M.S. Ramaiah Institute of Technology
(Autonomous Institute, Affiliated to VTU)
Department of Information Science & Engg.

gramme : B.E
rse: Information security
: I
x Marks: 30
rtions for Test: Unit-I and Unit-II(L1-L21)
structions to Candidates: Answer any 2 full questions.

Term: 30th August to 29th December 2018
Course Code: IS716
Sem: VII **Sec:** A,B,C
Time: 1Hr

#	Question	Marks	Course Outcome	Bloom's Level
1.	a) Discuss the security services defined by X.800 to ensure security of the systems.	10	CO1	Understand
	b) What are the properties of the Modulo operator. Demonstrate each of them with an example.	05	CO2	Under stand
2.	a) Encrypt the message "**Taj Mahal an ivory-white marble**" using the Hill cipher with the key **AGRA**: Show your calculations and the result. **OR** Explain the multiletter encryption using Playfair cipher. Encrypt the message "**Taj Mahal an ivory white marble**" using the Playfair cipher with the keyword **AGRA**.	10	CO1	Apply
	b) Discuss with examples the properties that the Modular Arithmetic exhibits.	05	CO2	Apply

P.T.O

Ramaiah Institute of Technology
(Autonomous Institute, Affiliated to VTU)
Department of Information Science & Engg.

Information Security(IS716)
Scheme and Solution

I#	Question	Marks	Course Outcome	Bloom's Level
1.	a) Discuss the security services defined by X.800 to ensure security of the systems.	10	CO1	Understand
	5 points – 2 Marks each • **Authentication** - assurance that the communicating entity is the one claimed • **Access Control** - prevention of the unauthorized use of a resource • **Data Confidentiality** –protection of data from unauthorized disclosure • **Data Integrity** - assurance that data received is as sent by an authorized entity • **Non-Repudiation** - protection against denial by one of the parties in a communication			
	b) What are the properties of the Modulo operator. Demonstrate each of them with an example.	05	CO2	Under stand
	Properties: - 3 Marks + Examples – 2 Marks The congruence relation satisfies all the conditions of an equivalence relation: • Reflexivity: $a \equiv a \pmod n$ • Symmetry: $a \equiv b \pmod n$ if and only if $b \equiv a \pmod n$ • Transitivity: If $a \equiv b \pmod n$ and $b \equiv c \pmod n$, then $a \equiv c \pmod n$ If $a_1 \equiv b_1 \pmod n$ and $a_2 \equiv b_2 \pmod n$, or if $a \equiv b \pmod n$, then: • $a + k \equiv b + k \pmod n$ for any integer k (compatibility with translation) • $k\,a \equiv k\,b \pmod n$ for any integer k (compatibility with scaling) • $a_1 + a_2 \equiv b_1 + b_2 \pmod n$ (compatibility with addition) • $a_1 - a_2 \equiv b_1 - b_2 \pmod n$ (compatibility with subtraction) • $a_1\,a_2 \equiv b_1\,b_2 \pmod n$ (compatibility with multiplication) • $a^k \equiv b^k \pmod n$ for any non-negative integer k (compatibility with exponentiation) • $p(a) \equiv p(b) \pmod n$, for any polynomial $p(x)$ with integer coefficients (compatibility with polynomial evaluation) If $a \equiv b \pmod n$, then it is false, in general, that $k^a \equiv k^b \pmod n$. However, one has: • If $c \equiv d \pmod{\varphi(n)}$, where φ is Euler's totient function, then $a^c \equiv a^d \pmod n$ provided a is coprime with n			
2.	a) Encrypt the message "**Taj Mahal an ivory white marble**" using the Hill cipher with the key **AGRA**: Show your calculations and the result. **OR** Explain the multiletter encryption using Playfair cipher. Encrypt the message "**Taj Mahal an ivory white marble**" using the Playfair cipher with the keyword **AGRA**.	10	CO1	Apply
	Hill Cipher:			

Solution:

Number the alphabet:

A=0, B=1,....., Z=25.

Keyword represented as a matrix: 01 Mark

A	G
R	A

0	6
17	0

The plaintext split into column vector: 02 Marks

T	J	A	A	A	I	O	Y	H	T	M	R	L
A	M	H	L	N	V	R	W	I	E	A	B	E

Plaintext represented as appropriate column vector:

19	9	0	0	0	8	14	24	7	19	12	17	11
0	12	7	11	13	21	17	22	8	4	0	1	4

Encrypted values:

0	20	16	14	0	22	24	2	22	24	0	6	24
11	23	0	0	0	6	4	18	15	11	22	3	24

Sample Calculation: 05 Marks

$$\begin{matrix} 0 & 6 \\ 17 & 0 \end{matrix} \begin{matrix} 19 \\ 0 \end{matrix} = \begin{matrix} 0 \\ 323 \end{matrix} = \begin{matrix} 0 \\ 11 \end{matrix}$$

Similarly other calculations to be shown:

Encrypted Text message: 02

ALUKQACAAAAWGYECSWPYLAWGDYY

Play Fair Cipher: - 6 Marks

Matrix:

A	G	R	B	C
D	E	F	H	I/J
K	L	M	N	O
P	Q	S	T	U
V	W	X	Y	Z

Encrypted Message:

PBFDBDGKBKDZWBXEYHUFLGBGNK

Encrypting and Decrypting 04 Marks

- plaintext is encrypted two letters at a time
  1. if a pair is a repeated letter, insert filler like 'X'
  2. if both letters fall in the same row, replace each with letter to right (wrapping back to start from end)
  3. if both letters fall in the same column, replace each with the letter below it (wrapping to top from bottom)

otherwise each letter is replaced by the letter in the same row and in the column of the other

letter of the pair

c) Discuss with examples the properties that the Modular Arithmetic exhibits.	05	CO2	Apply

Properties of Congruences – 03 Marks

Congruences have the following properties:

1. $a \equiv b \pmod{n}$ if $n|(a - b)$.
2. $a \equiv b \pmod{n}$ implies $b \equiv a \pmod{n}$.
3. $a \equiv b \pmod{n}$ and $b \equiv c \pmod{n}$ imply $a \equiv c \pmod{n}$.

To demonstrate the first point, if $n \mid (a - b)$, then $(a - b) = kn$ for some k.
So we can write $a = b + kn$. Therefore, $(a \bmod n)$ = (remainder when $b + kn$ is divided by n) = (remainder when b is divided by n) = $(b \bmod n)$.

Examples- 2 Marks

23 = = 8 (mod 5) because 23 - 8 = 15 = 5 * 3
-11 = = 5 (mod 8) because -11 - 5 = -16 = 8 * (-2)
81 == 0 (mod 27) because 81 - 0 = 81 = 27 * 3

a) List out the various Block Cipher Modes of Operation and explain any two of them in detail.	05	CO1	Understand

Listing – 1 Mark

- Electronic Code Book (ECB) – Electronic code book is the easiest block cipher mode of functioning. ...
- Cipher Block Chaining – Cipher block chaining or CBC is an advancement made on ECB since ECBcompromises some security requirements
- Cipher Feedback Mode (CFB)
- Output Feedback Mode
- Counter Mode

Explanation for any two of them: 2*2 Marks each

3.	b) With a neat diagram, discuss the stages involved in key generation of S-DES. Generate the two keys K1 and K2 using the same, given the input 0101110111. Consider the following permutations:	07	CO1	Apply

P10
3 ,5,2,7,4,10,1,9,8,6

P8
6,3,7,4,8,5,10,9

Writing Key generation Diagram **– 3 Marks**
Using the S-DES Key generation method, we get,
K1=00111011 **- 2 Marks**
K2=10111110 **- 2 Marks**

c) Discuss the properties of Modular Arithmetic for Integers in Z_n.	03	CO2	Understand

Discuss on residue classes – 02 Marks
Example – 01 Marks

Ramaiah Institute of Technology
(Autonomous Institute, Affliated to VTU)
Department of Information Science and Engineering
Programme: B.E in Information Science and Engineering

m	:	30th August – 29th December, 2018	**Course Code**	:	IS716
urse Name	:	Information Security	**Semester**	:	VII
E	:	Test - II	**Max. Marks**	:	30
te	:	20.11.2018	**Time Allotted**	:	9.30AM - 10.30AM

rtions for the Test: Lecture Numbers from 22 to 36 as per lesson plan
truction to Candidates: Use of mobile phones is strictly prohibited

Answer any two questions

Questions	Marks	Bloom's Level	CO
escribe Diffle-Hellman Key Exchange algorithm. Users A and B use the Diffie-lellman Key Exchange technique with parameters common prime q=71 and rimitive root α = 7. If users A and B have private keys X_A = 5 and X_B= 12 espectively, compute their public keys Y_A and Y_B. Also, calculate the shared ecret key K.	09	A	CO2
ompare and contrast Dumpster diving, Wardriving and Wardialing.	06	Az	CO3
Determine the result of encryption using RSA algorithm if the following arameters are used. =17, q=11, e=7, M=88. lso, perform decryption using the parameter 'd' which must be computed.	08	A	CO2
utline the features of ICMP (Ping) in detail.	07	U	CO3
lucidate OS Fingerprinting in detail.	08	U	CO3
lustrate and explain Digital Signature Algorithm with neat sketches.	07	U	CO2

– Understand; A – Apply; Az-Analyze

User B Key Generation: Private Key $X_B < q$
Public Key,

$$Y_B = \alpha^{X_B} \bmod q$$

Exchange of public keys between both users

Shared Key:
At A:

$$K = Y_B{}^{X_A} \bmod q$$

At B:

$$K = Y_A{}^{X_B} \bmod q$$

(5 Marks)

Public keys Y_A and Y_B

$$Y_A = \alpha^{X_A} \bmod q$$
$$Y_A = 7^5 \bmod 71 = 51$$

$$Y_B = \alpha^{X_B} \bmod q$$
$$Y_B = 7^{12} \bmod 71 = 4$$

(2 Marks)

Shared secret key

At A:

$$K = Y_B{}^{X_A} \bmod q$$
$$K = 4^5 \bmod 71 = 30$$

OR

At B:

$$K = Y_A{}^{X_B} \bmod q$$
$$K = 51^{12} \bmod 71 = 30$$

(2 Marks)

Compare and contrast Dumpster diving, Wardriving and Wardialing. **Answer:** Based on location, all the three attacks can be launched. *Dumpster Diving:* Only address of the victim is known. Garbage from the company may contain personal financial data, operation manuals, passwords, guides, etc., One can reduce this risk by shredding old CDs, wiping hard drives, etc., Electronic dumpster diving involves automatic archiving of web pages in Wayback Machine or leaking of information by disgruntled employees for free or premium. *Wardriving:* Act of finding and marking location and status of wireless networks. This risk can be ereduced by turning on encryption and physically protecting wireless access points. Also a form of wardialling. *Wardialling:* Act of automatically scanning telephone numbers using a modem, usually dialing every telephone number in a local area. Attacks are launched using modem that have weak or no authentication at all. (2 Marks for every attack)	06	Az	CO3
Determine the result of encryption using RSA algorithm if the following parameters are used. p=17, q=11, e=7, M=88. Also, perform decryption using the parameter 'd' which must be computed. **Answer:** Calculate n=pq=17*11=187 Calculate $\varphi(n)$ = (p-1)(q-1) = 16*10 =160 $d = e^{-1} \bmod \varphi(n) = 7^{-1} \bmod 160 = 23$ Extended Euclidean Algorithm to calculate d	08	A	CO2

Step (k)		q	p
0	160 = 22*7 +6	22	$p_0 = 0$
1	7=1*6+1	1	$p_1 = 1$
2	6=6*1+0	6	$p_2 = (p_0 - p_1 q_0) \bmod 160$ =(0-1*2) mod 160 = 138
3			$p_3 = (p_1 - p_2 q_1) \bmod 160$ =(1-138*1) mod 160 = 23

(2 Marks)

Encryption: $C = M^e \bmod n = 88^7 \bmod 187 = 11$ (3 Marks) Decryption: $M = C^d \bmod n = 11^{23} \bmod 187 = 88$ (3 Marks)			
Outline the features of ICMP (Ping) in detail. **Answer:** Internet Control Message Protocol (ICMP) • Designed to aid in network diagnostics and to send error messages (1 Mark) • Any network device using TCP/IP can send, receive and process ICMP messages • ICMP messages have no priority and do not flood the network • Sometimes, these messages are discarded if devices consider them as interruptions • These messages cannot be sent as response to other ICMP messages. Not sent in case of broadcast / multicast address. • If there is fragmentation of traffic, ICMP message is sent for first segment only (3 Marks) • Common type of ICMP message is ping which is designed to verify connectivity • Ping sweep to determine active hosts. • Drawback of ping: Can ping one system at a time and does not identify services in the system (3 Marks)	07	U	CO3
Elucidate OS Fingerprinting in detail. **Answer:** • OS in a system can be detected in active / passive manner • Passive tool monitors network traffic without interacting with the target • Active tool interacts with the target, sends triggers, analyses responses (2 Marks) • Passive Fingerprinting: IP addresses, active systems and open ports have been identified from the packets in the network • Commonly examined items – IP TTL value, TCP window size, IP DF option and IP TOS option • P0f (Linux based tool) (2 Marks) • Active Fingerprinting: Attacker injects packets voluntarily instead of waiting for random packets to analyze • Methods used: FIN probe, Bogus Flag Probe, Initial Sequence Number Sampling, IPID Sampling, TCP Initial Window, ACK value, ToS, TCP Options, Fragmentation Handling (2 Marks) • OS Fingerprinting Tools: Nmap, Xprobe2 (2 Marks)	08	U	CO3

Illustrate and explain Digital Signature Algorithm with neat sketches. **Answer:** Diagram for signing and verifying – 03 Marks Explanation – 04 Marks Global Public Key Components p, prime number and q, prime divisor of (p-1) $g = h^{\frac{(p-1)}{q}} \bmod p$ User's Private Key x, random / pseudo-random integer with 0<x<q User's Public Key $y = g^x \bmod p$ User's Per-Message Secret Number k, random / pseudo-random integer with 0<k<q Signing $r = (g^k \bmod p) \bmod q$ $s = [k^{-1} H(M) + xr] \bmod q$ Signature(r,s) Verifying $w = (s')^{-1} \bmod q$ $u_1 = [H(M')w] \bmod q$ $u_2 = [(r')w] \bmod q$ $v = [(g^{u1} y^{u2}) \bmod p] \bmod q$ TEST : v=r' M- Message to be signed H(M) - Hash of M using SHA-1 M', r', s' – Received version of M,r,s	07	U	CO2

M S RAMAIAH INSTITUTE OF TECHNOLOGY

Department of Information Science and Engineering

Term:	30/08/2018 to 29/12/2018	**Course Code:**	IS716
Course:	Information Security	**Semester:**	7
CIE:	Test-III	**Max Marks:**	30
Date:	21-12-2018	**Time:**	9.30AM - 10.30AM

rtions for Test: Lecture Nos. from 33 to 53 as per lesson plan

structions to Candidates: Answer any **Two** out of Three questions. **Mobile phones are banned**

	Questions	Marks	Bloom's Level	CO
	List System Assessment Tools and write a note on Nessus	8	R	CO4
.	Explain types of spread-spectrum technology.	4	U	CO5
.	Write a note on wireless LAN threats	3	R	CO5
.	Explain the design principles of firewall and list its limitations	7	U	CO4
.	Compare WiFi security protocols WEP and WPE	4	An	CO5
	Explain WPA authentication protocol defined in RFC 3758.	4	U	CO5
.	Illustrate Metasploit attack platform	4	U	CO4
.	Compare Viruses and worms	4	U	CO4
.	What are IDS? Explain different types of IDS	7	Ap	CO5

–Remembering, U – Understanding, Ap – Applying, An – Analyzing

1a. List System Assessment Tools (any 4 with simple explanation)

- GFI LANguard
- ISS Internet Scanner
- MBSA
- NetRecon
- Retina
- QualysGuard
- SARA
- SAINT
- VLAD
- X-Scan

write a note on Nessus

The client/server model
The basic components of Nessus:

- The Nessus client/server model
- The Nessus plug-ins
- The Nessus Knowledge Base

The basic steps in reviewing:

1. Inventory network devices.
2. Identify targets.
3. Create a plug-in policy.
4. Launch a scan.
5. Analyze the reports.
6. Remediate and repair.

1b. Explain types of spread-spectrum technology

spread spectrum technology – definition / purpose

Types of spread-spectrum technology

1. direct-sequence spread spectrum (DSSS),
2. frequency-hopping spread spectrum (FHSS), and
3. orthogonal division multiplexing (ODM).

1c. Write a note on wireless LAN threats

- Wardriving
- Warchalking
- War flying

2a. Explain the design principles of firewall and list its limitations

Firewall Design Principles, goals / aim and techniques

Limitations

). Compare WiFi security protocols WEP and WPE **(2x2=4)**

EP	WPA
ired Equivalent Privacy	WiFi Protected Access protocol
C4	AES
ey – 64/128 bits wit 24 bit IV	Key – 256 with 40 bit IV
	Secures the wireless systems quickly

:. Explain WPA authentication protocol defined in RFC 3758. **(4)**

ctensible Authentication Protocol (EAP) is a WPA authentication protocol defined in RFC 3758.EAP rides on top of e Ethernet protocol to facilitate authenticationbetween the client requesting to be authenticated and the server erformingthe authenticating. There is also EAP over LAN (EAPOL), which the IEEEapproved as a transmission ethod to move packets from the client to anauthentication serverThere are four basic types of EAPOL packets:
he **EAPOL packet** — This message type is simply a container for transporting EAP packets across a LAN.
he **EAPOL start** — This message is used by the client to inform the authenticator it wants to authenticate to the etwork.
he **EAPOL logoff** — The message informs the authenticator that the client is leaving the network.
he **EAPOL key**— This message type is used with 802.1x for key distribution.

emporal Key Integrity Protocol (TKIP) is used to address the known cipher attack vulnerability that WEP was ılnerable to.
KIP's role is to ensure each data packet is sent with its own unique encryption key. TKIP uses the RC4 algorithm.

a. Illustrate Metasploit attack platform. **(4)**

is a vulnerability assessment tools
ne basic approaches are
Selecting the exploit module to be executed
Choosing the configuration options for the exploit options
Selecting the payload and specifying the payload options to be entered
Launching the exploit and waiting for a response
etasploit has three basic ways that it can be controlled:
ne msfweb — A simple point-and-click interface
ne msfconsole — A console-based interface
ne msfcli —A command-line interface

b. Compare Viruses and worms **(2x2=4)**

linimum of 2 differences between virus and worms

:.What are IDS? Explain different types of IDS. **(1+2x3=7)**

efine IDS.
wo types of IDS are Host-based and Network-based IDS with examples.

M S Ramaiah Institute of Technology

Department of Information Science and Engineering

Term:	30/08/2018 to 29/12/2018	**Course Code:**	IS711
Course :	Data Mining	**Semester:**	7
CIE:	Test 1	**Max Marks:**	30
Date:	01/10/18	**Time:**	9:30 –10:30 AM

Portions for Test: Lecture Nos. from 1 to 16 as per lesson plan

Instructions to Candidates: Answer any **Two** out of Three questions.

Questions	Marks	Bloom's Level	CO
Discuss any three challenges that motivate the development of data mining.	6	U	1
Consider the following market basket transaction table. If the minimum support is 30%, obtain the frequent itemsets using the Apriori Algorithm and also find the strong rules if the minimum confidence is 80%.	9	Ap	2

TID	Item list
T1	a,b,d,e
T2	b,c,d
T3	a,b,d,e
T4	a,c,d,e
T5	b,c,d,e
T6	b,d,e
T7	c,d
T8	a,b,c
T9	a,d,e
T10	b,d

Discuss the issues on measurement and data collection aspect of data quality.	6	U	
I)Consider the following set of frequent 3 item sets: {1,2,3},{1,2,4},{1,2,5},{1,3,4},{1,3,5},{2,3,4},{2,3,5},{3,4,5} Assume that there are only five items in the data set. i) List all candidate 4-itemsets obtained by a candidate generation procedure using the $F_{k-1} \times F_1$ ii)List all candidate 4-itemsets obtained by the candidate generation procedure in Apriori. II) Describe the alternative methods for generating freguent itemsets using a. equivalence classes b. Breadth- first vs Depth-First Draw suitable diagrams.	9	Ap	
Discuss at least three ways of dealing with missing data listing and their advantages and disadvantages.	6	U	
Given below is the transaction table. i)Obtain the frequent itemsets using the Apriori Principle. Given that the minimum support is 2. ii) obtain the strong association rules, given that the confidence threshold is 70%.	9	Ap	

TID	Item list
T1	I1, I2, I5
T2	I2,I4
T3	I2,I3
T4	I1,I2,I4
T5	I1,I3
T6	I2,I3
T7	I1,I3
T8	I1,I2,I3,I5
T9	I1,I2,I3

Mobile phones are banned

M S Ramaiah Institute of Technology

Department of Information Science and Engineering

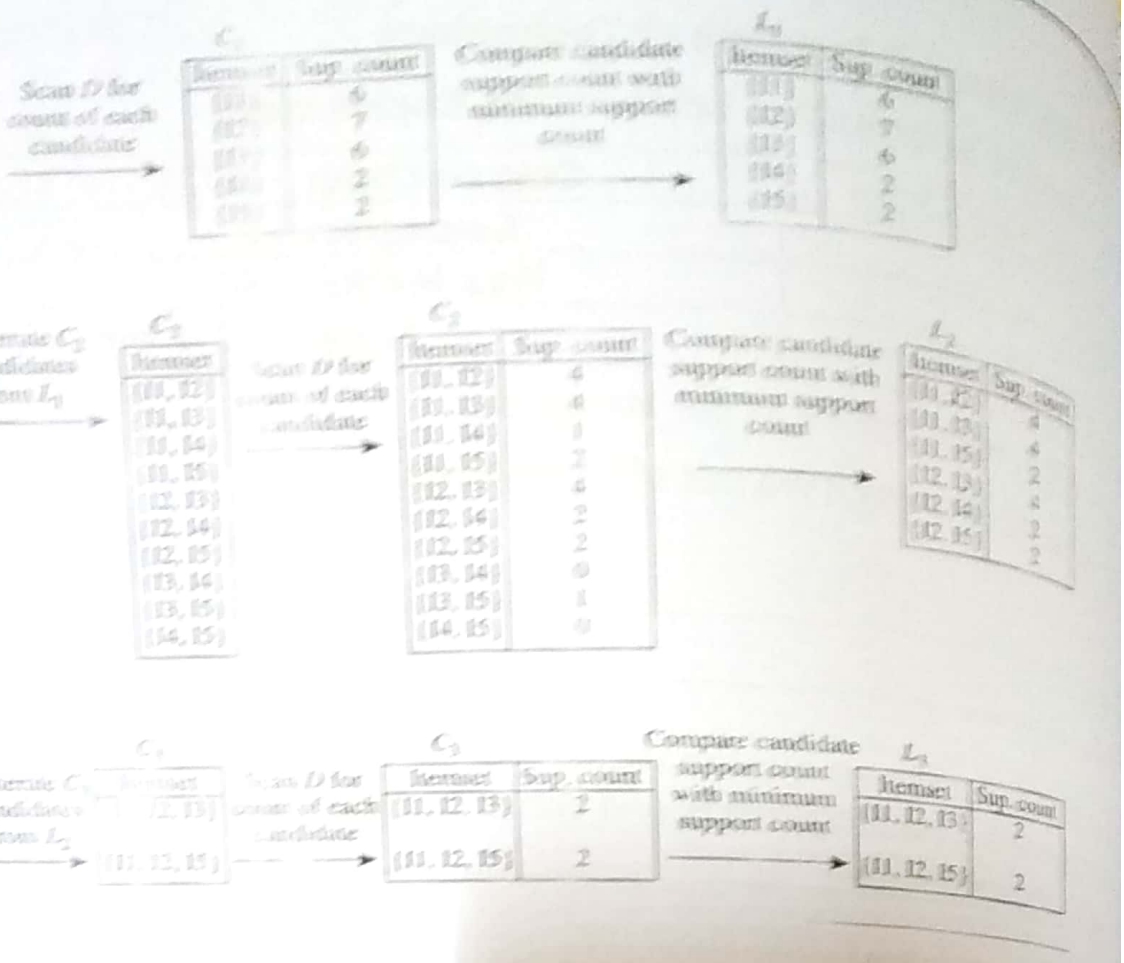
Term:	30/08/2018 to 29/12/2018	**Course Code:**	IS721
Course:	Data Mining	**Semester:**	7
CIE:	Test 1	**Max Marks:**	30
Date:	**03/10/18**	**Time:**	**12-1 pm**

ortions for Test: Lecture Nos. from 1 to 16 as per lesson plan

structions to Candidates: Answer any **Two** out of Three questions.

cheme of Solution

Questions	Marks
Challenges (Any 3*2m Each with an explanation) Scalability, High Dimensionality, Heterogeneous and complex data, Data Ownership and distribution, Nontraditional Analysis	6
Min support = 3 Generating frequent itemsets: **------(5m)** {a},{b},{c},{d},{e},{a,b}, {a,d},{a,e},{b,c},{b,d},{b,e},{c,d},{d,e}, {a,d,e},{b,d,e} **Rules -----(4m)** **{a,d}->{e} confidence =4/4=100%** **{a,e}->{d} confidence =4/4=100%** {d,e}->{a} confidence =4/6=67% **{a}->{d,e} confidence =4/5=80%** {d}->{a,e} confidence =4/9=45% {e}->{a,d} confidence =4/6=67% {b,d}->{e} confidence =4/6=67% **{b,e}->{d} confidence =4/4=100%** {d,e}->{b} confidence =4/6=67% {b}->{d,e} confidence =4/7=67% {d}->{b,e}confidence =4/9=67% {e}->{b,d} confidence =4/6=67% **Strong rules are highlighted.**	9
Issues on measurement and data collection aspect of data quality.(Each with an explanation carries -1m) Measurement and data collection errors ,Noise and Artifacts, Precision, Bias AND accuracy, Outliers, Missing values, Inconsistent values, Duplicate values	6
I) (Each carries 2m) **i)** {1,2,3,4},{1,2,4,5},{1,2,3,5},{2,3,4,5},{1,3,4,5} **ii)** {1,2,3,4}, {1,2,3,5},{2,3,4,5} **II) Each exaplantion + digram---2.5m**	9
• Eliminate Data Objects or attributes • Estimate missing values • Ignore the missing value during analysis **Each with advantages and disadvantages carries 2m**	6
i)	9



Single itemset ---(1m)
2-itemset---(2m)
3-itemset –(2m)

i) Obtain the strong association rules, given that the confidence th
70%. --------(4m)

Rule Generation
{I1,I2}->I3 Confidence =2/4 =50%
{I1,I3}->I2 Confidence =2/4 =50%
{I2, I3}->I1 Confidence =2/4 =50%
{I1} ->{I2, I3} Confidence =2/6 =33%
{I2}->{I1,I3} Confidence =2/7 =29%
{I3}->{I1,I2} Confidence =2/6 =50%
{I1,I2}->I5 Confidence =2/4 =50%
{I1,I5}->I2 Confidence =2/2=100%
{I2, I5}->I1 Confidence 2/2=100%
{I1} ->{I2, I5} Confidence =2/6=33%
{I5}->{I1,I2} Confidence = 2/2=100%
If the minimum confidence threshold is 70% then 2nd,3rd, and last rules ar
rules

Ap- Apply

M S RAMAIAH INSTITUTE OF TECHNOLOGY
(Autonomous Institute, affiliated to VTU)
DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

Term:	10/8/2018 to 29/12/2018	**Course Code:**	IS721
Course:	Data Mining	**Semester:**	VII – A, B & C
CIE:	Test – II	**Max Marks:**	30
Date:	19-11-2018		

Instructions to Candidates: Answer any two out of three questions, Mobile phones and programmable calculators are banned.

No	Questions	Marks	Bloom's Level	CO
	Generate frequent item sets for the following transactions in the given table using FP growth approach. Given min support =3.	7	AP	2

Trans Id	Items Bought
1.	{a,d,e}
2.	{a,b,c,e}
3.	{a,b,d,e}
4.	{a,c,d,e}
5.	{b,c,e}
6.	{b,d,e}
7.	{c,d}
8.	{a,b,c}
9.	{a,d,e}
10.	{a,b,e}

a. Discuss some of the characteristics of Decision Tree Induction	8	AP	3
b. Illustrate two potential causes of model over fitting with suitable examples.	7	U	2
Describe the properties of the objective measures. Indicate a measure for each.	8	AP	3

Consider the following transaction table

Tid	1	2	3	4	5	6	7	8	9	10
A1	S	M	S	M	D	M	D	S	M	S
A2	Y	Y	N	N	Y	Y	N	N	Y	Y
Class label	no	no	no	no	yes	no	no	yes	no	yes

Compute the Gini index for the multiway split of attribute A1 and binary split of A2 and indicate the best attribute to split.
Compute the classification error for the binary split of attribute A1 and A2. Indicate best attribute to split.

Consider the following Market basket transactions	7	AP	2

TID	Item list
T1	a,b,d,e
T2	b,c,d
T3	a,b,d,e
T4	a,c,d,e
T5	b,c,d,e
T6	b,d,e
T7	c,d
T8	a,b,c
T9	a,b,d
T10	b,d

a) Draw the contingency table for each of the following rules using the transactions shown in the above table
Rules: {b}->{c}, {a}->{d}, {b}->{d},{e}->{c},{c}->{a}
b) Use the contingency tables in part (a) to compute and rank the rules in decreasing order according to the following measure
i)Lift ii) Confidence

Describe the algorithm for Decision Tree Induction. Also discuss the design issues of Decision tree Induction	8	U	3

U-Understand, AP- Apply

Tree Generation and Writing all the frequent itemsets.

3(a) a. Characteristics of Decision Tree Induction

3(b) b. Illustrate two potential causes of model over fitting with suitable examp

2a) Properties of the objective measures namely: Inversion Property
Null Addition Property
Scaling Property

2 b) Consider the following transaction table

Tid	1	2	3	4	5	6	7	8	9	
A1	S	M	S	M	D	M	D	S	M	S
A2	Y	Y	N	N	Y	Y	N	N	Y	Y
Class label	no	no	no	no	yes	no	no	yes	no	ye

Compute the Gini index for the multiway split of attribute A1 and bina
of A2 and indicate the best attribute to split.

$Gini(A1=S) = 1-(2/4)^2-(2/4)^2 = 0.5$

$Gini(A1=M) = 1-(0/4)^2-(4/4)^2 = 0$

$Gini(A1=D) = 0.5$

$Gini(A1) = (4/10)*0.5 +(4/10)*0 + (2/10)*0.5 = 0.3$

$Gini(A2 = Y) = 0.44$

$Gini(A2= N) = 0.375$

$Gini(A2) = 0.6*0.44 + 0.4 *0.375 = 0.414$

A1 is the best attribute to split.

Compute the classification error for the binary split of attribute A1 and A
Indicate best attribute to split.

Classification error(A1={S,M}) = 1-max[2/8,6/8] = 2/8

Classification error(A1={D}) = 1-max[1/2,1/2] = ½

Classification error(A1={S,M},{D}) = 8/10*2/8 + 2/10*1/2 = 3/10 = 0.

Classification error(A1={S,D},{M}) = 0.7

Classification error(A1={D,M},{S}) = 0.3

Best combination is either A1={S,M},{D} or A1={D,M},{S}

Classification error (A2 = {Y}) = 1-max[2/6,4/6] = 2/6

Classification error (A2 = {N})= 1-max[1/4,3/4] = 1/4

Classification error(A2) = 6/10*2/6+4/10*1/4 = 0.3

Best split as per classification error is either A1 or A2

3 a) Contingency tables for the rules:

{b}->{c}

	c	c'
b	3	5
b'	2	0

{a}->{d}

	d	d'
a	4	1
a'	5	0

{b}->{d}

	d	d'
b	7	1
b'	2	0

{e}->{c}

	c	c'
e	2	3
e'	3	2

{c}->{a}

	a	a'
c	2	3
c'	3	2

Decreasing order of Lift and Confidence

Lift	Confidence
{b}->{d}	{b}->{d}
{a}->{d}	{a}->{d}
{e}->{c}	{e}->{c}
{c}->{a}	{b}->{c}
{b}->{c}	{c}->{a}

Hunt's Algorithm

Design Issues

M S RAMAIAH INSTITUTE OF TECHNOLOGY

(Autonomous Institute, affiliated to VTU)

DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

Term:	30 Aug to 29th Dec 2018	**Course Code:**	IS721
Course:	Data Mining	**Semester:**	VII – A, B & C
CIE:	Test – III	**Max Marks:**	30
Date:	21-12-1018		

Instructions to Candidates: **Answer any two out of three questions, Mobile phones are banned.**

Portions for Test: Lecture Nos. from 33 to 50 as per lesson plan

Questions	Marks	Bloom's Level	CO
Explain k-nearest neighbor algorithm with an example.	5	U	3
Given a relational table where patients are described by binary attributes. Excluding name, all other attributes are asymmetric binary variables. Name | Fever | Cough | Test1 | Test2 | Test3 | Test4 J | 1 | No | P | N | N | N M | 1 | No | P | N | P | N K | 1 | Yes | N | N | N | N Compute the distance between each pair of the 3 patient objects.	6	AP	4
Explain the basic measures for text retrieval.	4	U	5
Write a note on bagging.	4	U	3
Find out the three clusters using K-mean clustering for below points with k is 3 and initial centroids are A1(2,10), B1(5,8) and C1(1,2). Use Euclidean distance function and show only 1) The three cluster centers after the first round execution 2) The final three clusters. A1(2,10) , A2(2,5),A3(8,4), B1(5,8), B2(7,5),B3(6,4), C1(1,2),C2(4,9)	7	AP	4
Elaborate on the keyword based approach.	4	U	5
The following table shows the midterm and final exam grades obtained for students in a database course. Predict the final examination grade of a student who received an 86 on the midterm examination. x(MidTerm) | y(Final Exam) 72 | 84 50 | 63 81 | 77 74 | 78 94 | 90	5	AP	3
What is cluster analysis? Explain any two types of clustering methods	6	U	4
Document/term | t1 | t2 | t3 | t4 | t5 | t6 | t7 d1 | 0 | 4 | 10 | 8 | 0 | 5 | 0 d2 | 5 | 19 | 7 | 16 | 0 | 0 | 32 d3 | 15 | 0 | 0 | 4 | 9 | 0 | 17 d4 | 22 | 3 | 12 | 0 | 5 | 15 | 0 d5 | 0 | 7 | 0 | 9 | 2 | 4 | 12 From the above table calculate the TF-IDF value of a term t4 in d2.	4	AP	5

AP- Apply , U- Understand

M S RAMAIAH INSTITUTE OF TECHNOLOGY

(Autonomous Institute, affiliated to VTU)

DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

Term:	30th Aug to 29th Dec 2018	**Course Code:**	IS721
Course:	Data Mining	**Semester:**	VII – A, B & C
CIE:	Test – III	**Max Marks:**	30
Date:	20-12-2018		

SCHEME

Questions	Marks	Bloom's Level	CO
Explanation of k-nearest neighbot algorithm with an example	5	U	3
Given a relational table where patients are described by binary attributes. Excluding name, all other attributes are asymmetric binary variables.	2*3=6	AP	4

Name	Fever	Cough	Test1	Test2	Test3	Test4
J	1	0	1	0	0	0
M	1	0	1	0	1	0
K	1	1	0	0	0	0

D(J,M) = 0+1/2+0+1= 0.33
D(J,K)= 1+1/1+1+1= 0.67
D(M,K)=1+2/1+1+2=0.75

Explanation the basic measures for text retrieval. 1) Precision 2) recall	2*2=4	U	5
Note on bagging.	4	U	3
1) The three cluster centers after the first round execution C1=(2,10) C2=(6,6) C3=(1.5,3.5) 2) Final clusters G1={A1} G2={A3,B1,B2,B3,C2} G3={A2,C1}	3 4	AP	4
Explanation on the keyword based approach.	4	U	5

3a) The following table shows the midterm and final exam grades obtained for students in a course. Predict the final examination grade of a student who received an 86 on the midterm examination.

x(MidTerm)	y(Final Exam)
72	84
50	63
81	77
74	78
94	90

x mean= 74.2

y mean = 78.4

w1=0.598

w0=34.83

y= 34.83+0.598*x

when x=86

y=86.2

3b) Definition of cluster analysis

Explanation any two types of clustering methods

3c)

Document/term	t1	t2	t3	t4	t5	t6	t7
d1	0	4	10	8	0	5	0
d2	5	19	7	16	0	0	32
d3	15	0	0	4	9	0	17
d4	22	3	12	0	5	15	0
d5	0	7	0	9	2	4	12

From the above table calculate the TF-IDF value of a term t4 in d2.

TF(d2,t4)= 1.3432

TF(t4)=0.1760

TF_IDF(d2,t4)= 0.2355

AP- Apply , U- Understand

RAMAIAH INSTITUTE OF TECHNOLOGY
(Autonomous Institute, affiliated to VTU)
DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

Term:	28.8.18 – 30.12.2018	**Course Code:**	18722
Course:	Distributed Computing	**Semester:**	VII
CIE:	1	**Max Marks:**	30
Date:	3.9.18	**Time:**	3.30-4.30 PM

Instructions to Candidates: Answer any two questions,Portions:1.1-1.14 , Modules banned

S.No.	Questions	M	Bloom's level	CO
Q1	**A.** List out the different parallel computers. Describe logical organization of a BlueGene / L node. **B.** With the help of example explain different types of data dependence	8 7	Remember Understand	1 2
Q2	**A.** Define false sharing. Write a program to remove false sharing in count3 computation. **B.** Enumerate the different performance metrics	8 7	Understand Analyze	1 2
Q3	**A.** Compute the sum of given sequence number using prefix sum {7,3,15,10,13,18,6,4} **B.** . Illustrate the different sources of performance loss.	8 7	Apply Understand	1 2

RAMAIAH INSTITUTE OF TECHNOLOGY
(Autonomous Institute, affiliated to VTU)
DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

Term:	28.8.18 – 30.12.2018	**Course Code:**	18722
Course:	Distributed Computing	**Semester:**	VII
CIE:	1	**Max Marks:**	30
Date:	3.9.18	**Time:**	3.30-4.30 PM

Answer scheme

Ans 1 a.

List of the different parallel computers : chip multiprocessors, symmetric multiprocessor arch, heterogeneous chip design,clusters,supercomputers **2 Marks**

Logical organization of a BlueGene / L node.

Diagram : **3 Marks**
Explanation **4 Marks**

Ans 1 b.

Data dependence **4 Marks**

– Preserved to maintain correctness

– Flow dependence: read after write

– Anti dependence: write after read

– Output dependence: write after write

– Input dependence: read after read (memory reuse)

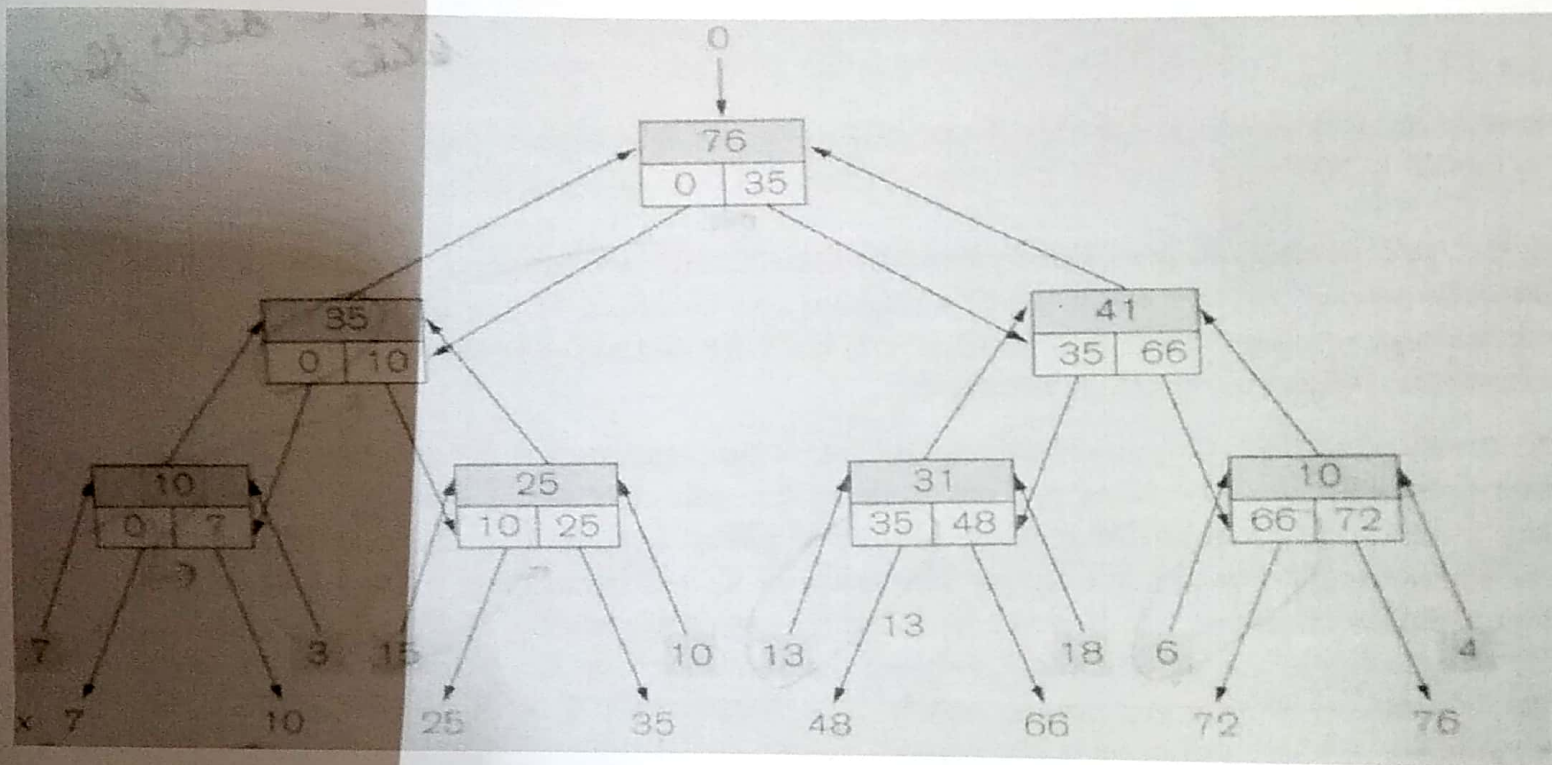
Example : **2 Marks**

Ans 2 a. :
Definition of false sharing : **3 marks**
Program on count 3 : **6 Marks**

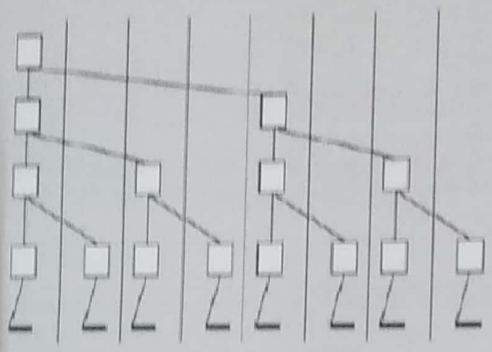
```
1  struct padded_int
2  {
3    int value;
4    char padding[60];
5  } private_count[MaxThreads];
6
7  void count3s_thread(int id)
8  {
9    /* Compute portion of the array this thr
       work on */
10   int length_per_thread=length/t;
11   int start=id*length_per_thread;
12
13   for(i=start; i<start+length_per_thread
14   {
15     if(array[i] == 3)
16     {
17       private_count[id]++;
18     }
19   }
20   mutex_lock(m);
21   count+=private_count[id].value;
22   mutex_unlock(m);
23 }
```

Ans 2 b.
Explanation of the different sources of performance loss : overhead, non-parallelizable code, contention, ideal time **6 marks**

Ans 3 a. :
Compute the sum of given sequence number using prefix steps **3 marks**
sum. {7,3,15,10,13,18,6,4} **6 marks**



M S Ramaiah Institute of Technology
Department of Information Science and Engineering 15 73 2
Distributed Computing Scheme

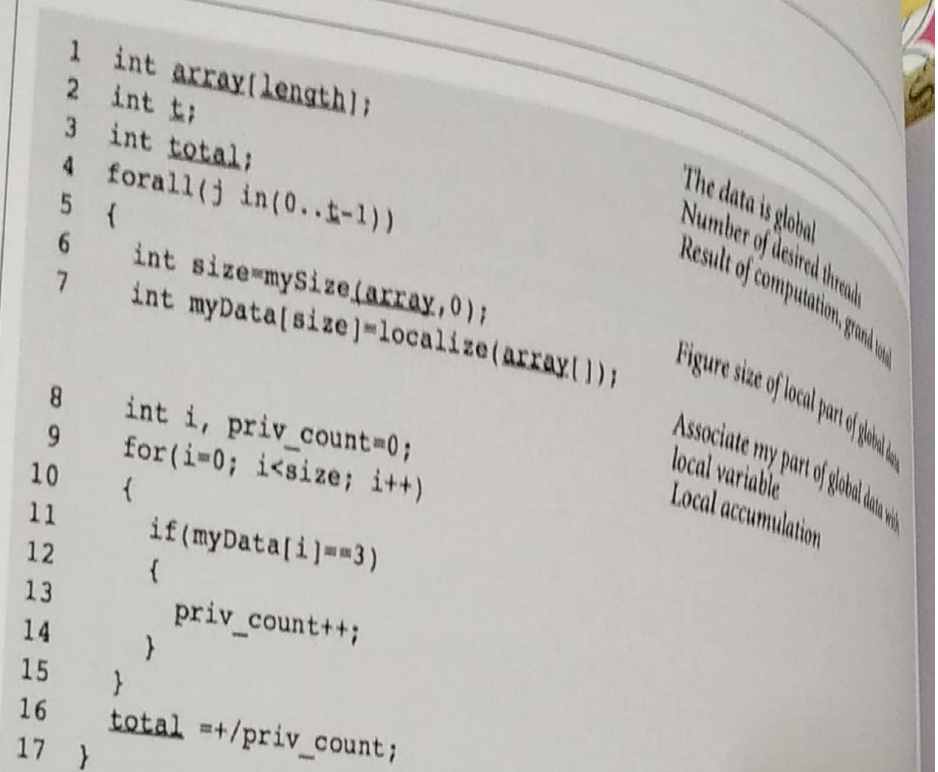
Q. No	Questions	Marks
1a.	1. Fixed Parallelism 2. Unlimited Parallelism 3. Scalable Parallelism	3 3 4
1b.	Identify and explain any five uses of Generalized Reduces and Scans. To understand the power of reduces and scans, consider the following problems, along with a description of their solution. Here, we describe each solution sequentially, but each can be solved in parallel using customized reduce operations: • Second smallest array element • Histogram (k-way) • Length of longest run of 1s • Index of first occurrence of x • Team Standings • Keep the longest sequence of 1s • Index of Last Occurrence	5
2a.	Schwartz' Algorithm 	3

	1 int nodeval'[P];
	2
	3 forall(index in(0..P-1))
	4 {
	5 int tally;
	6 stride=1;
	7 ...
	8
	9 while(stride<P)
	10 {
	11 if(index%(2*stride)==0)
	12 {
	13 tally=tally+nodeval'[index+stride];
	14 stride=2*stride;
	15 }
	16 else
	17 {
	18 nodeval'[index]=tally;
	19 break;
	20 }
	21 }
	22
	23 }
	
	Global full/empty variables for saving the value from right child
	COMPUTE tally *HERE*
	Begin logic for tree
	Send initially to tree node
	Exit, if no longer a parent
		4

	Explanation	3

- Reading and Writing Global Memory
- Connecting Global and Local Memory
- Synchronized Memory

Diagram



3a.

```
1  int array[length];
2  int t;
3  int total;
4  forall(j in(0..t-1))
5  {
6    int size=mySize(array,0);
7    int myData[size]=localize(array[]);

8    int i, priv_count=0;
9    for(i=0; i<size; i++)
10   {
11     if(myData[i]==3)
12     {
13       priv_count++;
14     }
15   }
16   total =+/priv_count;
17 }
```

The data is global
Number of desired threads
Result of computation, grand total

Figure size of local part of global data

Associate my part of global data with local variable
Local accumulation

ompute grand total

3b.

i. Block Allocations

If our goal is to exploit locality, it follows that for most computations contiguous portions of a data structure should be allocated together on the same process. This is the familiar spatial and temporal locality that caches exploit.

ii. Cyclic Allocations

In algorithms in which the amount of work is not proportional to the amount of data, block allocations may suffer from poor load balance. For example, consider LU decomposition, a linear algebra computation that decomposes a matrix into the product of two matrices as a way to solve a system of linear equations.

iii. Overlap Regions

Our principle of operating on large blocks of computation leads to the concept of an overlap region, which is a type of software cache.

iv. Irregular Allocations

Of course, there are many algorithms that use data structures other than arrays. Because the data references are irregular—and often not known until runtime—it can be a very inefficient process to fetch the non-local values: Discover that a reference is non-local, fetch it, discover the next non-local reference, fetch it, and so on. To avoid this fine-grained, serial solution—and hopefully to apply sound parallel concepts—a technique called inspector/executor has been developed.

v. Applying the Generalized Scan

To illustrate the operation of the generalized scan, imagine an array A of integers from the sequence 1,k. The scan
lastOccurrence\ A
returns in index position i the index of the most recent occurrence of A [i] or it returns 0 if this is the first occurrence. We use as a tally an array of k elements, which is initialized to 0s; if A[i] is;, the accum() function stores i in tally [j] as the last occurrence; the combine () function takes the maximum of each element of the two arrays; and the scan generator reprocesses the block of data, using the ptally as its initial value.

M S RAMAIAH INSTITUTE OF TECHNOLOGY

Department of Information Science and Engineering

Term:	30/08/2018 to 29/12/2018	**Course Code:**	IS732
Course:	Distributed Computing	**Semester:**	7
CIE:	Test-III	**Max Marks:**	30
Date:	20-12-2018	**Time:**	2-3 PM

Instructions to Candidates: Answer any **Two** out of Three questions. **Mobile phones are banned**

Q.No	Questions	Marks	Bloom's Level	CO
1a.	What are POSIX threads? Define following POSIX thread functions i. Thread creation and destruction ii. Thread synchronization iii. Compare threads iv. Adding mutual exclusion	8	U	CO4
1b.	Compare blocking and non-blocking communication in MPI and Explain Other Communication Modes used in MPI.	7	An	CO5
2a.	Make use of openMP programming to parallelize Count 3s program and summarize the performance issues of POSIX threads while implementing parallelism	7	Ap	CO4
2b.	Explain following: i. Unified Parallel C ii. Titanium iii. Hidden Parallelism iv. Attached Processors	8	U	CO5
3a.	Explain following with respect to openMP: i. Reduction ii. Semantic Limitations on parallel for iii. Thread Behavior and Interaction iv. Sections	8	U	CO4
3b.	Explain any five MPI routine functions with their parameters	7	U	CO5

U – Understanding, Ap – Applying, An – Analyzing

1a. What are POSIX threads? Define following POSIX thread functions (4x2 = 8)

i. Thread creation and destruction

intpthread_create(pthread_t *tid,constpthread_attr_t *attr,void *(*start_routine)(void *

intpthread_join(pthread_ttidr, void **status);

ii. Thread synchronization (any 2 functions)

intpthread_cond_signal(pthread_cond_t *cond); // Condition to signal
intpthread_cond_broadcast(pthread_cond_t *cond);

intpthread_cond_wait(pthread_cond_t *cond,pthread_mutex_t *mutex);
intpthread_cond_timedwait(pthread_cond_t *cond,pthread_mutex_t *mutex,conststruct
*abstime);

iii. Compare threads

intpthreadequal(pthread t tl, pthread t t2);

Arguments:

Two thread IDs

Return value:

- Nonzero if the two thread IDs are the same (following the C convention).
- 0 if the two threads are different.

iv. Adding mutual exclusion

Acquiring and Releasing Mutexes

intpthread_mutex_lock(pthread_mutex_t *mutex);
intpthread_mutex_unlock(pthread_mutex_t *mutex);
intpthread_mutex_trylock(pthread_mutex_t *mutex);

OR

Dynamic allocation

intpthread_mutex_init(pthread_mutex_t *mutex,pthread_mutexattr_t *attr);
intpthread_mutex_destroy(pthread_mutex_t *mutex);
intpthread_mutexattr_init(pthread_mutexattr_t *attr);
intpthread_mutexattr_destroy(pthread_mutexattr_t *attr);

1b. Compare blocking and non-blocking communication in MPI

Blocking

intMPI_Send(void *buffer, int count, MPI_Datatype type, intdest, int tag, MPI_Comm *com
This routine has blockingsemantics, which means that the routine does not return until the m
been sent

int MPI Recv(void *buffer, int count, MPI Datatype type, int source, int tag, MPI Commcom
Status *status);
This routine has blockingsemantics—it does not return until the message is received.

Non-Blocking
MPI_Isend() - non-blocking version of the send operation and
MPI_Irecv() - non-blocking version of the receive operation

(3x1=3)

Other Communication Modes
Synchronous Send: MPI_Ssend() andMPi_issend()
Buffered send: MPI_Bsend( ) and MPI_Ibsend()
Ready send: MPI_Rsend() and MPI_Irsend()

(4)

2a. Make use of openMP programming to parallelize Count 3s program

```
int count3s()
{
  int i, count_p;
  count=0;
  #pragma omp parallel shared(array, count, length)\
    private(count_p)
  {
    count_p=0;
    #pragma omp parallel for private(i)
    for(i=0; i<length; i++)
    {
      if(array[i]==3)
      {
        count_p++;
      }
    }
    #pragma omp critical
    {
      count+=count_p;
    }
  }
  return count;
}
```

Summarize the performance issues of POSIX threads while implementing parallelism (3x1=3)
Granular issues
Thread Scheduling
Priority inversion

2b. Explain following:
i. Unified Parallel C (4x2=8)
- Extends c programming (pointer explanation)
- memory model

ii. Titanium

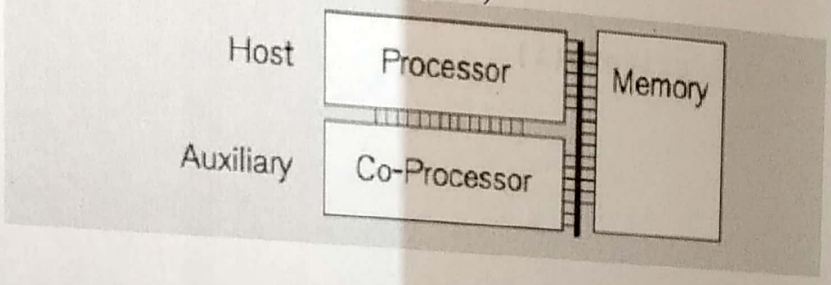
- an extension of Java that executes on distributed memory parallel computers
- memory model

iii. Hidden Parallelism (any four features)

the following features thatdeliver hidden parallelism, listed roughly in chronological o

- Bit-parallel functional units
- Multiple functional units
- Pipelined execution
- Out of order execution
- Increasingly wide data-paths
- DMA controllers
- Prefetch units
- Trace caches
- Simultaneous multi-threading
- Vector processors
- Chip multi-processors
- Co-processors: I/O controllers, network controllers, graphics co-processors

iv. Attached Processors (explain)



3a. Explain following with respect to openMP:

i. Reduction

- Explain reduction with an operation
- eg. #pragma omp parallel for reduction (+, count)

ii. Semantic Limitations on parallel for

- List the limitations of parallel for

iii. Thread Behavior and Interaction

- thread behavior at the beginning and ending of pragma

iv. Sections

- sections pragma explanation

3b. Explain any five MPI routine functions with their parameters

Listing five MPI routine functions
Defining them

M S Ramaiah Institute of Technology

Department of Information Science and Enginerering

Term:30/08/18 to 29/12/18.
Course: Virtual and Augmented Reality.
CIE: Test 1.
Date:04/10/18

Course Code:IS71C3
Semester:7
Max Marks:30
Time:3 30-4 30

Scheme of Evaluation

1a. Explain the features and characters of human eye relevant to the design of display systems.

->Features and characters of human eye relevant to the design of display systems are as follows:

- Accomodation
- The pupil
- The retina
- Visual acuity
- Light and dark adaption
- Peripheral vision
- Persistence of vision
- Stereopsis
- Visual field
- Synthetic images versus reality

1b.Discuss the basic principles associated with sound transmission, the ear's action and how we utilise sounds in 3D space.

-> The basic principles associated with sound transmission, the ear's action and how we utilise sounds in 3D space are as follows:

- Sound
- Sound perception
- Frequency range
- Sound intensity
- Sound direction
- The sound stage
- Head-related transfer functions
- Measuring HRTFs
- Ambisonics

2a.Design a data glove from the context of haptic technology for tactile sensation.

->It must be tracked in three dimensions ,

- The finger positions may be monitored and the glove may even incorporate pressure pads that inflate when a virtual touch condition is detected.
- To monitor the glove's position , a sensor is needed to provide the x-, y- and z- coordinates.
- Pressure pads must give a tactile stimulus when activated.

2b.How tracking sensors and hardwareworks,explain with the help of Flock of Birds and Space ball or any two techniques of your choices.

->Spaceball:

- It is an input device for measuring simultaneous translations and rotations about the X-, Y- and Z- axes.
- When it is gripped by the user, forces and torques are measured and input to host computer.
- The approximate force range is 0.5-20N and the approximate torque range 15-600N mm.

Logitech Head Tracker:

- It measures absolute position and orientation using ultrasonic speakers and microphones.
- Three speakers are mounted in a large fixed triangle, while three microphones are housed in a small triangle that is normally attached to the user's head.
- The technique of triangulation enables the position of the mobile triangle to be specified relative to the fixed triangle.

3a.Write a short note on Functions for segmenting the display file.

-> Functions for segmenting the display file areas follows:

- Segment creation
- Closing a segment
- Deleting a segment
- Renaming a segment

3b.What is double buffering explain the free storage allocation from the context of display file compilation.

->Double buffering:

It is a technique for drawing graphics that shows no stutter, tearing and other artifacts.

- Free storage allocation explanation

Virtual and Augmented Reality

Test-2 Scheme & solutions

1 a. With the help of neat sketches eplain the integration of the various elements of a generic VR system.

Explain with diagram vairious comonents.

->VE can take many forms,for example it could be a realistic representation of some physical environment like interior of building,a kitchen etc.

->Interactive 3D modelling software can be used to construct such an interactive environment.

->They also include engineering drawings,architectural plans and the incorporation of important physical dimensions.

->The environment or parts may already exist as a CAD database.

->In the above case some form of conversion software is required to translate it to suitable form.

->It might be 3D database of goographical,heirarchical network.

->It could also be multidimensional dataset associated with stock transactions.

->What ever the nature of underlying data,a geometric model is required to represent atomic entities and their relationships with each other.

1 b.What are the stages of Haptic Rendering?Explain the pipeline with the help of appropriate example.

->The three stages of Haptic rendering are:

.Application stage

.Geometry stage

.Rasterizer stage

->The application stage is done entirely in software by CPU.

->It reads the world geometry database as well aas the users input mediated by devic
mice,trackballs,trackers,or sensing gloves.

->In response the application stage may change the view of the simulation.

->The application stage results are fed to the geometry stage.

->Geometry stage is either implemented in software or hardware.

->This stage consists of model transformations,lighting computations,scene
projection,clipping, and mapping.

->The lighting substage calculates the surface color based on the type and the number
simulated light sources in the scene.

->The last stage is the rasterizing stage,which is done in hardware,in order to gain spee

->This stage converts the vertex information output by geometry stage into pexel inform
needed by video display.

->It also performs antialiasing in order to smooth out the jagged appearence of polygon
edges.

->Example: the HP visualize fx pipeline architecture.

2 a. With the help of PC VR Engine adopted Intel Co.explain how to build PC graphics

->Most important for the real time requirements of the VR simulations are PC CPU speed and the rendering capability of reciding graphics card.

->The PC architecture untic 1990 had a boottleneck that hindered the performance.

->this was because of the slow PCI(peripheral component interface).

->Modern PC such as adopted by Intel solve PCI bandwidth problem by introducing Intel Accelerated graphics port.

->The AGP bus operates at much higher bandwidth.

->They transfers the textures and other graphics data directly from the system RAM to the video memory on the graphics card.

->The transfer is mediated by amemory controller chip.

->The AGP transfer rate on current PC is over 2GBPS using AGP 8x bus.

->At the above rates it is possible to reverse part of the RAM to serve as secondary video memory.

->because of that less memory needs to be allocated for the graphics card.

->PC graphics accelerators-they are typically cards produced by third parties to retrofit mid to high range pc's

->ex. The ATI Fire GL2

->The Elsa Gloria lil

->The xBox

2 b. List various methods by which virtual object surfaces/shapes can be constructed,Evaluate any three among them.

->The various methods are:

->Using a toolkit editor

->Import CAD files

->Creating surfaces with a3D Digitalizer

->Creating surfaces with a 3D scanner

->Using online 3D object database

using a toolkit editor :

- All graphical programming languages allow 3D polygonal object shape to be

->Text base specification of vertex coordinates and connectivity is necessary to
shape and normals need to be specified if smooth shading is required.

->The proces is based on trial and error iterations and requires a amount of time
proportional to the programming skills.

Importing CAD files:-

->These programs are the de facto standard in mechanical and architectural design.

->Thus the pre existing models of mechanical buildings or assemblies can be used.

->They have to be created and saved in separate files.

->Once they are saved they have to be converted to desired format.

Using online 3D database:

->This involved purchising already created 3D models from commercial house.

->Such models will have three levels of detail low(L),medium(M),high(H).

->Each model comes in two data files.

->It is then possible to use these graphical libraries to transform into wireframe

3 a. Explain the three parameters which are involved in kinematics modeling

->Homogeneous transformation matrices:

->4x4 transformation matrices were used to express object translations,rotations and scaling.

->Ahomogeneous transformation matrix is given by geometric equatoin.

$$T = \begin{matrix} R(3x3) & p(3x1) \\ 0\ \ 0\ \ 0 & 1 \end{matrix}$$

->Where R(3x3) is the rotation submatrix expressing the orientation of the system of coordinates B with respect to the coordinates of A.

->the submatrix represents the projection of B system of coordinates along the tried of vectors of A coordinates.

->Object position:

->We know that object surface model uses (x,y,z) coordinates expressed in an object system of coordinates.

->This system of coordinates is attached into object,usually at the center of gravity,and oriented along the object axex of symmetry.

->When the object moves in the virtual world.its system of coordinates moves with it .

->There fore the position and orientation of object vertices in the object system of coordinated remains invarient.

->The above fact is true as long as the object surface does not deform or it is not cut.

->Object hierarchies:

->Object heirarchies define groups of objects which mov together as a whole but whose parts can also move independently.

->A heirarchy implies atleast two levels of virtual objects.

->The higher level objects are called parents.

->The motion of parent objects is replicated by all its children objects.

->However a child object can move without effecting the parent objects position.

->Child objects can usually be replicated several times in the heirarchy.

->The parent -child heirarchy is mathematically described by the same kind of homogen transformations as in the first point.

3 b. Design Behaviour modelling and model management from the context of VR.AR

->Behaviour modelling:

->An virtual human is a 3D model character that has a human behaviour.group of such ag called as crowds and have crowd behviour.

->Whenever object interacted,it was assumed that one was controlled by the user.

->It is also possible to model object behaviour that is independent of usres actions.

->Example, a guided agent needs a human specified path in order to travel from one locatio another.A guided door has to have its degree of opening controlled either by the user or by and so on.

->At the other extreme an autonomous agent percieves information about surrounding virtu environment and decides what path to follow accordingly.

->An autonomous door is in control of its motion and can even guide an agent to control it

->Fully autonomous agents,such as virtual football player in the light uniform need to perce environment in order to take appropriate actions.

->The agent behaviour has its heirarchy.reflex behaviour being at the lowest heirarhy.

=>However a child object can move without effecting the parent object.

=>Child objects can usually be replicated several times in the heirarchy.

=>The parent -child heirarchy is mathematically described by the same kind of

transformations as in the first point.

3 b. Design Behaviour modelling and model management from the context of V.E.

=>Behaviour modelling:

=>An virtual human is a 3D model character that has a human behaviour group of

called as crowds and have crowd behviour.

=>Whenever object interacted,it was assumed that one was controlled by the user.

=>It is also possible to model object behaviour that is independent of users action.

=>Example, a guided agent needs a human specified path in order to travel from one

another.A guided door has to have its degree of opening controlled either by the user

and so on.

=>At the other extreme an autonomous agent percieves information about surrounding

environment and decides what path to follow accordingly.

=>An autonomous door is in control of its motion and can even guide an agent to cross

=>Fully autonomous agents,such as virtual football player in the light uniform need to

environment in order to take appropriate actions.

=>The agent behaviour has its heirarchy.reflex behaviour being at the lowest heirarchy

->Model Management:-

->Geometrical,kinematics,phisical and behaviour modelling of a highly populated virtual world will result in a very complex model.the resultant large computations load is difficult to manage.

->These problems are faced by architects having to virtualize large buildings with many floors,offices etc.

->The same problems are faced by surgical simulators.

->Model management combines techniques designed to help with the VR engine render complex models as descrined erlier.

->There are several model management approaches like cell segmentation,off-line precomputation,database management etc.

->Level of detail (LOC) combines methods used to impove the graphics pipeline throughput by selecting object level of detail appropriately.

->Static methods are based on object discrete geometry,alpha blending,morphing.

->Discrete geometry LOD managements is the simplest and was the first to be implemented.

->Objects are loaded by the application stage of the graphics pipeline based on the distance from the camera.

->Cell segmentation-This involves partitioning the large model into smaller ones.

->Automatic cell segmentation -This method partitions the VR universe into the samller ones.only the objects in the current cell will be rendered and rest will be omitted.

MOBILES ARE BANNED

M.S.Ramaiah Institute of Technogy
(Autonomous Institute, Affiliated to VTU)
Department of Information Science & Engineering

Term:	**30.08.2018 to 29.12.2018**	**Course Code:**	**IS71C3**
Course:	**Virtual & Augmented Reality**	**Semester :**	**VII**
CIE:	**Test – III**	**Max Marks:**	**30**
Date:	**21.12.2018**	**Time:**	**2 to 3 pm**

Portions for Test: Lecture Nos. from36 to 54 as per lesson plan .

Instructions to Candidates:Answer any two Questions

Questions	**Marks**	**Bloom's Level**	**CO**
With the help of neat sketches, explain the Simplified optics model of an HMD, which suits to human Visual System.	7	Remember	2
Explain any 4 mechanism of the following: a) Mechanical Trackers b) Magnetic Trackers c) Ultrasonic Trackers d) Optical Trackers e) Hybrid inertial Trackers	8	Understand	2
Design and draw the architecture of the CAVE with explanation	7	Create	2
Discuss 1. Conventional, 2. Computer assisted Animation, 3.Interpolation 4.Simple animation effects.	8	Analyze	5
Explain the 3 categories of animation languages.	7	Understand	3
Distinguish the methods of controlling animation	8	Analyze	3

1.A)

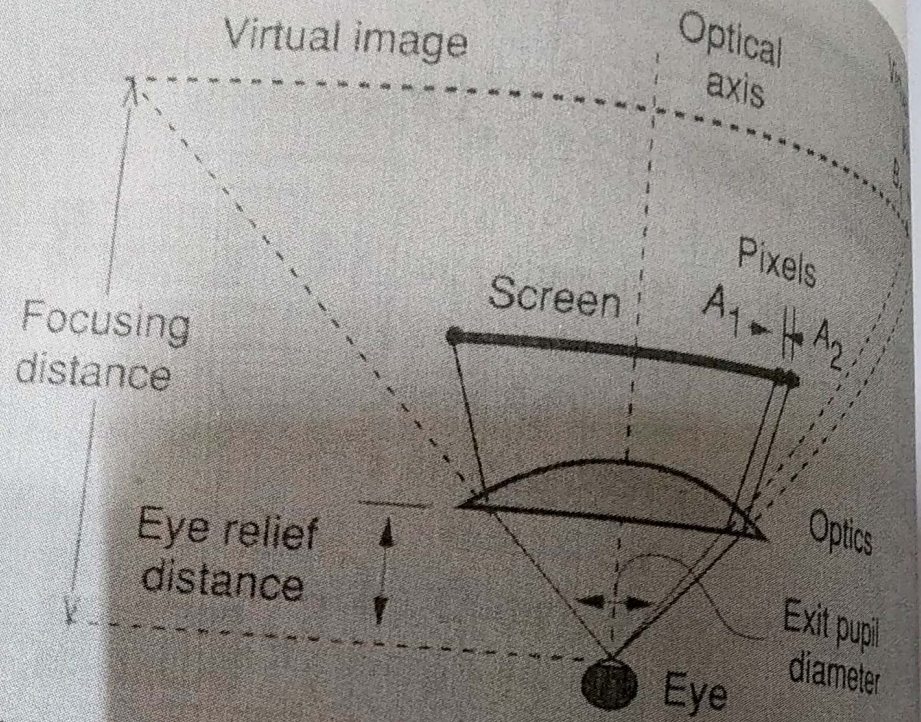


Fig. 3.2 Simplified optics model of an HMD. Adapted from Robinett
©1992 Massachusetts Institute of Technology. Reprinted by permission

- Human visual perception mechanism
- Field of view
- Depth perception

1.B)

2.a)

Tracking System	Characteristics	Advantages	Disadvantages
Optical Passive Markers	Use of cameras and reflective markers.	- Precision < 1mm - Wireless - Less burden	- Position only - Limited measurement space - Occlusions - Post-processing latency
Optical Active Markers	Use of cameras and infrared emitting diodes.	- Precision < 1mm - Wireless - Higher range than passive	- Position only - Limited measurement space - Occlusions - Post-processing latency - Frequency divided by sensors - Need of wires to connect the markers.
Optical Markerless	Use of cameras but without markers (based on image segmentation).	- Wireless - Flexible - No sensor burden - Contextual information	- High noise - Occlusions - High post-processing cost - Generally not real-time - High sensitivity to lighting
Electromagnetic	Based on the use of electromagnetic devices, with an emitter antenna and a tracking sensor.	- Portable - Wireless - Flexible sensor arrangement	- Limited range - No reference position - Magnetic disturbances
Mechanical	Based on the use of mechanical elements to determine the relative orientation of the elements.	- Portable - Wireless - Robust, reliable	- Restrictive movement - No reference position - Relative orientation only.
Inertial	System using accelerometers and potentiometers to measure the orientation of the object	- Accelerations - Precision < a degree - Portable - Wireless - Fast calibration	- No reference position - Post-processing – external contacts - Noise - Magnetic disturbances

2.a)

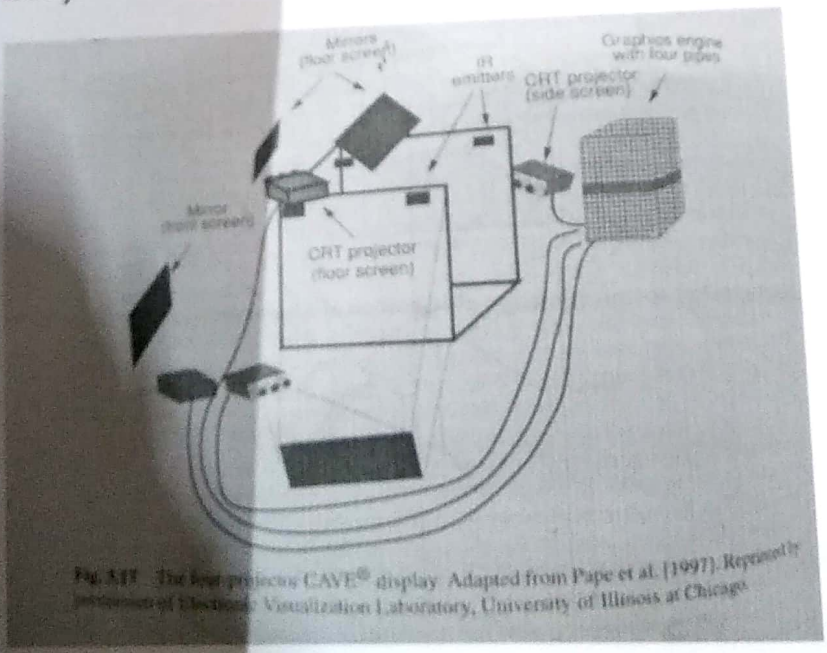


Fig. 3.17 The four-projector CAVE® display. Adapted from Pape et al. [1997]. Reprinted by permission of Electronic Visualization Laboratory, University of Illinois at Chicago.

CAVE architecture.

- Four projections are used
- Users wearing active glasses see a ver scene
- Newer version is called RAVE

2.B)

1.Conventional: Conventional animation is an stage of animations where sketch pencils and to create the animation

2. Computer assisted: Computer software will animation which will be more easier and faster conventional.

3.Interpolation:

3.A)

- Linear-List Notations

  Each event in the animation is described by a starting point and ending frame number and the action to take place is event.

- General-Purpose Language

  Another way to describe animation is to embed animation capability within a general purpose programming languages.

- Graphical Languages

  Graphical animation languages describe animation in a more visual way. These languages are used for expressing, editing, and comprehending the simultaneous changes taking place in an animation.

Methods of controlling animations are:

- Full Explicit control
- Procedural control
- Constraint-Based Systems
- Tracking live Action
- Actors
- Kinematics and Dynamics

RAMAIAH INSTITUTE OF TECHNOLOGY
(Autonomous Institute, affiliated to VTU)
DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

Term:	30.08.2018 to 29.12.2018	**Course Code:**	18IS81
Course:	Deep Learning	**Semester:**	8th
CIE:	Test – 1	**Max Marks:**	30
Date:	3/10/18	**Time:**	12.00 – 01.00 pm

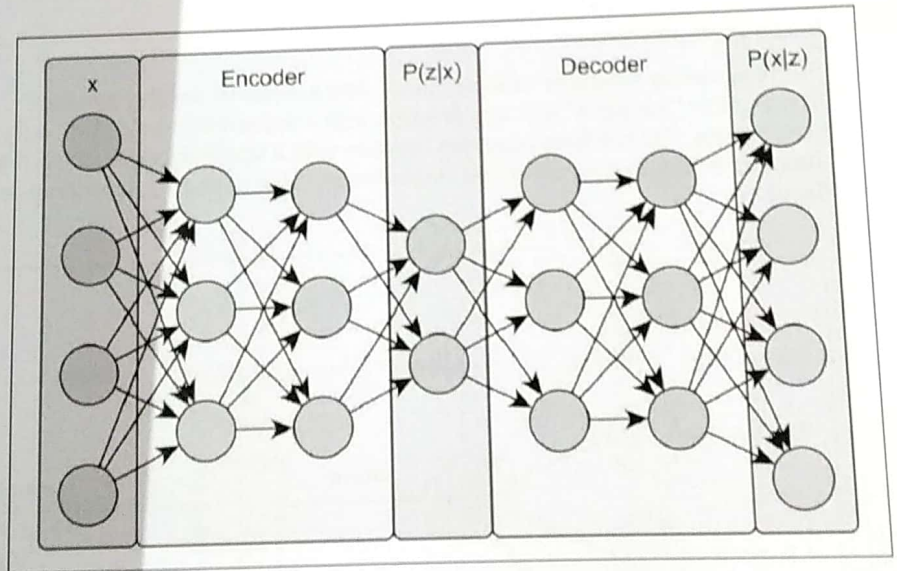
Portions for Test: Lecture Nos. from 01 to 15 as per lesson plan. **Instructions to Candidates:** Answer any two questions. **Note:** Mobiles and Programmable Calculators are strictly prohibited.

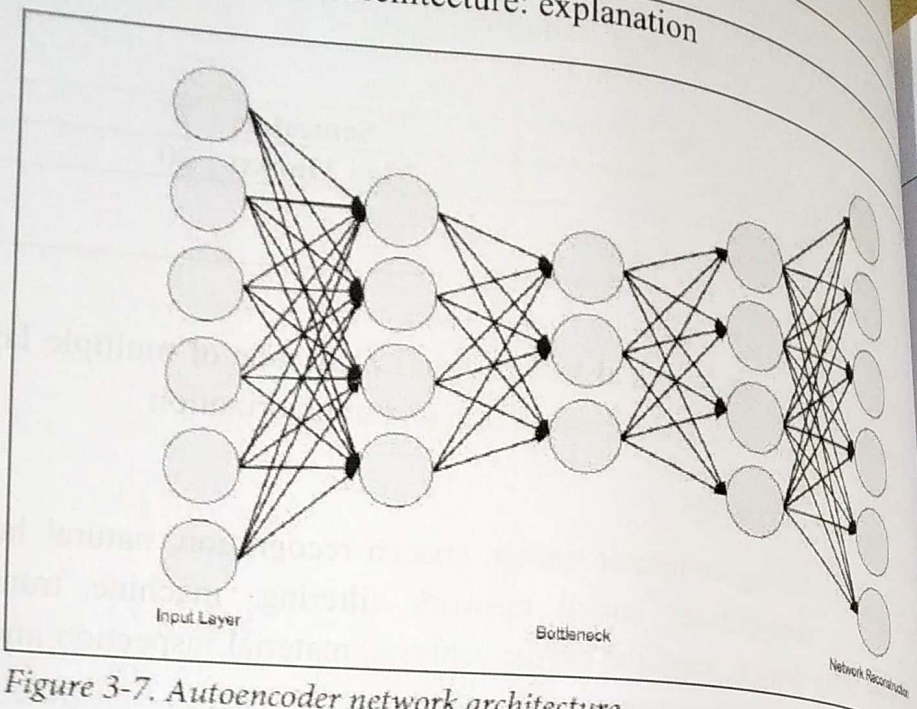
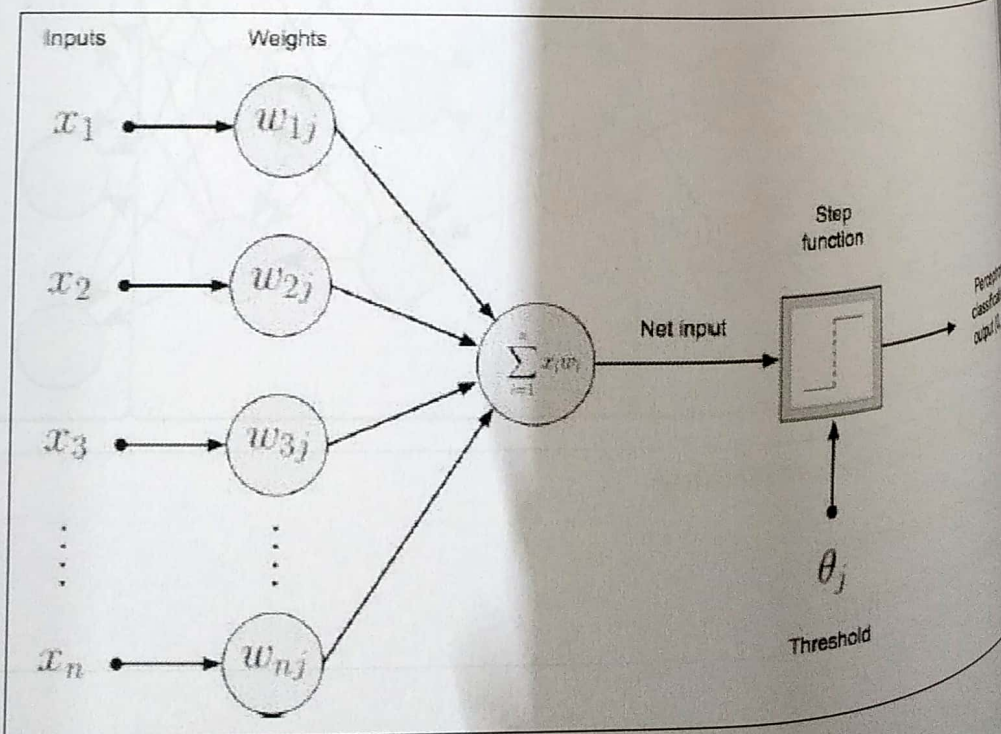
Sl. #		Question	Marks	Bloom's Level #	Cos
1.	a.	Discuss Deep Learning. List out the applications of deep learning.	8	U	CO 1
	b	Describe Variational Autoencoder network architecture & its mechanism of training.	7	R	CO 2
2.	a	List out the functions which is used to propagate the output of one layer's nodes forward to the next layer. Explain any two of them.	8	U	CO 1
	b	Describe Autoencoder network architecture.	7	R	CO 2
3.	a.	With block diagram explain Single-layer perceptron.	8	U	CO 1
	b.	Identify and Explain different parts of a basic Restricted Boltzmann Machines.	7	U	CO 2

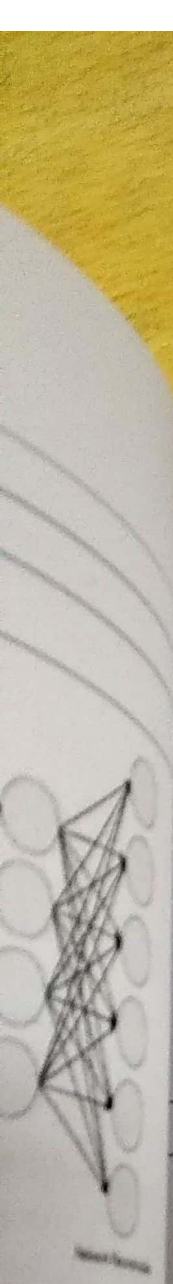
#R = Remember; U = Understand; A = Apply; An = Analyze;

RAMAIAH INSTITUTE OF TECHNOLOGY
(Autonomous Institute, affiliated to VTU)
DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING
Answer Scheme
Test 1

ourse/Code:	Deep Learning / IS71D4	**Semester:**	7th
CIE:	Test – 1	**Max Marks:**	30

		Marks
Answer a.	**Definition**: deep structured learning, hierarchical learning, use a cascade of multiple layers of nonlinear processing units for feature extraction and transformation **applications of deep learning:** applied to fields including computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics, drug design, medical image analysis, material inspection and board game program	5 3
b	Variational Autoencoder :  *Figure 3-8. VAE network architecture* + supporting statements Mechanism of training:	3 2 2
Answer a.	Following are the functions which is used to propagate the output of one layer's nodes forward to the next layer: Linear, Sigmoid, Tanh, Hard Tanh, Softmax, Rectified Linear A detail explanation of any two	2

	Answer b.	Autoencoder network architecture: explanation  *Figure 3-7. Autoencoder network architecture* Diagram:
3.	Answer a.	Single-layer perceptron: we're summing n number of inputs times their associated weights and then sending this "net input" to a step function with a defined threshold. Typically perceptrons, this is a Heaviside step function with a threshold value of 0.5. The function will output a real-valued single binary value (0 or a 1), depending on input.  *Figure 2-3. Single-layer perceptron*



Answer b.	There are five main parts of a basic RBM: • Visible units • Hidden units • Weights • Visible bias units • Hidden bias units + Explanation of each points	7

RAMAIAH INSTITUTE OF TECHNOLOGY
(Autonomous Institute, affiliated to VTU)
DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

Term:	30.08.2018 to 29.12.2018	**Course Code:**	IS71D4
Course:	Deep Learning	**Semester:**	7th
CIE:	Test – II	**Max Marks:**	30
Date:	22/11/18	**Time:**	9.30 – 10.30 am

Portions for Test: Lecture Nos. from 16 to 35 as per lesson plan. **Instructions to Candidates:** Answer any two questions. **Note:** Mobiles and Programmable Calculators are strictly prohibited.

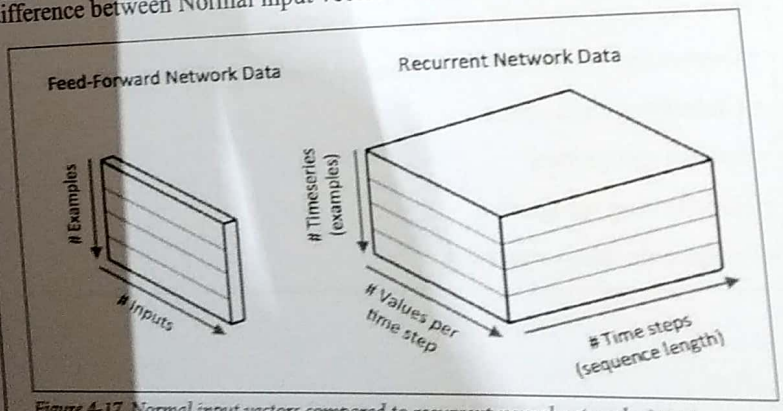
#		Question	Marks	Bloom's Level #	Cos
	a.	Describe Convolutional Neural Networks and its applications.	7	R	CO 2
	b	Is Recurrent Neural Networks (connectionist models) are better than Markov models? Justify your answer.	8	U	CO 3
	a	List and Summarize any two the major components of convolutional layers.	7	U	CO 2
	b	Discuss the difference between Normal input vectors and Recurrent neural networks input with the help of block diagram.	8	U	CO 3
	a.	With block diagram describe Deep Belief Networks architecture.	7	R	CO 2
	b.	Explain LSTM Networks and its properties.	8	U	CO 3

#R – Remember; U – Understand;

RAMAIAH INSTITUTE OF TECHNOLOGY
(Autonomous Institute, affiliated to VTU)
DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

Answer scheme **Semester:** 7th
Test 2
Course Code: IS71D4

Course: Deep Learning

1.	Answer a.	Convolutional Neural Networks and its applications. The goal of a CNN is to learn higher-order features in the data via convolutions. They are well suited to object recognition with images and consistently top image classification competitions. They can identify faces, individuals, street signs, platypuses, and many other aspects of visual data. Definition + application	4+3
	Answer b	Yes, Recurrent Neural Networks (connectionist models) are better than Markov models (and other time-window limited models) because they can capture the long-range time dependencies in the input data. Recurrent Neural Networks accomplish this because their hidden state captures information from an arbitrarily long context window and does not have the limitation of the other techniques. Moreover, the number of states they can model is represented by the hidden layer of nodes, and these states grow exponentially with the number of nodes in the layer	1+7
2.	Answer a	Major components of convolutional layers: • Filters • Activation maps • Parameter sharing • Layer-specific hyperparameters Explanation of any two the major components of convolutional layers.	1 3+3
	Answer b	difference between Normal input vectors and RNN input: explanation + diagram  *Figure 4-17. Normal input vectors compared to recurrent neural networks input*	5+3
3.	Answer a.	Explanation + diagram DBNs are composed of layers of Restricted Boltzmann Machines (RBMs)	4+3

for the pretrain phase and then a feed-forward network phase

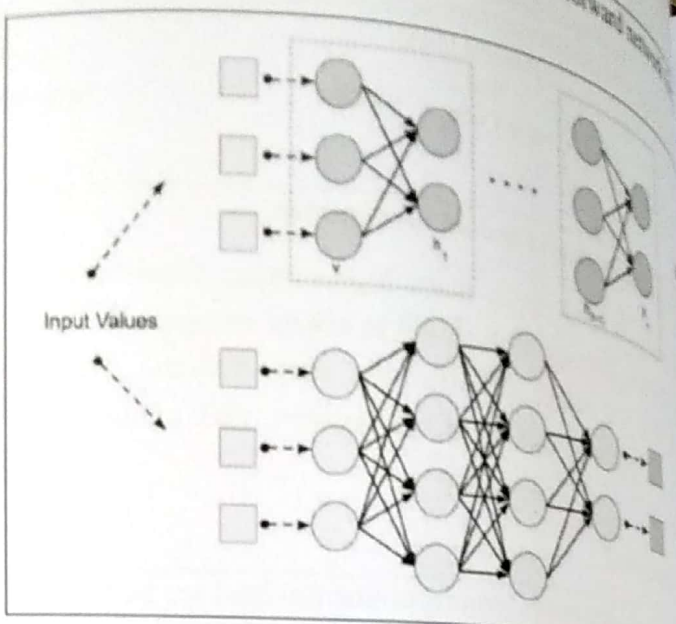


Figure 4-1. DBN architecture

Answer b.	LSTM Networks + properties. LSTM networks are the most commonly used variation of Networks. LSTM networks were introduced in 1997 by Schmidhuber. The critical component of the LSTM28 is the memory cell are modulated by the input gates and forget gates. Properties of LSTM networks LSTMs are known for the following: • Better update equations • Better backpropagation Here are some example use cases of LSTMs: • Generating sentences (e.g., character-level language models) • Classifying time-series • Speech recognition • Handwriting recognition • Polyphonic music modeling

RAMAIAH INSTITUTE OF TECHNOLOGY
(Autonomous Institute, affiliated to VTU)
DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

Term:	30.08.2018 to 29.12.2018	**Course Code:**	IS71D4
Course:	Deep Learning	**Semester:**	7th
CIE:	Test – III	**Max Marks:**	30
Date:		**Time:**	

Portions for Test: Lecture Nos. from 36 to 52 as per lesson plan. **Instructions to Candidates:** Answer any two questions. **Note:** Mobiles and Programmable Calculators are strictly prohibited.

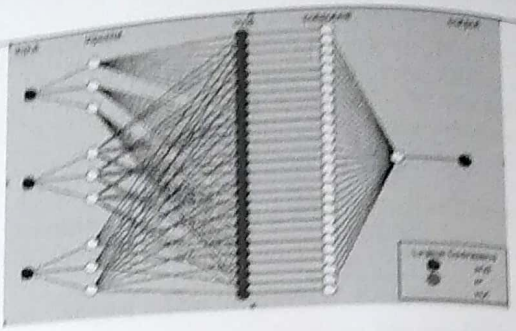
Q.#		Question	Marks	Bloom's Level #	Cos
1.	a.	Why Recurrent Neural Networks are good fit for Time-series and sequential data? Illustrate with the help of example.	7	U	CO 4
	b	Explain Vectorization. List out the Needs of data Vectorize.	8	U	CO 5
2.	a	With the help of block diagram explain Multilayer Perceptions.	7	R	CO 4
	b	Describe Classification Model Output Layer with the help of example.	8	R	CO 5
3.	a.	Discuss the step by step process in Building the Intuition.	7	U	CO 4
	b.	Illustrate Feed-Forward Multilayer Neural Networks with the help of block diagram.	8	R	CO 5

#R – Remember; U – Understand;

RAMAIAH INSTITUTE OF TECHNOLOGY
(Autonomous Institute, affiliated to VTU)
DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING
Answer scheme
Test 3

Course: Deep Learning **Course Code: IS71D4** **Semester: 7th**

Answer a.	Time-series and sequential data represent a waveform when graphed on a two dimensional plot. Often, we are looking for areas of the graph that make particular but is arranged in a two-dimensional grid to form a picture. However, in both time series and images, we are looking for specific objects that might occur in the data. These objects might be scaled in size and usually do not appear in the same place every time in the data, providing us with a challenge. Recurrent Neural Networks have evolved from multilayer perceptrons to better model the time domain for time-series data. Recurrent Neural Networks are better able to model the time domain by allowing a sequence of input vectors to be treated as a single logical input for a Recurrent Neural Network model. Example:	4 3
Answer b	**Vectorization:** Vectorization techniques (along with the data-handling process itself) is core to the process of data science and many times grossly overlooked The vectorization phase of machine learning can last from hours to days, depending on your comfort level with programming and vectorization. This impedance factor tends to slow down many new users beginning work with statistical models **Needs of data Vectorize:** In the course of working in machine learning and data science, we need to analyze all types of data. A key requirement is being able to take each data type and represent it as a numerical vector (or in some cases, a multidimensional array of numbers). Neural networks still need to represent the input data as vectors and matrices because they cannot work directly on text, graph, and other non-vector /matrix representations.	4 4
Answer a	A multilayer perceptron (MLP) is a class of feedforward artificial neural network. An MLP consists of, at least, three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training.	



Explanation + diagram

	Answer b	Describe Classification Model Output Layer with the help of example. In classification models, we have N number of output units in the output lay... we produce a class score for each output unit. If N = 1, then we have a model... single label. In this case, we'd be classifying the presence of a condition or ... condition (e.g., spam versus not spam). If N > 1 then we're scoring the input ... respect to each of the classes, and we use a different output layer configurati... this variation, we might be classifying a document as to what category—sport... business, politics it belongs to. There is also the case for which the documen... belong to multiple categories—for example, sports and business.
3.	Answer a.	step by step process in Building the Intuition: 1. Determine what our input data should be: a. The input data type informs us as to architecture 2. Determine what our intended result should be: a. Gives us guidance on configuring architecture b. Determines the output layer type 3. Set up architecture of network to support problem: a. Choice of model, architecture, and cost function are all important b. Depending on the architecture, we'll select a number of hidden layers c. Choose activations per layer based on the overall architecture of the netw... and the intent of the specific layer 4. Work with training data to handle the following tasks: a. Clean data b. Produce visualizations c. Perform vectorization and normalization d. Balance classes (as necessary) e. Create test, train, and validate splits 5. Develop a hyperparameter tuning strategy with a balanced subset of the da... a. Increase the subset data size and tweak hyperparameters as needed 6. If the final training dataset is large, use Spark to train on more data (if ... appropriate).

Answer b. Feed-Forward M... feed-forward mu... have the same ... simply be equal... regression. In t... and neuron

x_1

x_2

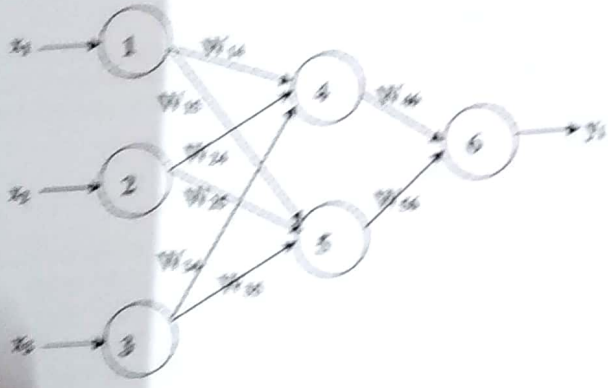
x_3

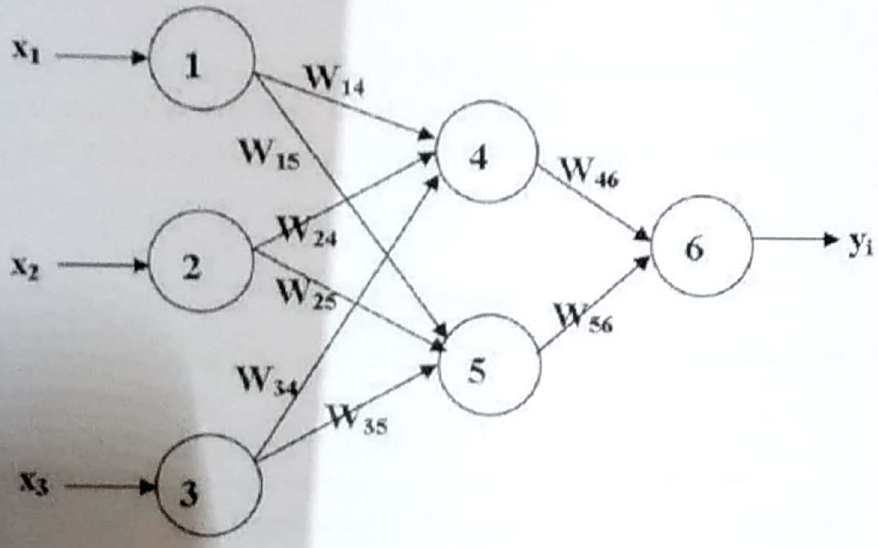
5+3

Answer is.

Feed-Forward Multilayer Neural Networks: explanation + diagram

feed-forward multilayer perceptron neural networks, input layers are required to have the same number of input units as the input vector. The output layer will simply be equal to the number of labels for classification and a single neuron for regression. In the sections that follow, we discuss strategies for determining layer and neuron



wer	Feed-Forward Multilayer Neural Networks: explanation + diagram feed-forward multilayer perceptron neural networks, input layers are required to have the same number of input units as the input vector. The output layer will simply be equal to the number of labels for classification and a single neuron for regression. In the sections that follow, we discuss strategies for determining layer and neuron  	5+3

RAMAIAH INSTITUTE OF TECHNOLOGY
(Autonomous Institute, affiliated to VTU)
DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

Term:	30.08.2018 to 29.12.2018	**Course Code:**	IS71E1
Course:	Data Science	**Semester:**	VII
CIE:	Test-I	**Max Marks:**	30
Date:	05.10.2018	**Time:**	2.30PM -3.30PM

Portions for Test:Lecture Nos from 17 to 36 as per lesson plan
Instructions to Candidates:Answer any two question
Note:Mobiles and programmable Calculators are strictle prohibited

Sl.#	QUESTION	Marks	Bloom's Level#	COs
1	a) Illustrate the importance of PMF and write a pseudo code to derive them.	8	U	CO1
	b) Walk through the Chi Square Test with an example .	7	A	CO2
2	a) Discuss the importance of Central Limit Theorem	8	A	CO2
	b) Explain the popularity of Normal distribution in real world	7	U	CO1
3	a) State and analyse the implications of PMF of Binomial distribution and its coefficient	8	An	CO1
	b) Hypothesis testing and the meaning of Type I and Type II errors	7	A	CO2
	#U-Understand; A-Apply; An-Analyze			

EST1 -SCHEME & SOLUTIONS: DATA SCIENCE(IS71E1)

erm:	30.08.2018 to 29.12.2018	Course Code:	IS71E1
ourse:	Data Science	Semester:	VII
IE:	Test-I	Max Marks:	30
ate:	05.10.2018	Time:	2.30PM -3.30PM

.no	QUESTION	MARKS
1a)	Illustrate the importance of PMF and write a pseudo code to derive them.	8

	Solution:
	The probability mass function or PMF is a function that maps observations to probabilities. It is a function
	that gives the probability that a discrete random variable is exactly equal to some value. It is is often the
	primary means of defining a discrete probability distribution, and such functions exist for either scalar or
	multivariate random variables whose domain is discrete. They are also very useful for exploratory data
	analysis as they can help reveal patterns, differences and other features that can be otherwise non-
	obvious.
	We can obtain a PMF for a discrite sample, by dividing the frequency of a value by the size of the sample
	- i.e the normalized frequencies of the sample.
	The following psuedocode illustrates this:
	In [6]: #let sample be the list of observations/values
	sample = [2,3, 4,4,4,4,4,5,2,10]
	freq = {} # freq is a dictionary/map
	for x in sample:
	freq[x] = freq.get(x,0)+1
	n = float(len(sample))
	pmf = {}
	for key in freq:
	pmf[key] = freq[key]/n
	print pmf
	# PMF is the resulting probability mass function for sample.
	{10: 0.1, 3: 0.1, 4: 0.5, 2: 0.2, 5: 0.1}	

b)	Walk through the Chi Square Test with an example	7
	Solution: Let us take an example of rolling a 6 sided Die, 60 times. If a table of outcomes with frequencies in this experiment is given, and we need to check if the die is biased or not, we can use the Chi-Square test to test the significance of getting the given frequencies, and if the	

Term:	30.08.2018 to 29.12.2018	**Course Code:**	
Course:	Data Science	**Semester:**	IS71E1
CIE:	Test-I	**Max Marks:**	VII
Date:	05.10.2018	**Time:**	30
			2.30PM-3.30PM

conclusion is that the effect (seeing the given frequencies) is statistically significant, then the
we can conclude that the Hypothesis is true, and the Die is biased.Say the outcome frequencies are:
Outcome 1 2 3 4 5 6
Frequency 8 9 19 6 8 10
The sum of frequencies given add up to 60.
We make a null hypothesis that the die is fair.
We calculate the Expected Frequency from a theoretical standpoint = 1/6 * 60 = 10
If E is the expected frequency of an outcome and O is the observed frequency, compute
the Chi-Square statistic, by calculating
for each observation, and sum the result to get
Outcome 1 2 3 4 5 6
Frequency 8 9 19 6 8 10
(E-O) 2 /2 0.4 0.1 8.1 1.6 0.4 0.0
Which is approximately Chi-Square distributed with k-1 degrees of freedom (where k is the number
of categories/outcomes). Thus we get the chi-square test statistic for this problem as as $\chi 2$ =
10.6 which has 5 degrees of freedom. Reading the value from a Chi-Square table for 10.6, at
5 degrees of freedom, we get the P-value to be 0.05 Thus the null hypothesis that the dice is
fair is false, and we conclude that the die is biased.

2a)	Discuss the importance of Central Limit Theorem
	Solution: Normal distributions are closed under linear transformations, but we add values drawn from two normal distributions, the resulting values will be normally distributed. However this property does not necessarily hold for other distributions. The Central Limit Theorem, states that if we add up a number of values drawn from almost any distribution, the resulting values converge to a normal distribution. Specifically if the mean and variance of the distribution of values is μ and σ, and the values are drawn independently from a distribution with a finite mean and variance, then the distribution of the sum of values is approximately N(nμ,nσ). The Central Limit Theorem explains, at least in part, the prevalence of normal distributions in the natural world. Most characteristics of animals and other

Term:	30.08.2018 to 29.12.2018	**Course Code:**	18ITE1
Course:	Data Science	**Semester:**	VII
CIE:	Test-I	**Max Marks:**	30
Date:	05.10.2018	**Time:**	2.30PM -3.30PM

	forms are affected by a large number of genetic and environmental factors whose effect is additive. The characteristics we measure are the sum of a large number of small effects, so their distribution tends to be normal. The theorem is a key concept in probability theory because it implies that probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions. Even when we don't know the original distribution from which the data was drawn, assuming it is a normal distribution will still help find useful insights from the data.	
b)	Explain the popularity of Normal distribution in real world	7
	Solution: This is mainly because of the Central Limit Theorem. In probability theory, the central limit theorem (CLT) establishes that, in most situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution (informally a "bell curve") even if the original variables themselves are not normally distributed. The theorem is a key concept in probability theory because it implies that probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions. It turns out that if we add up a large number of values from almost any distribution, the distribution of the sum converges to normal. More specifically, if the distribution of the values has mean and standard deviation μ and σ, the distribution of the sum is approximately $N(n\mu, n\sigma^2)$ provided that: The values have been drawn independently. The values have to come from the same distribution (although this requirement can be relaxed) The values have to be drawn from a distribution with finite mean and variance, so most Pareto distributions are out The number of values you need before you see convergence depends on the skewness of the distribution. Sums from an exponential distribution converge for small sample sizes. Sums from a lognormal distribution do not. This explains the prevalence of normal distributions in the natural world. Most characteristics of animals and other life forms are affected by a large number of genetic and environmental factors whose effect is additive. The characteristics we measure are the sum of a large number of small effects, so their distribution tends to be normal	
3a)	State and analyse the implications of PMF of Binomial distribution and its coefficient.	8
	The PMF of Binomial distribution is given by PMF(k) = $n \cdot p^k (1-p)^{n-k}$	

Term:	30.08.2018 to 29.12.2018	Course Code:	IS71E1
Course:	Data Science	Semester:	VII
CIE:	Test-I	Max Marks:	30
Date:	05.10.2018	Time:	2.30PM-3.30PM

(k)
The binomial coefficient is pronounced "n choose k", and it can be compu
directly like this:
n
n!=(k) k!(n − k)!
or recursively:
n
n − 1
n − 1
=
+
(k) (k) (k − 1)
The implications of the PMF of a binomial distribution is that if the probabi
an event occuring in the
distribution is given by p, and the probability of the event not occuring is giv
by q, then the probability
of the event occuring k times for all values of k ∈ [0, n] is given by the PMF
the binomial distribution.
Where n is the number of trials.
The value of the coefficient is also the same as the number of ways to select k
items from a set of n
items. Which is an important measure in combinatorics and computer science.
also gives the
expansion of (x + y) n , the pascal triangle, finding the number of paths in a gr
etc.

3 b) Hypothesis testing and the meaning of Type I and Type II errors

Hypothesis testing is the process of determining whether an observed effect is
statistically
significant or not. The underlying logic is similar to a proof by contradiction. T
prove a
mathematical statement, A, we assume temporarily that A is false. If that
assumption leads to
a contradiction, we conclude that A must actually be true. Similarly, we say that
an effect is
not significant. This is called the null hypothesis , and based on that assumption
we calculate
the probability of the apparent effect, known as the p-value , if the p-value is low
enough, we

Term:	30.08.2018 to 29.12.2018	Course Code:	IS71E1
Course:	Data Science	Semester:	VII
CIE:	Test-I	Max Marks:	30
Date:	05.10.2018	Time:	2.30PM -3.30PM

conclude that the null hypothesis is unlikely to be true.
In hypothesis testing, there are two kinds of errors that occur, Type I errors and Type II
errors.

- **Type I errors** are also called false positives , which is when we conclude that a hypothesis is true (accept), while in reality it is false. In other words we consider an effect to be significant, when it was actually due to chance.
- **Type II errors** are also called false negative , which is when we conclude that a hypothesis is false (reject), while in reality it is true. In other words we consider an effect to be due to chance, when it was actually significant.

RAMAIAH INSTITUTE OF TECHNOLOGY
(Autonomous Institute, affiliated to VTU)
DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

Term:	30.08.2018 to 29.12.2018	Course Code:	IS71E1
Course:	Data Science	Semester:	VII
CIE:	Test-II	Max Marks:	30
Date:	22.11.2018	Time:	2.30PM -3.30PM

Portions for Test:Lecture Nos from 17 to 36 as per lesson plan
Instructions to Candidates:Answer any two question
Note:Mobiles and programmable Calculators are strictle prohibited

Sl.#	QUESTION	Marks	Bloom's Level#	COs
1a)	What is the biggest strength of linear regression model? What is its weakness?	8	U	CO3
b)	PCA is used mainly for noise reduction in dataset.Justify.	7	A	CO4
2a)	Write the pseuodocode for hierarchical clustering algorithm	8	A	CO4
b)	Why do Naive bayes classifiers perform better than linear models?	7	U	CO3
3a)	Discuss the relative importance of random forests over decision trees.	8	An	CO3
3 b)	What are the parameters that can be used to compare clustering algorithms?	7	A	CO4
	#U-Understand; A-Apply; An-Analyze			

Term:	30.08.2018 to 29.12.2018	**Course Code:**	IS71E1
Course:	Data Science	**Semester:**	VII
CIE:	Test-II scheme solutions	**Max Marks:**	30
Date:	22.11.2018	**Time:**	2.30PM -3.30PM

Q.no	QUESTION	MARKS
1a)	What is the biggest strength of linear regression model? What is its weakness?	8

Solution:

Linear regression, or ordinary least squares (OLS), is the simplest and most classic linear method for regression.

General formula: $\hat{y} = w[0] * x[0] + w[1] * x[1] + \ldots + w[p] * x[p] + b$

Linear regression finds the parameters w and b that minimize the mean squared error between predictions and the true regression targets, y, on the training set. The mean squared error is the sum of the squared differences between the predictions and the true values.

Strengths: - Simple model, easy for beginners to understand - No parameters to tune - Useful as many problems can be transformed into linear regression problems

Weaknesses: -

Overfitting: Linear regression has no parameters, which is a benefit, but it also has no way to control model complexity. For a one-dimensional dataset, there is little danger of overfitting, as the model is very simple (or restricted). However, with higher-dimensional datasets, linear models become more powerful, and there is a higher chance of overfitting. - Linear Regression Is Limited to Linear Relationships: By its nature, linear regression only looks at linear relationships between dependent and independent variables. That is, it assumes there is a straight-line relationship between them which may not always be true. - Sensitive to outliers:

b)	PCA is used mainly for noise reduction in dataset.Justify.	7

Solution:

Principal Component Analysis (PCA) is used to a) denoise and to b) reduce dimensionality.

It does not eliminate noise, but it can reduce noise.

Basically an orthogonal linear transformation is used to find a projection of all data into k dimensions, whereas these k dimensions are those of the highest variance. The eigenvectors of the covariance matrix (of the dataset) are the target dimensions and they can be ranked according to their eigenvalues. A high eigenvalue signifies high variance explained by the associated eigenvector dimension.

Lets take a look at the usps dataset, obtained by scanning handwritten digits from envelopes by the U.S. Postal Service.

First, we compute the eigenvectors and eigenvalues of the covariance matrix and plot all eigenvalues descending. We can see that there a few eigenvalues which could be named principal components, since their

	eigenvalues are much higher than the rest.
2a)	Write the pseudocode for hierarchical clustering algorithm **Solution:** Given a set of N items to be clustered, and an N*N distance (or similarity) matrix, the basic process of hierarchical clustering is this: 1. Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain. 2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less. 3. Compute distances (similarities) between the new cluster and each of the old clusters. 4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. (*)
b)	Why do Naive bayes classifiers perform better than linear models? **Solution:** Naive Bayes classifiers tend to be faster in training compared to linear models. The reason that naive Bayes models are so efficient is that they learn parameters by looking at each feature individually and collect simple per-class statistics from each feature. The models work very well with high-dimensional sparse data and are relatively robust to the parameters. Naive Bayes models are great baseline models and are often used on very large datasets, where training even a linear model might take too long. The price paid for this efficiency is that naive Bayes models often provide generalization performance that is slightly worse than that of linear classifiers like LogisticRegression and LinearSVC.
3a)	Discuss the relative importance of random forests over decision trees.
	Solution: A main drawback of decision trees is that they tend to overfit the training data. Random forests are one way to address this problem. A random forest is essentially a collection of decision trees, where each tree is slightly different from the others. The idea behind random forests is that each tree might do a relatively good job of predicting, but will likely overfit on part of the data. If we build many trees, all of which work well and overfit in different ways, we can reduce the amount of overfitting by averaging their results. This reduction in overfitting, while retaining the predictive power of the trees, can be shown using rigorous mathematics. To implement this strategy, we need to build many decision trees. Each tree should do an acceptable job of predicting the target, and should also be different from the other trees. There are two ways in which the trees in a random forest are randomized: by selecting the data points used to build a tree and by selecting the features in each split test. Random forest gives nonzero importance to many more features than the single tree. The randomness in building the random forest forces the algorithm to consider many possible explanations, the

	result being that the random forest captures a much broader picture of the data than a single tree. Random forests for regression and classification are currently among the most widely used machine learning methods. They are very powerful, often work well without heavy tuning of the parameters, and don't require scaling of the data. Random forests usually work well even on very large datasets, and training can easily be parallelized over many CPU cores within a powerful computer	
3 b)	What are the parameters that can be used to compare clustering algorithms?	7
	Solution: There are metrics that can be used to assess the outcome of a clustering algorithm relative to a ground truth clustering, the most important ones being the adjusted rand index (ARI) and normalized mutual information (NMI), which both provide a quantitative measure between 0 and 1. The adjusted rand index provides intuitive results, with a random cluster assignment having a score of 0 and DBSCAN (which recovers the desired clustering perfectly) having a score of • A common mistake when evaluating clustering in this way is to use accuracy_score instead of adjusted_rand_score, normalized_mutual_info_score, or some other clustering metric. The problem in using accuracy is that it requires the assigned cluster labels to exactly match the ground truth. • Using metrics like ARI and NMI usually only helps in developing algorithms, not in assessing success in an application. There are scoring metrics for clustering that don't require ground truth, like the silhouette coefficient. However, these often don't work well in practice. • The silhouette score computes the compactness of a cluster, where higher is better, with a perfect score of 1. While compact clusters are good, compactness doesn't allow for complex shapes. • A slightly better strategy for evaluating clusters is using robustness-based clustering metrics. These run an algorithm after adding some noise to the data, or using different parameter settings, and compare the outcomes. The idea is that if many algorithm parameters and many perturbations of the data return the same result, it is likely to be trustworthy. Unfortunately, this strategy is not implemented in scikit-learn at the time of writing.	

RAMAIAH INSTITUTE OF TECHNOLOGY
(Autonomous Institute, affiliated to VTU)
DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

Term:	30.08.2018 to 29.12.2018	Course Code:	IS71E1
Course:	Data Science	Semester:	VII
CIE:	Test-III	Max Marks:	30
Date:	22.12.2018	Time:	12:00 to 1:00PM

Portions for Test: Lecture Nos from 17 to 36 as per lesson plan
Instructions to Candidates: Answer any two question
Note: Mobiles and programmable Calculators are strictly prohibited

Sl.#	QUESTION	Marks	Bloom's Level#	COs
1	a)Consider a family which has three children with the sample space S={BBB,BBG,BGB,GBB,GGG,GGB,GBG,BGG} where B= boy and G= girl and suppose the probability of having a boy is the same as the probability of having a girl. Consider the random variable X to be the number of boys and find the PMF and CDF.	8	An	CO2
	b) A company conducted a test for effectiveness of Vitamin C tablets in controlling cold and here are the results. Use Chi Squared test to verify the result of the experiment with P=0.05.		A	CO2

	Vitamin C tablet taken	Vitamin C tablet not taken
Got Cold	15	22
Did not get cold	100	93

Sl.#	QUESTION	Marks	Bloom's Level#	COs
2	a) Consider the following training set for a text classifier. Using Naive Bayes predict the tag of the sentence "A very close game".	8	A	CO3

Text	Tag
"A great game"	Sports
"The election was over"	Not sports
"Very clean match"	Sports
"A clean but forgettable game"	Sports
"It was a close election"	Not sports

	b) Explain the k-NN algorithm with sample code	7	U	CO3
3	a) Compare various unsupervised learning techniques and bring out their pros and cons.	8	U	CO4
	b) Use Agglomerative clustering to group the data described by the following distance matrix		A	CO4

	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

RAMAIAH INSTITUTE OF TECHNOLOGY
(Autonomous Institute, affiliated to VTU)
DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

Term:	30.08.2018 to 29.12.2018	Course Code:	IS71E1
Course:	Data Science	Semester:	VII
CIE:	Test-III(SCHEME & SOLUTION)	Max Marks:	30
Date:	22.12.2018	Time:	12:00 to 1:00PM

Portions for Test: Lecture Nos from 17 to 36 as per lesson plan
Instructions to Candidates: Answer any two question

Sl.#	QUESTION	Marks	Bloom's Level#	COs
1	a)Consider a family which has three children with the sample space S={BBB,BBG,BGB,GBB,GGG,GGB,GBG,BGG} where B= boy and G= girl and suppose the probability of having a boy is the same as the probability of having a girl. Consider the random variable X to be the number of boys and find the PMF and CDF. suppose the probability of having a boy is the same as the probability of having a girl. Let the random variable XX be the number of boys. Then XX will have the following pmf: t: 0 | 1 | 2 | 3 P(X=t): 1/8 | 3/8 | 3/8 | 1/8 Then, we can use the pmf to find the cdf. t: 0 | 1 | 2 | 3 FX(t)=P(X≤t): 1/8 | 1/8+3/8=4/8 | 4/8+3/8=7/8 | 7/8+1/8=1 **SCHEME: 4 MARKS FOR PMF(FOR EACH ENTRY 1 MARK) ,4 MARKS FOR CDF(FOR EACH ENTRY 1 MARK)**	8	An	CO2
	b)Explain the k-NN algorithm with sample code **SOLUTION:** The k-NN algorithm is arguably the simplest machine learning algorithm. Building the model consists only of storing the training dataset. **Algorithm:** Step 1: Determine parameter K = number of nearest neighbours Step 2: Calculate the distance between new data point and all the training examples Step 3: Sort the distances and determine the K nearest neighbours based on minimum distance. Step 4: Gather the category Y of the K nearest neighbours Step 5: Use simple majority of the category of K nearest neighbors	7	U	CO3

as the prediction value
for the new data point.

Sample Code:

Split data into training and test set so as to evaluate generalization performance: from sklearn.from sklearn.datasets import load_breast_cancer from sklearn.model_selection import train_test_split

```
cancer = load_breast_cancer()
X_train, X_test, y_train, y_test =
train_test_split(cancer.data, cancer.target,
stratify=cancer.target, random_state=0)
```

Import and instantiate the classifier. Set parameters i.e k = 3 (number of neighbours to use)

```
from sklearn.neighbors import KNeighborsClassifier
clf = KNeighborsClassifier(n_neighbors=3)
```

Fit the classifier using the training set. For KNeighborsClassifier this means storing the dataset, so as to compute neighbors during prediction:

```
clf.fit(X_train, y_train)
```

To make predictions on the test data, the predict method is called. For each data point in the test set, this computes its nearest neighbors in the training set and finds the most common class among these.

```
print("Test set predictions: {}".format(clf.predict(X_test)))
```

To evaluate how well the model generalizes, the score method is called with the test data together with the test labels:

```
print("Test set accuracy: {:.2f}".format(clf.score(X_test,
y_test)))
```

If the model is say about 86% accurate, it means the model predicted the class correctly for 86% of the samples in the test dataset.

2	a)Consider the following training set for a text classifier. Using Naive Bayes predict the tag of the sentence "A very close game".	8

Text	Tag
"A great game"	Sports
"The election was over"	Not sports
"Very clean match"	Sports
"A clean but forgettable game"	Sports
"It was a close election"	Not sports

SOLUTION:

WORD	P(WORD|SPORTS)	P(WORD|NOT_SPORTS)
A	(2+1)/(11+14)	(1+1)/(9+14)
VERY	(1+1)/(11+14)	(0+1)/(9+14)
CLOSE	(0+1)/(11+14)	(1+1)/(9+14)
GAME	(2+1)/(11+14)	(0+1)/(9+14)

P(A|SPORTS) * P(VERY|SPORTS) * P(CLOSE|SPORTS) *P(GAME|SPORTS)*P(SPORTS)=0.0000276

P(A|NOT SPORTS) * P(VERY|NOT SPORTS) * P(CLOSE|NOT SPORTS) *P(GAME|NOT SPORTS)*P(NOT SPORTS)=0.0000572

Excellent! Our classifier gives "A very close game" the **Sports** tag

SCHEME:

4 MARKS FOR TABLE ENTRIES,2 MARKS FOR SPORTS,2 MARKS FOR NOT SPORTS

b)How to interpret the principal components?.	7	U	CO4

SOLUTION:

Step 1: Determine the number of principal components

Determine the minimum number of principal components that account for most of the variation in your data, by using the following methods.

Proportion of variance that the components explain

Use the cumulative proportion to determine the amount of variance that the principal components explain. Retain the principal components that explain an acceptable level of variance. The acceptable level depends on your application. For descriptive purposes, you may only need 80% of the variance explained. However, if you want to perform other analyses on the data, you may want to have at least 90% of the variance explained by the principal components.

Eigenvalues

You can use the size of the eigenvalue to determine the number of principal components. Retain the principal components with the largest eigenvalues. For example, using the Kaiser criterion, you use only the principal components with eigenvalues that are greater than 1.

Scree plot

The scree plot orders the eigenvalues from largest to smallest. The ideal pattern is a steep curve, followed by a bend, and then a straight line. Use the components in the steep curve before the first point that starts the line trend.

Principal Component Analysis: Income, Education, Age, Residence, Employ, ...

Eigenanalysis of the Correlation Matrix

Eigenvalue 3.5476 2.1320 1.0447 0.5315 0.4112 0.1665 0.1254

0.0411
Proportion 0.443 0.266 0.131 0.066 0.051 0.021 0.016 0.005
Cumulative 0.443 0.710 0.841 0.907 0.958 0.979 0.995 1.000
Eigenvectors
Variable PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8
Income 0.314 0.145 -0.676 -0.347 -0.241 0.494 0.018 -0.030
Education 0.237 0.444 -0.401 0.240 0.622 -0.357 0.103 0.057
Age 0.484 -0.135 -0.004 -0.212 -0.175 -0.487 -0.657 -0.052
Residence 0.466 -0.277 0.091 0.116 -0.035 -0.085 0.487 -0.662
Employ 0.459 -0.304 0.122 -0.017 -0.014 -0.023 0.368 0.739
Savings 0.404 0.219 0.366 0.436 0.143 0.568 -0.348 -0.017
Debt -0.067 -0.585 -0.078 -0.281 0.681 0.245 -0.196 -0.075
Credit cards -0.123 -0.452 -0.468 0.703 -0.195 -0.022 -0.158 0.058

Step 2: Interpret each principal component in terms of the original variables

To interpret each principal components, examine the magnitude and direction of the coefficients for the original variables. The larger the absolute value of the coefficient, the more important the corresponding variable is in calculating the component. How large the absolute value of a coefficient has to be in order to deem it important is subjective. Use your specialized knowledge to determine at what level the correlation value is important.

Principal Component Analysis: Income, Education, Age, Residence, Employ, ...

Eigenanalysis of the Correlation Matrix

Eigenvalue 3.5476 2.1320 1.0447 0.5315 0.4112 0.1665 0.1254 0.0411
Proportion 0.443 0.266 0.131 0.066 0.051 0.021 0.016 0.005
Cumulative 0.443 0.710 0.841 0.907 0.958 0.979 0.995 1.000
Eigenvectors
Variable PC1 PC2 PC3 PC4 PC5 PC6 PC7 PC8
Income 0.314 0.145 -0.676 -0.347 -0.241 0.494 0.018 -0.030
Education 0.237 0.444 -0.401 0.240 0.622 -0.357 0.103 0.057
Age 0.484 -0.135 -0.004 -0.212 -0.175 -0.487 -0.657 -0.052
Residence 0.466 -0.277 0.091 0.116 -0.035 -0.085 0.487 -0.662
Employ 0.459 -0.304 0.122 -0.017 -0.014 -0.023 0.368 0.739
Savings 0.404 0.219 0.366 0.436 0.143 0.568 -0.348 -0.017
Debt -0.067 -0.585 -0.078 -0.281 0.681 0.245 -0.196 -0.075
Credit cards -0.123 -0.452 -0.468 0.703 -0.195 -0.022 -0.158 0.058

Key Results: PC, Loading plot

In these results, first principal component has large positive associations with Age, Residence, Employ, and Savings, so this component primarily measures long-term financial stability. The second component has large negative associations with Debt and Credit cards, so this component primarily measures an applicant's credit history. The third component has large negative associations with income, education, and credit cards, so this component primarily measures the applicant's academic and income qualifications.

The loading plot visually shows the results for the first two

components. Age, Residence, Employ, and Savings have large positive loadings on component 1, so this component measure long-term financial stability. Debt and Credit Cards have large negative loadings on component 2, so this component primarily measures an applicant's credit history.

Step 3: Identify outliers

Use the outlier plot to identify outliers. Any point that is above the reference line is an outlier. Outliers can significantly affect the results of your analysis. Therefore, if you identify an outlier in your data, you should examine the observation to understand why it is unusual. Correct any measurement or data entry errors. Consider removing data that are associated with special causes and repeating the analysis.

SCHEME:EACH STEP CARRIES 2 MARKS

3	a) A company conducted a test for effectiveness of Vitamin C tablets in controlling cold and here are the results. Use Chi Squared test to verify the result of the experiment with P=0.05.	7	A	CO2

	Vitamin C tablet taken	Vitamin C tablet not taken
Got Cold	15	22
Did not get cold	100	93

SOLUTION:

Choose statistical test. The chi-square test is used when: data are nominal (in this case people either got a cold, or did not get a cold) the observed number in each category can be compared to an expected number none of the expected frequencies is less than five. State the null hypothesis. NH0 : Vitamin C does not affect the chance of catching a cold. State the level of significance. P = 5 per cent (0.05). Calculate the expected results. Assuming the null hypothesis to be true, we would expect the proportion of people taking vitamin C tablets who caught a cold to be the same as the proportion of people not taking vitamin C tablets who caught a cold. Calculate the totals of the rows and columns in the results table

	Vitamin C tablet taken	**Vitamin C tablet not taken**	**Row total**
GOT A COLD	15	22	37
DID NOT GET A COLD	100	93	193
COLUMN TOTAL	115	115	230

Calculate the proportion of students who caught a cold. Proportion of students who got a cold = 37/230 = 0.16

Calculate the number of students expected to catch a cold and not catch a cold in each sample. We would expect 16 per cent of

students taking vitamin C to catch a cold: 18.4 students. We would expect 84 per cent of students taking vitamin C to not catch a cold: 96.6 students. We would expect 16 per cent of students not taking vitamin C to catch a cold: 18.4 students. We would expect 84 per cent of students not taking vitamin C to not catch a cold: 96.6 students.

CHI-SQUARE RESULT= 1.61

*Yates' correction applies for a 2 x 2 table. What are the degrees of freedom? DF = (number of rows – 1) x (number of columns – 1) DF = (2–1) x (2–1) DF = 1

Compare the calculated value with the critical value.

The critical value of chi-squared at 5% significance and 1 degree of freedom is 3.84. Our calculated value is 1.61. The calculated value is smaller than the critical value at the 5% level of probability.

b)Use Agglomerative clustering to group the data described by the following distance matrix

	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

SOLUTION:

Single link: distance between two clusters is the shortest distance between a pair of elements from the two clusters.

We apply the algorithm ,

At the beginning, each point A,B,C, and D is a cluster ! c1 = {A}, c2={B}, c3={C}, c4={D}

Iteration 1 The shortest distance is d(c1,c2)=1 !

c1 and c2 are merged ! the clusters are c3={C}, c4={D}, c5={A,B}

The distances from the new cluster to the others are d(c5,c3) = 2, d(c5,c4)=5

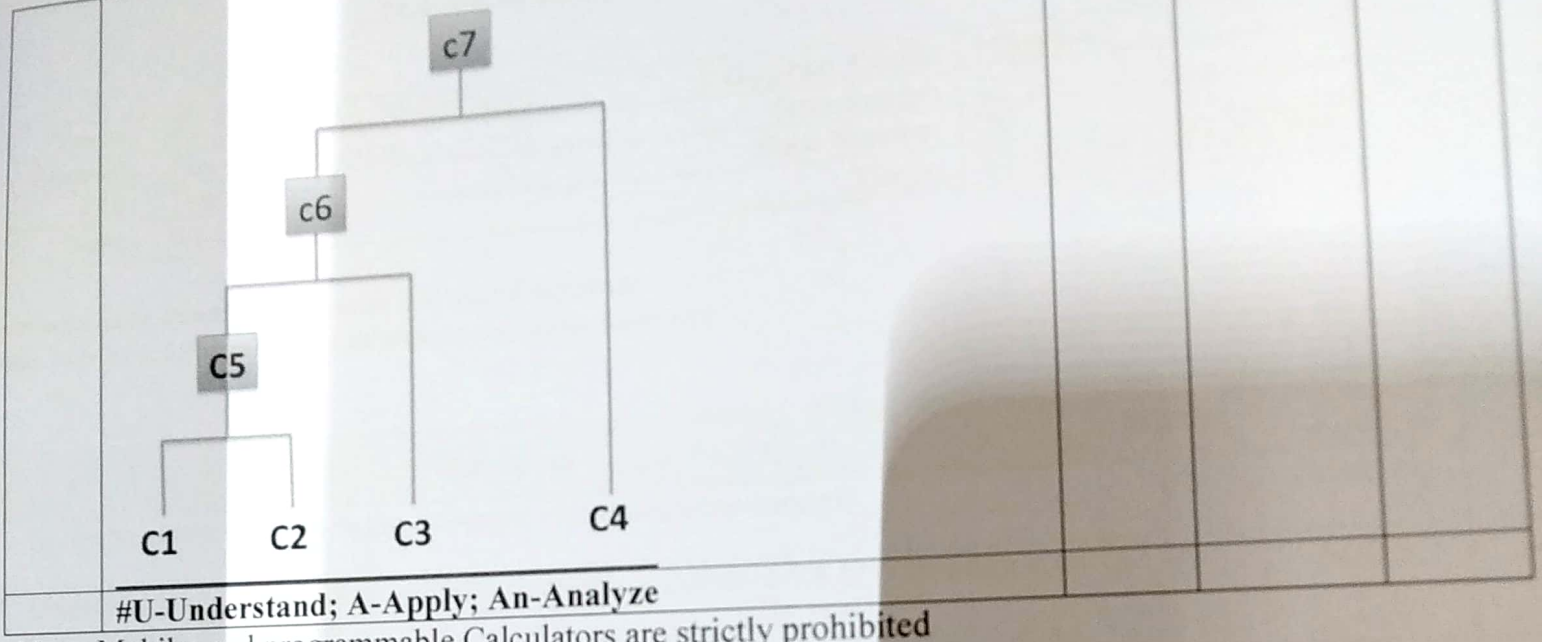
Iteration 2 The shortest distance is d(c5,c3)=2 ! c5 and c3 are merged ! the clusters are c6={A,B,C}, c4={D}

The distances from the new cluster to the others are:

d(c6,c4)=3

Iteration 3 c6 and c4 are merged ! the final cluster is c7={A,B,C,D}

RAMAIAH INSTITUTE OF TECHNOLOGY
(Autonomous Institute, affiliated to VTU)
DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING



#U-Understand; A-Apply; An-Analyze

Note: Mobiles and programmable Calculators are strictly prohibited

M S RAMAIAH INSTITUTE OF TECHNOLOGY
(Autonomous Institute, affiliated to VTU)
DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

M.Tech in Software Engineering

Term:	01-10-2018 to 30-01-2019	**Course Code:**	MSWEA1
Course:	Cloud Computing	**Semester:**	I
CIE:	Test – I	**Max Marks:**	30
Date:	04.12.2018	**Time:**	02:00 PM – 03:00 PM

Instructions to Candidates: Answer any two full questions.
Note: Mobiles and Programmable calculators are strictly prohibited.

Question	Marks	Bloom's Level #	COs
a) Give the comparison among different cloud delivery models.	05	Analyze	1
b) Illustrate the process that violates liveness requirements and deadlock in operation of a workflow.	05	Apply	2
c) Discuss the problems faced by virtualization of the x86 architecture.	05	Understand	3
a) Enumerate how RAID-5 configuration is applied to the cloud systems in handling cloud storage diversity and vendor lock-in.	05	Apply	1
b) Enumerate the organization of the GrepTheWeb application that uses MapReduce and AWS: EC2, Simple DB, S3 and SQS.	05	Apply	2
c)Explain the layering and interfaces between layers of a computer system.	05	Understand	3
a) Explain different services offered by Amazon Web Service and write the template for the creation of an EC2 instance.	05	Understand	1
b) Illustrate how zookeeper processes the read and write commands.	05	Apply	2
c)Distinguish between Full virtualization and para virtualization.	05	Analyze	3

M S RAMAIAH INSTITUTE OF TECHNOLOGY
(Autonomous Institute, affiliated to VTU)
DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

M.Tech in Software Engineering

Term:	01-10-2018 to 30-01-2019	**Course Code:**	MSWEA1
Course:	Cloud Computing	**Semester:**	I
CIE:	Test - I	**Max Marks:**	30
Date:	04.12.2018	**Time:**	02:00 PM - 03:00 PM

Instructions to Candidates: Answer any two full questions.
Note: Mobiles and Programmable calculators are strictly prohibited.

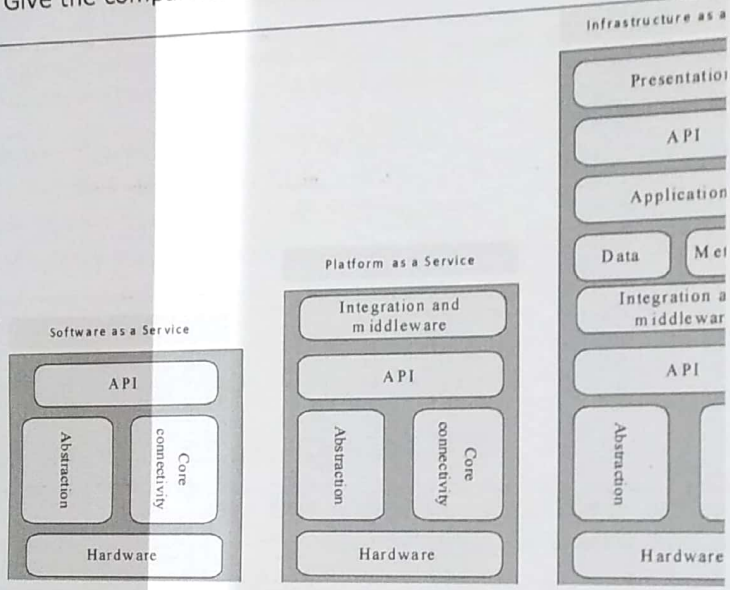
Question	Marks	Bloom's Level #	COs
a) Give the comparison among different cloud delivery models.	05	Analyze	1
b) Illustrate the process that violates liveness requirements and deadlock in operation of a workflow.	05	Apply	2
c) Discuss the problems faced by virtualization of the x86 architecture.	05	Understand	3
a) Enumerate how RAID-5 configuration is applied to the cloud systems in handling cloud storage diversity and vendor lock-in.	05	Apply	1
b) Enumerate the organization of the GrepTheWeb application that uses MapReduce and AWS: EC2, Simple DB, S3 and SQS.	05	Apply	2
c)Explain the layering and interfaces between layers of a computer system.	05	Understand	3
a) Explain different services offered by Amazon Web Service and write the template for the creation of an EC2 instance.	05	Understand	1
b) Illustrate how zookeeper processes the read and write commands.	05	Apply	2
c)Distinguish between Full virtualization and para virtualization.	05	Analyze	3

M S RAMAIAH INSTITUTE OF TECHNOLOGY
(Autonomous Institute, affiliated to VTU)
DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING
M.Tech in Software Engineering

Term:	01-10-2018 to 30-01-2019	**Course Code:**	MSWEA1
Course:	Cloud Computing	**Semester:**	I
CIE:	Test – I	**Max Marks:**	30
Date:	04.12.2018	**Time:**	02:00 PM – 03:00 PM

Instructions to Candidates: Answer any two full questions.
Note: Mobiles and Programmable calculators are strictly prohibited.

	Question	Marks	Bloom's Level #	COs
a)	Give the comparison among different cloud delivery models.	05		1



b)	Illustrate the process that violates liveness requirements and deadlock in operation of a workflow.	05		2

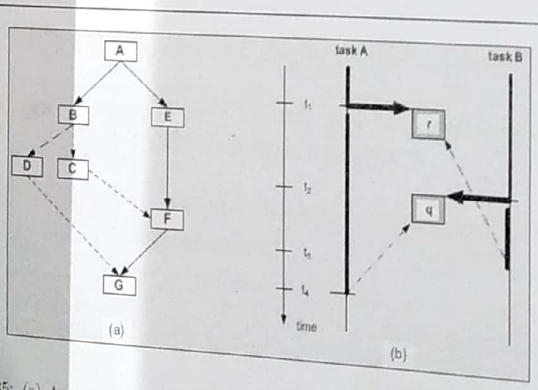


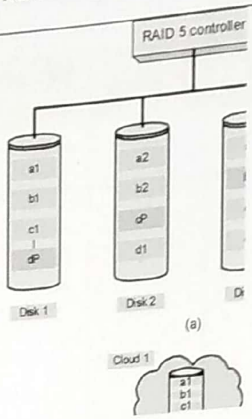
Figure 35: (a) A process description which violates the liveness requirement; if task C is chosen after completion of B, the process will terminate after executing task G; if D is chosen, then F will never be instantiated because it requires the completion of both C and E. The process will never terminate because G requires completion of both D and F. (b) Tasks A and B need exclusive access to two resources r and q and a deadlock may occur if the following sequence of events occur: at time t_1 task A acquires r, at time t_2 task B acquires q and continues to run; then, at time t_3, task B attempts to acquire r and it blocks because r is under the control of A; task A continues to run and at time t_4 attempts to acquire q and it blocks because q is under the control of B.

c) Discuss the problems faced by virtualization of the x86 architecture

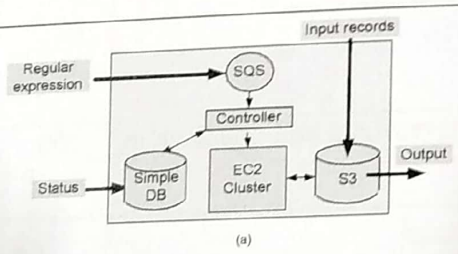
Problems faced by virtualization of the x86 architecture:

Ring de-privileging, Ring aliasing, Address space compression, Non-faulting access to privileged state, Interrupt virtualization, Access to hidden state, Ring compression.

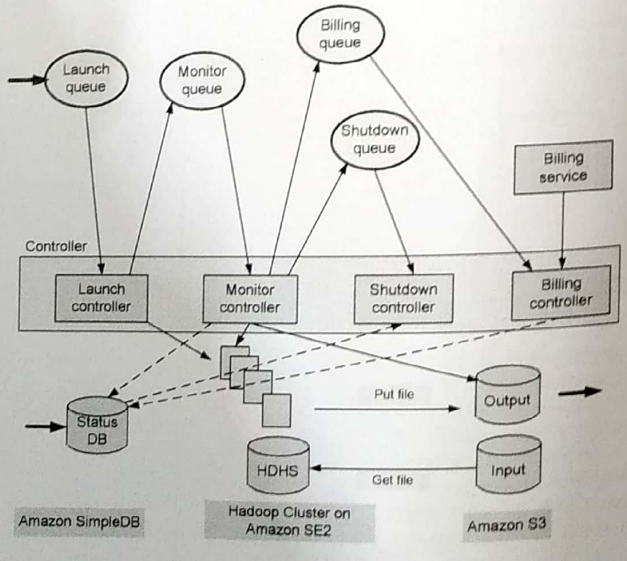
2. a) Enumerate how RAID-5 configuration is applied to the cloud systems in handling cloud storage diversity and vendor lock-in.



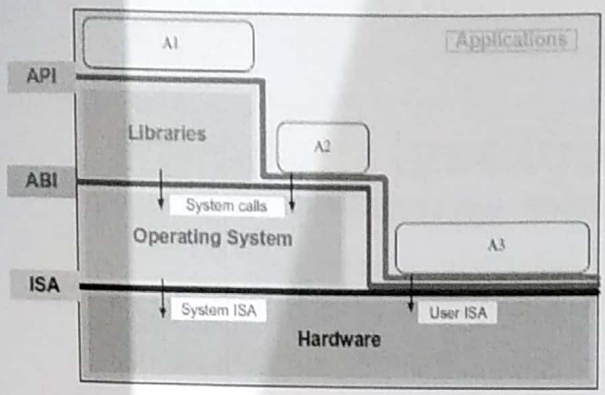
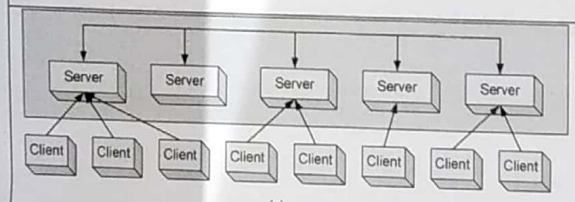
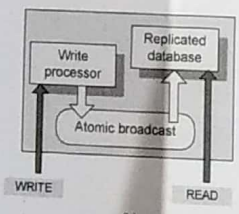
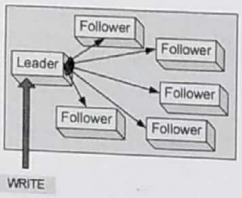
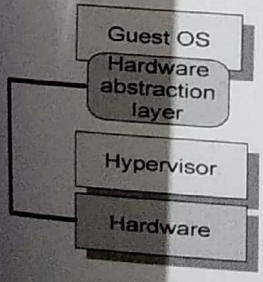
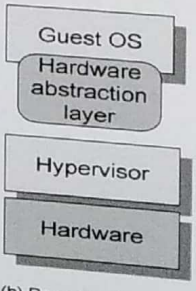
b) Enumerate the organization of the GrepTheWeb application that uses MapReduce Software and AWS: EC2, Simple DB, S3 and SQS. 05



(a)



(b)

c)Explain the layering and interfaces between layers of a computer system.	05		3
			
a) Explain different services offered by Amazon Web Service and write the template for the creation of an EC2 instance.	05		1
Services offered by Amazon Web Service: EC2, S3, EBS, Simple DB, SQS, Template for the creation of an EC2 instance.			
b) Illustrate how zookeeper processes the read and write commands	05		2
 (a)  (b)  (c)			
c) Distinguish between Full virtualization and para virtualization.	05		3
Full virtualization – a guest OS can run unchanged under the VMM as if it was running directly on the hardware platform. Paravirtualization - a guest operating system is modified to use only instructions that can be virtualized.  (a) Full virtualization  (b) Paravirtualization			