INDIAN INSTITUTE OF TECHNOLOGY MADRAS

B

Roll No. B E 1 6 B 0 3 6

Name : L. Srinath Muralidharan

Total No. of Pages □

Quiz I ☑ Quiz II/ Mid-Sem □ End-Semester □ Make-up □ Date : 20/2/19

Semester & Degree : Sixth, Dual Course No. BT3041 Part : □

Question No.	1	2	3	4	5	6	7	8	9	10
Marks									20	

11	12	13	14	15	16	17	18	19	20	Total
										12.78/20

Answer on both sides of the paper including the space below

1.

	Buddhism	Buddhist/	Buddhists	enlightenment	Nirvana	Asia	Monasticism	Dharma	Sangha	Paramitras
Passage 1	1	2	0	2	1	0	0	0	0	0
Passage 2	2	1	3	0	1	1	1	1	1	1

$$\text{cosine similarity} = \frac{\bar{a}\cdot\bar{b}}{\|\bar{a}\|\,\|\bar{b}\|}$$

Assumption: "Buddhists" and "Buddhist" are ~~count~~ considered the same.

$$= \frac{\langle \overset{1,2,0,2}{\quad}, 1, 0, 0, 0, 0, 0\rangle \cdot \langle 2, \overset{1,3}{\quad}, 0, 1, 1, 1, 1, 1, 1\rangle}{\|a\|\ \ \|b\|}$$

½

Also case insensitive.
So, enlightenment is same as Englightenment.

$$= \frac{5}{\sqrt{10}\ \sqrt{20}}$$

11

1	2 = 2	
2	1 = 2	
0	3 = 0	
2	0 = 0	
1	1 = 1	
2	1	4
8	4	1
1	4	9
	1	6
	1	
	4	
	4	
	1	

$$\cos\theta = \boxed{0.35}$$

0 being dissimilar, 1 being most similar, 0.35 indicates a minimal amount of similarity b/w the two passages.

3. a) Start by adding a single cluster having all data points to the 'list of clusters'.

b) Bisect the clusters using simple basic k-means algorithm, again and again. (Iteration step). Bisect those clusters first which are 'loose' and have high SSE

c) Repeat step (b), until you obtain 2^{n+1} clusters, where $2^n < K < 2^{n+1}$.

d) Agglomerate or merge clusters that are close and have low SSE. This will reduce no. of clusters by one each time.

e) Repeat step (d) until you have 'k' clusters.

4

4. No. of matches (1, 1) = 2

No. of mismatches (1, 0) or (0, 1) = 2

Jaccard distance $= \frac{2}{4} = 0.5$

Jaccard similarity $= \frac{2}{4} = 0.5$

4

5. d) Cone trees

6. B) scale based clustering.

7. a) Asymmetric binary data

8. B) It has an associated objective function.

4

9. A) k-means clustering

2. Since grades are ordinal variables, they can be ranked based on their order.

~~The following ranks are given for the corresponding~~

Let us allocate the following ranks for corresponding grades :-

S	3
A	2
B	1
C	0

$(S > A > B > C)$

	Cell bio	Mol bio	Genetics	Data analysis	Bio info
Avyay	1	0	2	1	1
Pratima	2	3	1	0	2

Manhattan distance is (Norm when $p = 1$) :-

3

$$|1-2| + |0-3| + |2-1| + |1-0| + |1-2|$$

$$= 1 + 3 + 1 + 1 + 1$$

$$= \boxed{7}$$

1.

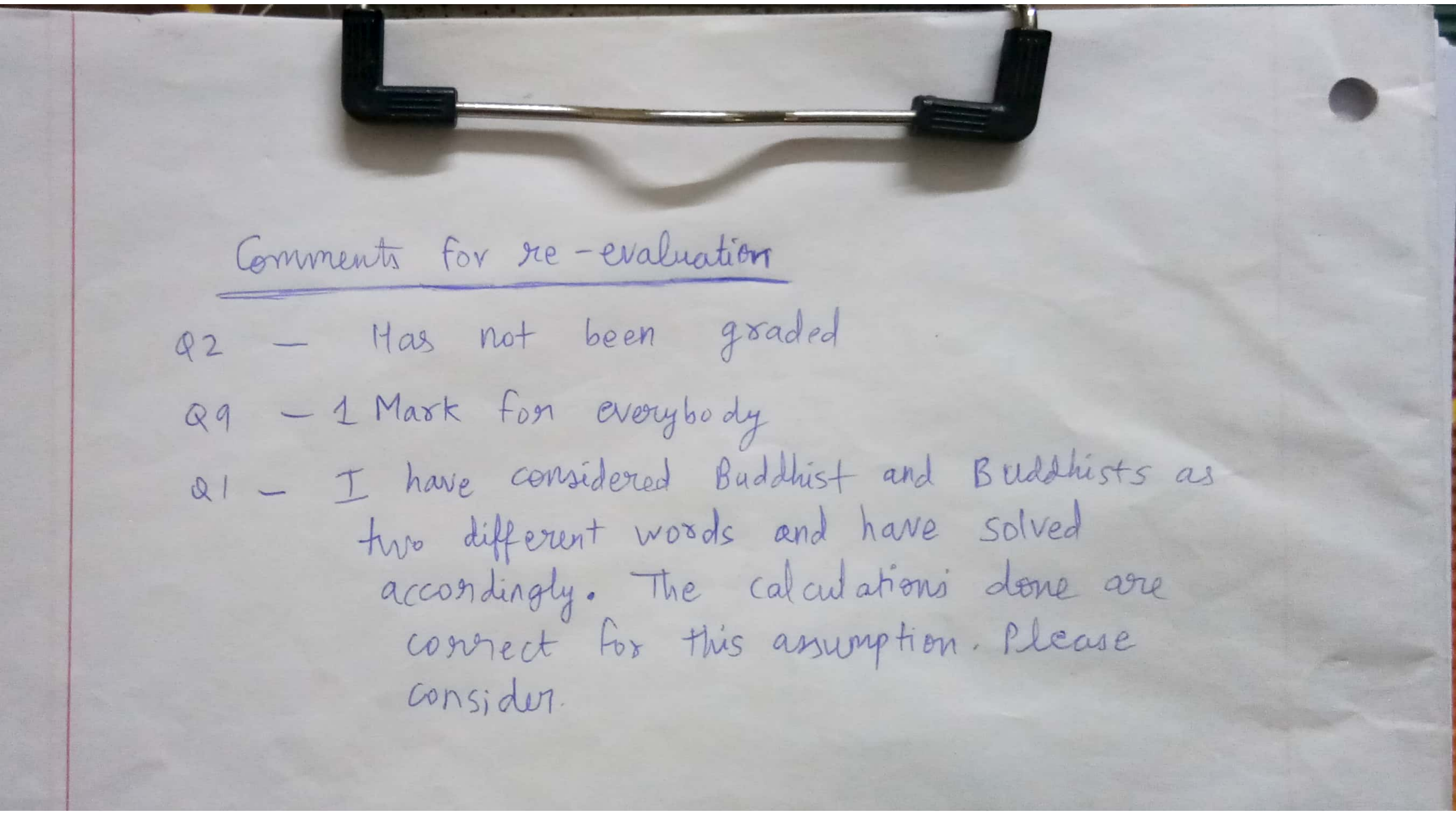
	Buddhism	Buddhists	Buddhist	Enlightenment	Nirvana	Asia	Monasticism	Dharma	Sangha	Paramitas
P1	1	2	0	2	1	0	0	0	0	0
P2	2	1	3	0	1	1	1	1	1	1

$$\text{cosine similarity} = \frac{\bar{a} \cdot \bar{b}}{||\bar{a}|| \; ||\bar{b}||}$$

$$= \frac{\overset{\bar{a}}{\langle 1,2,0,2,1,0,0,0,0,0 \rangle} \cdot \overset{\bar{b}}{\langle 2,1,3,0,1,1,1,1,1,1 \rangle}}{||\bar{a}|| \quad ||\bar{b}||}$$

$$\cos\theta = \frac{5}{\sqrt{10}\ \sqrt{20}} = \boxed{0.35}$$

This value indicates a minimal amount of similarity b/w P1 & P2.



Comments for re-evaluation

Q2 — Has not been graded

Q9 — 1 Mark for everybody

Q1 — I have considered Buddhist and Buddhists as two different words and have solved accordingly. The calculations done are correct for this assumption. Please consider.

L. Srinath Mu[illegible]ran
BE16B036

DEPARTMENT OF BIOTECHNOLOGY, IIT, MADRAS
CHENNAI – 36
BT 3041 Analysis and Interpretation of Biological Data

Class : Btech

Date : 27-3-2019

Time : 8:00-8:50 am QUIZ 2 Examination Marks: 20

Part A: Mark answers in the question paper itself and return it.

14 + 4 = 18/20

1. A dataset given as S = {(0,0), (0,1), (1,0), (2,1), (3,0), (3,-1), (4,0)} is clustered using DBSCAN. Let eps = 1.1, minPts = 2. The following 4 questions are based on the above problem. (4 marks)

1.1 The CORE points in the above data set are:

[✓ A) (0,0), (3,0)]

B) (2,1), (4,0)

C) (0,0), (0,1), (3,0), (3,-1)

D) (0,1), (2,1)

1.2. The BORDER points in the above data set are:

A) only (2,1)

B) (0,0), (0,1), (3,0), (3,-1)

[✓ C) (1,0), (0,1), (4,0), (3,-1)]

D) Only (4,0)

1.3 The NOISE points in the above data set are:

[✓ A) Only (2,1)]

B) (0,1), (2,1), (4,0)

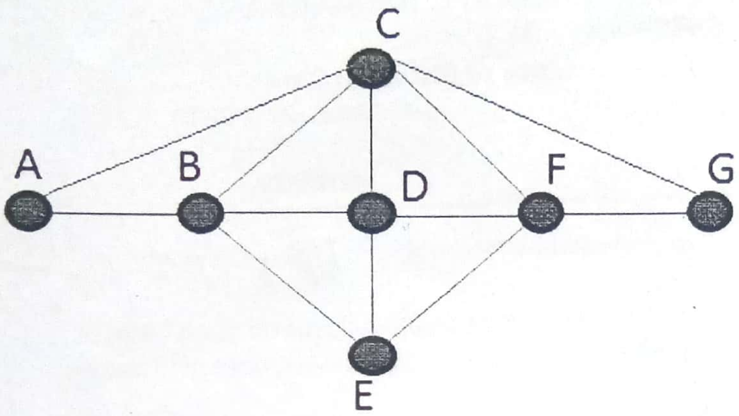
C) (0,1), (2,1), (3,-1), (4,0)

D) (0,1), (4,0)

1.4. The number of clusters that are picked up by DBSCAN in the above data set:

A) 1, [✓ B) 2], C) 3, D) 4

2. A graph representation of a dataset S consisting of 7 points (A to G) is shown below. A pair of points are connected by a link only if their similarity exceeds a certain threshold.

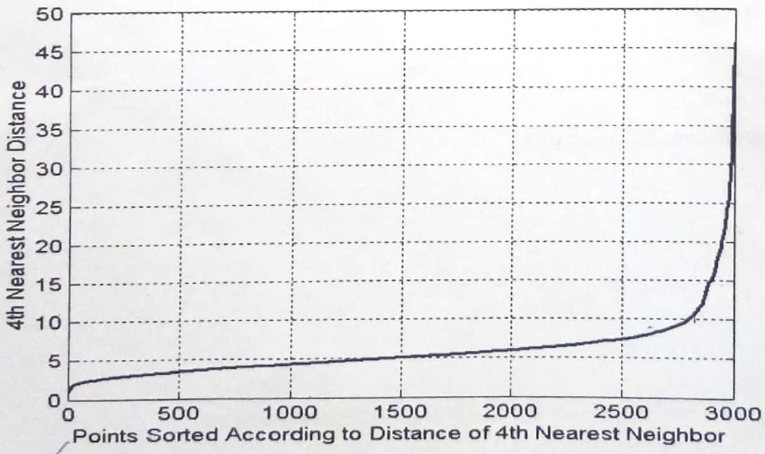


Construct the Shared Nearest Neighbor (SNN) graph of the above graph. The following 2 questions are based on the SNN graph that you constructed. (2 marks)

2.1 What is the strength of the link between nodes B and F in the SNN graph?
A) 1, B) 2, [✓ C)] 3, D) 4

2.2 What is the strength of the link between nodes D and G in the SNN graph?
B) 1, [✓ B)] 2, C) 3, D) 4

3. The plot below shows the 4th nearest neighbor distance of a data set in sorted order. If you wish to cluster the data using DBSCAN, what is the best value of eps given that MinPts = 4?



A) 5, [✓ B)] 10, C) 20, D) 42

4. Which of the following hierarchical clustering methods has an associated objective function?
A) Single linkage (MIN)
B) Complete linkage (MAX)
C) Ward's method
[✓ D)] Hierarchical clustering methods never have an objective function

3

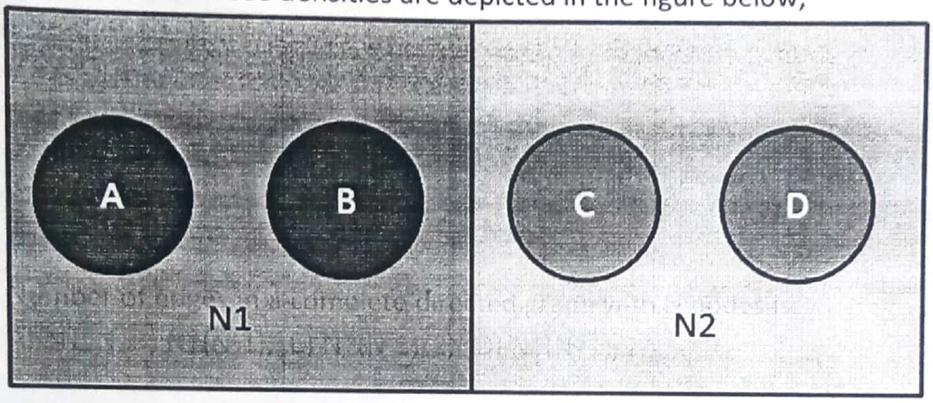
$^{N}C_2$

$\frac{N!}{(N-2)! \times 2} = \frac{N(N-1)}{2} + \frac{N(N-1)}{2}$

5. Number of edges in a complete directed graph with N nodes is:
   A) N*N, [✓ B) N*(N-1)], C) N*(N-1)/2, D) N*(N+1)/2

6. Which of the following is a density based clustering algorithm?
   A) Scale based clustering, [✓ B) Expectation Maximization using Gaussian Mixture Model], C) Bisecting K-means, D) none of the above

7. For a data set whose densities are depicted in the figure below,



Assume noise N1 has same density as clusters C and D. If Eps threshold is low enough to find clusters C and D, then:

A) The noise backgrounds N1 and N2 will be treated as a single cluster
B) A and B will be treated as separate clusters
C) A, B, C and D will be treated as separate clusters
[✓ D) A, B and N1 will be treated as a single cluster]

8. In fuzzy clustering, if w_{ij} denotes the weight that j'th data point belongs to i'th cluster, and if N is the number of data points, which of the following represents the constraints on w?

A) $\sum_i w_{ij} = 1$ and $\sum_j w_{ij} = 1$

B) $\sum_i w_{ij} = 1$ and $\sum_j w_{ij} = N$

[✓ C) $\sum_i w_{ij} = 1$ and $0 < \sum_j w_{ij} < N$]

D) $\sum_j w_{ij} = 1$ and $0 < \sum_i w_{ij} < N$

9. Which of the following are demerits of a multilayer perceptron? (multiple answers possible) (2 marks)

A) Parallelizable training algorithm

B) Slow training

2 C) Local minima

D) General approach to a wide range of problems

10. Which of the following statements are true for the solutions discovered by a perceptron (with step function nonlinearity)? (multiple answers possible) (2 marks)

A) Unique solution

B) Solution exists only if the training data is linearly separable

C) Non-unique solution

D) Solution exists even if the training data is linearly non-separable

NOTE : If its a MLP then solution exists even if data is linearly non-separable. But if we consider just an input + an output layer, solution exists only if data is linearly separable.

Based on my note, answer can be either (A),(B) or (A),(D).

(Part B) Answer in a separate answer sheet.

11. Design a multilayer perceptron (MLP) that can implement an XOR gate, by simple hand calculations (without using training). The MLP must have only 1 hidden layer and 2 neurons in that layer. The nonlinearity, g(x), for every neuron is the step function. The ~~XNOR~~ XOR gate is defined as:

X1	X2	D
0	0	0
0	1	1
1	0	1
1	1	0

(4 marks)

AND

0	0	0
0	1	0
1	0	0
1	1	1

OR

0	0	0
0	1	1
1	0	1
1	1	1

INDIAN INSTITUTE OF TECHNOLOGY MADRAS

A

Roll No. B E 1 6 B 0 3 6

Name : L. Srinath Muralidharan

Total No. of Pages

Quiz I □ Quiz II/ Mid-Sem ☑ End-Semester □ Make-up □ Date : 27/03/19

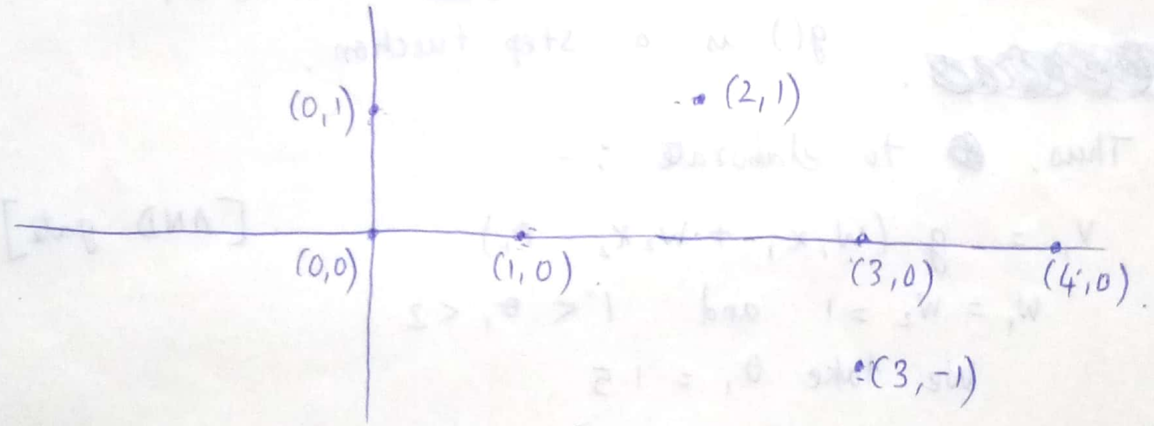
Semester & Degree : Sixth, Dual Course No. BT3041 Part :

Question No.	1	2	3	4	5	6	7	8	9	10
Marks										

11	12	13	14	15	16	17	18	19	20	Total

Answer on both sides of the paper including the space below

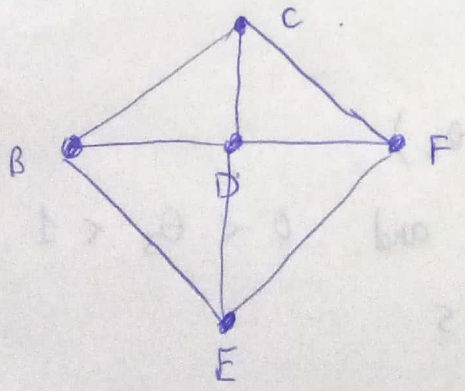
1.



2. Let the threshold T be ~~2~~ 2

T = 2.

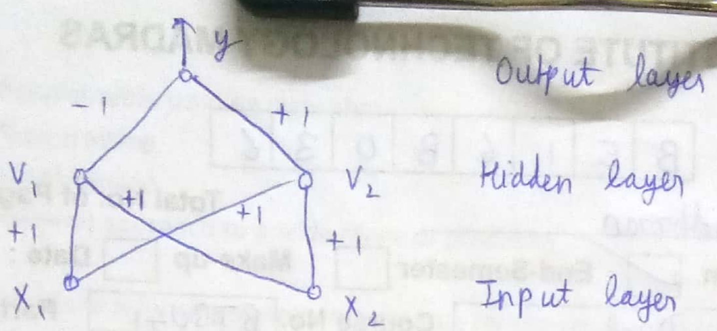
The SNN graph of above graph would then be :-



Pairs (A,B), (F,G), (A,C) and (C,G) share only one common neighbour. So their edges are broken.

11.

PART - B



$V_1 = g(x_1 + x_2 - 1.5)$ (AND gate) (Actually θ can be anything such that $1 < \theta < 2$)

$V_2 = g(x_1 + x_2 - 0.5)$ (OR gate) (θ should be such that $0 < \theta < 1$)

$y = g(V_2 - V_1 - \theta_0)$ where $0 < \theta_0 < 1$

g() is a step function.

4

Thus, to elaborate :-

$V_1 = g(W_1 x_1 + W_2 x_2 - \theta_1)$ [AND gate]

$W_1 = W_2 = 1$ and $1 < \theta_1 < 2$

We take $\theta_1 = 1.5$

$V_2 = g(W_1 x_1 + W_2 x_2 - \theta_2)$ [OR gate]

$W_1 = W_2 = 1$ and $0 < \theta_2 < 1$

We take $\theta_2 = 0.5$.

$y = g(W_1 V_1 + W_2 V_2 - \theta_3)$.

$W_1 = -1$, $W_2 = 1$ and $0 < \theta_3 < 1$.

We take $\theta_3 = 0.5$

Thus, $y = g(V_2 - V_1 - 0.5)$