

Duration: 3hrs

[Max Marks:80]

- N.B. :** (1) Question No 1 is Compulsory.
 (2) Attempt any three questions out of the remaining five.
 (3) All questions carry equal marks.
 (4) Assume suitable data, if required and state it clearly.

Q1 Attempt any **four** from following. [20]

- A How to choose the right ML algorithm?
- B Explain Regression line, Scatter plot, Error in prediction and Best fitting line.
- C Explain the concept of feature selection and extraction.
- D Explain K-means algorithm.
- E Explain the concept of Logistic Regression

Q2 A Explain any five applications of Machine Learning. [10]

B Explain Multivariate Linear regression method. [10]

Q3 A Create a decision tree using Gini Index to classify following dataset for profit. [10]

| Age | Competition | Type | Profit |
|-----|-------------|----------|--------|
| old | Yes | software | down |
| old | No | software | Down |
| old | No | hardware | Down |
| mid | Yes | software | Down |
| mid | Yes | hardware | Down |
| mid | No | hardware | Up |
| mid | No | software | Up |
| new | Yes | software | Up |
| new | No | hardware | Up |
| new | no | software | Up |

B Find SVD for $A = \begin{bmatrix} 2 & 2 \\ -1 & 1 \end{bmatrix}$ [10]

Q4 A Explain the Random Forest algorithm in detail. [10]

B Explain the concept of bagging and boosting. [10]

Q5 A Describe Multiclass classification. [10]

B Explain the concept of Expectation Maximization Algorithm. [10]

Q6 Write detailed note on following. (**Any two**) [20]

- A Linear Regression
- B Linear Discriminant Analysis for Dimension Reduction
- C DBSCAN

Time: 3 Hours

Marks: 80

- Note: 1. Question 1 is compulsory
 2. Answer any three out of the remaining five questions.
 3. Assume any suitable data wherever required and justify the same.

- Q1** a) Distinguish between Name node and Data node. [5]
 b) List and explain the core business drivers behind the NoSQL movement. [5]
 c) Mention four characteristics of big data. Elaborate these characteristics with respect to social media websites. [5]
 d) List and explain the different issues and challenges in data stream query processing. [5]

- Q2** a) What is a key-value store? What are the benefits of using a key-value store? [10]
 b) Write a map reduce pseudo code to multiply two matrices. Apply map reduce working to perform following matrix multiplication. [10]
- $$\begin{matrix} 1 & 2 & 6 & 7 \\ & & X & \\ 3 & 4 & 8 & 9 \end{matrix}$$

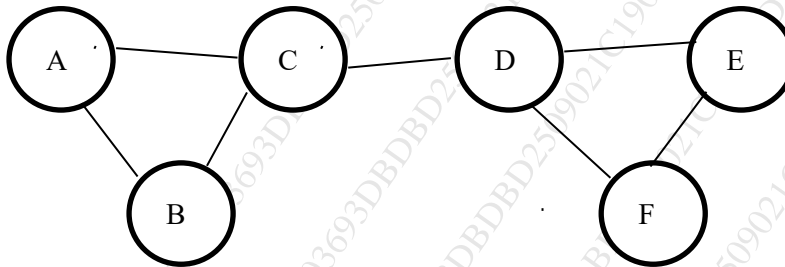
- Q3** a) Suppose the stream is $S = \{2, 1, 6, 1, 5, 9, 2, 3, 5\}$. Let hash functions $h(x) = ax + b \pmod{16}$ for some a and b , treat result as a 4-bit binary integer. Show how the Flajolet- Martin algorithm will estimate the number of distinct elements, $h(x) = 4x + 1 \pmod{16}$. [10]

- b) Consider the following data frame given below: [10]

| course | id | class | marks |
|--------|----|-------|-------|
| 1 | 11 | 1 | 56 |
| 2 | 12 | 2 | 75 |
| 3 | 13 | 1 | 48 |
| 4 | 14 | 2 | 69 |
| 5 | 15 | 1 | 84 |
| 6 | 16 | 2 | 53 |

- i. Create a subset of course less than 3 by using [] brackets and demonstrate the output.
 ii. Create a subset where the course column is less than 3 or the class equals to 2 by using subset () function and demonstrate the output.
- Q4** a) Explain natural join and grouping and aggregation relational algebraic operation using MapReduce. [10]
 b) With a neat sketch, explain the architecture of the data-stream management system. [10]

Q5 a) Determine communities for the given social network graph using Girvan-Newman algorithm. **[10]**



b) List and discuss various types of data structures in R. **[10]**

Q6 a) i. The following table shows the number of units of different products sold on different days: **[10]**

| Product | Monday | Tuesday | Wednesday | Thursday | Friday |
|----------------|--------|---------|-----------|----------|--------|
| Bread | 12 | 3 | 5 | 11 | 9 |
| Milk | 21 | 27 | 18 | 20 | 15 |
| Cola Cans | 10 | 1 | 33 | 6 | 12 |
| Chocolate bars | 6 | 7 | 4 | 13 | 12 |
| Detergent | 5 | 8 | 12 | 20 | 23 |

Create five sample numeric vectors from this data.

ii. Name and explain the operators used to form data subsets in R.

b) Define collaborative filtering. Using an example of an e-commerce site like flipkart or amazon describe how it can be used to provide recommendation to users. **[10]**

Time: 3 hours

Max. Marks: 80

N.B. (1) Question No. 1 is compulsory

(2) Assume suitable data if necessary

(3) Attempt any three questions from the remaining questions

Q.1 Solve any Four out of Five

5 marks each

- a Explain the challenges of Natural Language processing.
- b Explain how N-gram model is used in spelling correction
- c Explain three types of referents that complicate the reference resolution problem.
- d Explain Machine Translation Approaches used in NLP.
- e Explain the various stages of Natural Language processing.

Q.2 10 marks each

- a What is Word Sense Disambiguation (WSD)? Explain the dictionary based approach to Word Sense Disambiguation.
- b Represent output of morphological analysis for Regular verb, Irregular verb, singular noun, plural noun Also Explain Role of FST in Morphological Parsing with an example

Q.3 10 marks each

- a Explain the ambiguities associated at each level with example for Natural Language processing.
- b Explain Discourse reference resolution in detail.

Q.4 10 marks each

a

| | | | | | | |
|-----|--------|--------|-------|--------|------|-----|
| <S> | Martin | Justin | can | watch | Will | <E> |
| <S> | Spot | will | watch | Martin | <E> | |
| <S> | Will | Justin | spot | Martin | <E> | |
| <S> | Martin | will | pat | Spot | <E> | |

For given above corpus,

N: Noun [Martin, Justin, Will, Spot, Pat]

M: Modal verb [can , will]

V:Verb [watch, spot, pat]

Create Transition Matrix & Emission Probability Matrix

Statement is “Justin will spot Will”

Apply Hidden Markov Model and do POS tagging for given statements

- b Describe in detail Centering Algorithm for reference resolution.

Q.5 10 marks each

- a For a given grammar using CYK or CKY algorithm parse the statement

“The man read this book”

Rules:

| | |
|----------------------------|---|
| $S \rightarrow NP VP$ | $Det \rightarrow that this a the$ |
| $S \rightarrow Aux NP VP$ | $Noun \rightarrow book flight meal man$ |
| $S \rightarrow VP$ | $Verb \rightarrow book include read$ |
| $NP \rightarrow Det NOM$ | $Aux \rightarrow does$ |
| $NOM \rightarrow Noun$ | |
| $NOM \rightarrow Noun NOM$ | |
| $VP \rightarrow Verb$ | |
| $VP \rightarrow Verb NP$ | |

- b Explain Porter Stemmer algorithm with rules

Q.6 10 marks each

- a Explain information retrieval versus Information extraction systems
b Explain Maximum Entropy Model for POS Tagging

Duration: - 3 Hours

Marks: 80 Marks

NB: - Question 1 is compulsory

Solve any four questions from Question no. 1 .

Solve any three questions from the remaining.

- 1 a. Discuss the objectives of information retrieval systems? **20 (4x5)**
 - b. Explain the process of Structured Text retrieval model.
 - c. Explain the taxonomy of Information retrieval Model.
 - d. Explain the role of pattern matching in Information retrieval.
 - e. Explain multimedia indexing approach.
- 2 a. Illustrate information retrieval system? Discuss its relationship to DBMS, digital libraries and data warehouses **10**
 - b. Explain in detail about vector-space retrieval models with an example? **10**
- 3 a. What is local and global analysis and Differentiate between automatic local analysis and global analysis? **10**
 - b. What is the role of suffix array and suffix tree in information retrieval system with example. **10**
- 4 a. What is the Signature File ? Explain the structure of signature files with example? **10**
 - b. What is the significance of tf and idf ? How can you calculate tf and idf in vector model? **10**
- 5 a. Compare and contrast evaluation of ranked and unranked Retrieval Results ? **10**
 - b. Explain Query Processing in context of Distributed IR? **10**
- 6 Write short notes on **any two** **20**
 - a. Rocchio method for Query expansion
 - b. Parametric and zone indices
 - c. Latent Semantic Indexing Model
 - d. Flat browsing vs hypertext browsing model.

(3 Hours)

Total Marks: 80

Note:

1. **Question No. 1 is compulsory.**
2. Attempt any **THREE** out of the remaining **FIVE** questions.
3. Assume suitable data if necessary.

- 1 Answer the following (any 4) (20)
 - a) Define the terms: Mitigation, Triage 5
 - b) What is Climate Change? What are the effects of Global Warming? 5
 - c) What is Disaster Scenario of India? 5
 - d) What are Man-made hazards? Also discuss possibilities of chemical spills. 5
 - e) Explain Risk and Vulnerability. 5
- 2
 - a) Elaborate the guidelines laid down by NDMA for disaster management in India. 10
 - b) What are different types of flood? Enlist structural mitigation measures for flood. 10
- 3
 - a) Explain disasters related to industries. Discuss the socioeconomic impact of industrial disasters with a case study. 10
 - b) Enlist and explain various terms and concepts with respect to earthquakes. 10
- 4
 - a) Explain role of various softwares in disaster management. 10
 - b) What are different mitigation measures for Tsunami and Cyclones? 10
- 5
 - a) What is the role NGO in disaster management? Enlist major NGOs working on disaster management 10
 - b) Appraise the role of Geo-informatics in disaster management 10
- 6
 - a) Explain various means of raising finance for mitigating and managing disasters 10
 - b) Discuss the role of GIS and Remote Sensing in disaster management. 10
