60 Marks)

Instructions: Show your calculations whenever required. Just writing the answer will not be considered for award of marks.

Q1. (10 Marks) Apply agglomerative clustering to find two clusters from the following dataset (which has eight records) using (i) Single link (ii) Complete link and (iii) Average link inter-cluster proximity measures. For each of the above mentioned inter-cluster proximity measure you have to find clusters separately. Use Euclidean distance as the distance measure. Show all the steps of your computation.

Record	R1	R2	R3	R4	R5	R6	R7	R8
Value	5	9	25	30	49	10	48	44

Q2. (10 Marks) Suppose a dataset consists of 10 points (A, B, C, D, E, F, G, H, I, J) for which the pairwise distances are given in Table 1. Let the ***Eps*** and ***MinPts*** parameters be 8 and 3 (excluding the point itself). Apply the following version of DBSCAN algorithm on this dataset and show the output for each step of this algorithm.

Table 1: Distance between points

	A	B	C	D	E	F	G	H	I	J
A	0	5	9	2	11	6	12	18	9	25
B	5	0	22	14	9	2	10	25	26	9
C	9	22	0	7	5	14	15	24	12	14
D	2	14	7	0	9	4	10	15	11	11
E	11	9	5	9	0	24	13	3	7	23
F	6	2	14	4	24	0	10	26	13	18
G	12	10	15	10	13	10	0	9	18	14
H	18	25	24	15	3	26	9	0	10	10
I	9	26	12	11	7	13	18	10	0	12
J	25	9	14	11	23	18	14	10	12	0

DBSCAN algorithm:

Step 1. Label all points as core, border, or noise points.
Step 2. Eliminate noise points.
Step 3. Put an edge between all core points that are within Eps of each other.
Step 4. Make each group of connected core points into a separate cluster.
Step 5. Assign each border point to the closest of its associated core points.

Q3. (5+5 = 10 Marks) Answer the following questions.

a. We are interested to predict spam emails and for this we have developed a model M. Out of a test set consisting of 800 normal and 200 spam emails, M is able to detect 700 normal and 150 spam emails correctly and the rest incorrectly. Draw the confusion matrix and find out the false positive rate and F-measure of M.

b. Show that accuracy is a function of sensitivity and specificity.

Q4. (2+4+6+1+2 = 15 Marks) Consider the training examples shown in Table 2 for a binary classification problem.

Table 2: Data set

a_1	a_2	a_3	Class
T	T	1.0	P
T	T	6.0	P
T	F	5.0	N
F	F	4.0	P
F	T	7.0	N
F	T	3.0	N
F	F	8.0	N
T	F	7.0	P
F	T	5.0	N

a. What is the entropy of this collection of training examples with respect to the positive class (P)?

b. What are the information gains of a_1 and a_2 relative to these training examples?

c. For a_3, which is a continuous attribute, compute the information gain for every possible split and state the best split point for a_3.

d. What is the best split (among a_1, a_2, and a_3) according to the information gain?

e. What is the best split (between a1 and a2) according to the Gini index?

Q5. (6+1+8 = 15 Marks) The following table shows the midterm marks, endterm marks and final grades obtained by students in a course.

Midterm	72	50	81	74	94	86	59	83	65	33	88	81
Endterm	84	63	77	78	90	75	49	79	77	52	74	90
Grade	A	B	A	A	A	A	B	A	B	B	A	A

a. Use the method of least squares to find (derive) an equation for the prediction of a student's endterm mark based on the student's midterm mark in the course.

b. Predict the endterm mark of a student who received 86 in the midterm exam using the equation obtained in part a.

c. Consider Naïve Bayes classification method to predict grade of a student from his/her midterm and endterm marks. Use this classification method to predict grade of a student who has scored 70 and 80 in midterm and endterm exams respectively.