NEERAJ®

M.C.S.-226

Data Science and Big Data

Chapter Wise Reference Book
Including Many Solved Sample Papers

Based on

I.G.N.O.U.

& Various Central, State & Other Open Universities

By: *Anand Prakash Srivastava*

NEERAJ PUBLICATIONS
(Publishers of Educational Books)

Mob.: 8510009872, 8510009878 E-mail: info@neerajbooks.com
Website: www.neerajbooks.com

MRP ₹ 280/-

Content

DATA SCIENCE AND BIG DATA

S.No.	Chapterwise Reference Book	Page

BLOCK-1: BASICS OF DATA SCIENCE

BLOCK-2: BIG DATA AND ITS MANAGEMENT

■■

Sample Preview of the Solved Sample Question Papers

Published by:

NEERAJ PUBLICATIONS

www.neerajbooks.com

QUESTION PAPER

June – 2023

(Solved)

DATA SCIENCE AND BIG DATA

M.C.S.-226

Time: 3 Hours] ***[Maximum Marks : 100***
Weightage: 70%

Note: *Question* ***No. 1*** *is compulsory. Attempt any* ***three*** *questions from the rest.*

Q. 1. ***(a)*** **Define Data Science. Give advantages of Data Science in an organization.**

Ans. Ref.: See Chapter-1, Page No. 1, 'Data Science-Definition'.

(b) **Explain Bayes' Theorem with suitable equation and example.**

Ans. Ref.: See Chapter-2, Page No. 11, 'Bayes Theorem' and Page No. 18-19, Q. No. 3.

(c) **What is a Histogram? How does Histogram differ from Bargraph? Briefly discuss the utility of Histogram in Data Science.**

Ans. Ref.: See Chapter-4, Page No. 33, 'Histogram' and Page No. 36, Q. No. 1.

Also Add: Here are some key utilities of histograms in data science:

1. Understanding Data Distribution: Histograms allow data scientists to quickly understand the distribution of data, including the shape, central tendency, and variability. This is crucial for gaining insights into the nature of the data.

2. Identifying Patterns and Trends: Patterns, trends, and clusters within the data become visually apparent in histograms. Peaks, valleys, and shapes provide valuable information about the structure of the dataset.

3. Detecting Outliers: Outliers, which are values significantly different from the majority of the data, can be easily identified in histograms. This helps in assessing the impact of extreme values on the distribution.

(d) **What is Hadoop MapReduce? Give its advantages. Also, discuss how <key-value> pair mechanism facilitates MapReduce programming.**

Ans. Ref.: See Chapter-6, Page No. 59, 'Hadoop Map Reduce' and Page No. 61-62, Q. No. 2.

Also Add:

<key-value> Pair Mechanism in MapReduce:

The key-value pair mechanism is fundamental to the MapReduce programming model. In MapReduce, both the input and output of each stage (map and reduce) are in the form of key-value pairs. This mechanism facilitates the parallel processing of data and allows developers to express a wide range of data processing tasks.

- **Map Phase:** In the map phase, the input data is divided into key-value pairs. The map function processes each input record and produces a set of intermediate key-value pairs.
- **Shuffle and Sort Phase:** The intermediate key-value pairs are shuffled and sorted based on the keys. This ensures that all values associated with a particular key are grouped together.
- **Reduce Phase:** In the reduce phase, the sorted key-value pairs are passed to the reduce function. The reduce function aggregates the values associated with each key and produces the final output.

(e) **In context of Data Science, what is Apache SPARK? How does Apache SPARK differ from Hadoop?**

Ans. Ref.: See Chapter-7, Page No. 68, 'Apache Spark Framework's'.

Also Add: Apache Spark and Hadoop are both distributed computing frameworks designed to process and analyze large-scale data sets, but they differ in several key aspects. Here are some of the primary differences between Apache Spark and Hadoop:

1. Processing Model: Hadoop: Primarily uses the MapReduce programming model, which involves two phases (map and reduce) for processing data. It is well-suited for batch processing.

Spark: Offers a more flexible processing model, including batch processing, interactive queries, streaming, and machine learning. Spark introduces the concept of Resilient Distributed Datasets (RDDs), allowing for in-memory processing.

2. Ease of Use: Hadoop: Writing MapReduce jobs can be verbose and requires developers to handle low-level details, making it less user-friendly.

Spark: Provides high-level APIs in languages like Scala, Java, Python, and R. Spark's APIs are more expressive and user-friendly than Hadoop's MapReduce, allowing for concise and readable code.

3. In-Memory Processing: Hadoop: Typically relies on disk-based storage and processing, which can result in slower performance.

Spark: Utilizes in-memory processing, allowing intermediate data to be stored in memory between stages. This significantly speeds up iterative algorithms and interactive queries.

***(f)* What are Data Streams? How do Data Streams differ from Databases? Why mining of data streams is considered as a challenging process in Data Science?**

Ans. Ref.: See Chapter-10, Page No. 92, 'Data Stream' and Page No. 93-94, 'Data Stream Management' and 'Issues and Challenges of Data Stream'.

***(g)* Explain PageRank algorithm, with suitable example.**

Ans. Ref.: See Chapter-11, Page No. 102, 'Page Ranking'.

***(h)* What are Dataframes in 'R' programming? Give characteristics of Dataframes.**

Ans. Ref.: See Chapter-13, Page No. 121, 'Data Frame'.

Q. 2. *(a)* Write the syntax to create the following plots in 'R' :

***(i)* Bar charts**

Ans. Ref.: See Chapter-14, Page No. 128, 'Bar Charts and Syntax'.

***(ii)* Box plots**

Ans. Ref.: See Chapter-14, Page No. 128, 'Box Plots and Syntax'.

***(iii)* Histogram**

Ans. Ref.: See Chapter-14, Page No. 128, 'Histogram and Syntax'.

***(iv)* Line graphs**

Ans. Ref.: See Chapter-14, Page No. 128, 'Line Graphs and Syntax'.

***(v)* Scatter plots**

Ans. Ref.: See Chapter-14, Page No. 128, 'Scatter Plots and Syntax'.

***(b)* Differentiate between Linear Regression and Multiple Regression, with suitable example for each.**

Ans. Ref.: See Chapter-15, Page No. 136, Q. No. 3.

***(c)* What are Decision Trees? What are categorical variables and continuous variables? How do these two variables relate to decision trees? Explain the role of entropy and information gain in decision trees.**

Ans. Ref.: See Chapter-16, Page No. 140, 'Decision Trees'.

Also Add: Role of Entropy and Information Gain:

1. Entropy as a Measure of Impurity: Entropy is used to quantify the impurity or disorder in a dataset. Lower entropy indicates a more homogenous set.

2. Information Gain for Feature Selection: Information Gain measures the effectiveness of a feature in reducing uncertainty about the target variable.

The decision tree algorithm selects the feature that maximizes Information Gain to split the data, resulting in more homogenous subsets.

3. Recursive Splitting: Decision trees use entropy and Information Gain iteratively for recursive splitting. The process continues until a stopping criterion is met, such as a certain depth or the purity of the nodes.

Q. 3. *(a)* Compare qualitative data with quantitative data. What do you understand by the term "Measurement Scale of Data"? Give characteristics of measurement scales of data. List various measurement scales with suitable example for each.

Ans. Ref.: See Chapter-1, Page No. 2, 'Statistical Data Types' and 'Measurement Scale of Data'.

***(b)* What is a Random Variable? Differentiate between Discrete Random Variable and Continuous Random Variable.**

Ans. Ref.: See Chapter-2, Page No. 11-12, 'Random Variables and Basic Distributions'.

***(c)* What is a Heat Map? Give uses and best practices for Heat Maps.**

Ans. Ref.: See Chapter-4, Page No. 34, 'Heat Map'.

Q. 4. *(a)* Explain the following operations of map-reduce with suitable example and supporting block diagram :

***(i)* Splitting**

***(ii)* Mapping**

***(iii)* Shuffling**

***(iv)* Reducing**

Ans. Ref.: See Chapter-11, Page No. 104, 'Page-rank Computation Using Mapreduce'.

Sample Preview of The Chapter

Published by:

NEERAJ PUBLICATIONS

www.neerajbooks.com

DATA SCIENCE AND BIG DATA

Introduction to Data Science

1

INTRODUCTION

The growth of the Internet and communication technology over the past ten years has produced a significant volume of unstructured data. This unstructured data encompasses information that is generated as a result of people using social media and mobile technologies, such as unformatted textual, visual, video, and audio data, among other types of information. Additionally, a significant amount of semi-structured data, such as XML data, is being generated at a significant rate as a result of the great expansion in the digital eco system of organizations.

The vast amount of data derived from organizational databases and data warehouses is in addition to all of this data. The decision-making processes of various organizations may be supported by the real-time processing of this data. Data science is a field that focuses on the processes of gathering, integrating, and analyzing massive amounts of data to provide information that can be used for making decisions.

CHAPTER AT A GLANCE

DATA SCIENCE – DEFINITION

Data science can be applied in a variety of organizations, some of which are listed below. Data science is a multidisciplinary science that aims to analyses data to provide information that can be applied to making decisions. This information may take the shape of forecasting models, predictive planning models, or other models that use comparable patterns.

The categories listed below are just a few where data science can be helpful.

1. It assists in making business decisions, such as determining the viability of the company they intend to work with.
2. It might aid in developing more accurate future projections, such as helping businesses create strategic strategies based on current patterns.
3. It might spot trends in different data that are similar, leading to applications like targeted advertising and fraud detection.

Data science can be applied in a variety of organisations, some of which are listed below.

1. By recommending the ideal time and route for food transportation, it may help an organization lower its logistical costs.
2. By detecting comparable group buying trends and implementing targeted advertising based on the information acquired, it can save marketing expenses.
3. It can be useful in developing public health strategies, particularly in catastrophe situations.

TYPES OF DATA

The type of data is one of the key factors that dictates the kind of analysis that needs to be done on the data. The various forms of data that must be processed in data science include the following:

1. Structured Data
2. Semi-Structured Data
3. Unstructured Data
4. Data Streams

1. Structured Data: Computers have been utilised as data processing tools ever since the dawn of the computing era. However, businesses did not begin employing computers to process their data until the 1960s. Common Business-Oriented Language (COBOL) was among the most widely used languages at the time. The division in COBOL used to stand in for the data structure being processed.

2. Semi-Structured Data: Semi-structured has some structure as the term would imply. The employment of tags or key/value pairs gives semi-structured data its structure. Semi-structured data is frequently created using XML, JSON objects, server logs, EDI data, etc.

3. Unstructured Data: The unstructured data does not adhere to any defined schema. For instance, an unstructured written text like the material in this unit. For unstructured data, you can add specific headings or meta information. In actuality, Zettabytes of unstructured data have been produced as a result of the expansion of the internet.

4. Data Streams: A series of data collected over time is what makes up a data stream. Whether structured, semi-structured, or unstructu-red, this type of data is always being produced. at instance, Internet of Things (IoT) devices like weather sensors will produce a data stream of pressure, temperature, wind speed, wind direction, humidity, etc. at a specific location where they are installed.

Statistical Data Types

In statistical analysis, there are two different types of data that can be used.

These are – Categorical data and Quantitative data

Categorical or Qualitative Data: In order to define a category of data, categorical data is employed. For instance, a person's occupation could have values from the categories "Business," "Salaried," "Others," etc. Nominal and Ordinal are two different measurement scales that can be used for the categorical data.

Quantitative Data: Quantitative data is numerical information that can be used to define various data scales. Additionally, there are two primary categories of qualitative data: discrete, which represents distinct numbers like 2, 3, 5, etc., and continuous, which provides continuous values of a specific variable. For instance, a continuous scale can be used to assess your height.

Measurement Scale of Data: Data are unprocessed facts, such as student information such as name, gender, age, height, etc. Similar to a primary key in a database, the name is often a differentiating data that seeks to clearly identify two data objects. A psychologist named Stanley Stevens identified the following four qualities that any scale that can be assessed must possess:

- The term "identify of a value" (IDV) refers to the requirement that each representation of the measure be distinct.
- The magnitude (M), which can be used to compare values, is the second characteristic. For instance, a weight of 70.5 kg is greater than a weight of 70.2 kg.
- The third feature is related to the Equality of the Intervals (EI) used to describe the data. For instance, the difference between 25 and 30 is 5 intervals, and it is likewise 5 intervals between 41 and 46.
- The final features are about a specified minimum or zero value (MZV), for instance, temperature has a MZV in the Kelvin scale.

Sampling

The amount of data that needs to be handled today is generally pretty considerable. This prompts you to consider whether you should utilize all of the data or just a representative sample of it. In a number of data science methodologies, an exploratory model is also developed using sample data.

BASIC METHODS OF DATA ANALYSIS

Several data sources provide the information needed for data science. This data is first cleared of mistakes and duplicate entries, then aggregated, before being displayed in a way that allows for various types of analysis.

Descriptive Analysis

Basic summaries of data are presented using descriptive analysis, but no attempt is made to analyze the data. These summaries could incorporate various statistical results and graphs.

Descriptive of Categorical Data: In figure below the Gender variable is categorical. In this instance, a frequency table of multiple categories would serve as the summary. For instance, the frequency distribution for the provided data would be:

Gender	Frequency	Proportion	Percentage
Female (F)	5	0.5	50%
Male (M)	5	0.5	50%

Descriptive of Quantitative Data: Height is a quantitative variable that is described by quantitative data. There are two different ways to describe quantitative data:

1. Identifying the primary trends in the data.
2. Describing how the data were distributed.

Central tendencies of Quantitative Data: Two fundamental measurements, the mean and the median, distinguish the center of the data in various ways.

Mode: The most prevalent value of a group of observations is referred to as the mode. Any observational value may serve as the mode value; it need not be the midpoint.

Spread of Quantitative Data: The distribution or variability of the observed data is a key consideration when establishing the quantitative data.

Exploratory Analysis

John Turkey of Princeton University proposed exploratory data analysis as a series of techniques that can be used to discover potential links between data in 1960. Following are a few typical techniques you can use in an exploratory analysis:

1. You can start by performing a descriptive analysis of your data's numerous category and qualitative variables.
2. Next, you may do some bi-variate analysis after finishing the univariate analysis.
3. As a third option, you might consider investigating the potential for multi-variate correlations between data.

Inferential Analysis

The purpose of inferential analysis is to determine the likelihood that the findings of a study may be generalized to the entire population.

Predictive Analysis

Advanced predictive analysis has been made possible by the availability of massive amounts of data and cutting-edge algorithms for mining and analyzing large data. Today's predictive analysis makes predictions for organizational strategic planning and policies using methods from artificial intelligence, machine learning, data mining, data stream processing, data modeling, etc. Large amounts of data are used in predictive analysis to find potential dangers and support decision-making.

COMMON MISCONCEPTIONS OF DATA ANALYSIS

Correlation analysis establishes a link between two variables, not causality. Consider three factors, such as a student's attendance, the grades they received, and the number of hours per week they dedicated to their studies. You discovered via data analysis that there is a significant association between the variables attendance and grades. Similar to this, a driven student who is devoting more time to studying can also be attending classes on a regular basis.

Correlation is not Causation: Correlation analysis establishes a link between two variables, not causality. Consider three factors, such as a student's attendance, the grades they received and the number of hours per week they dedicated to their studies. You discovered via data analysis that there is a significant association between the variable's attendance and grades. Similar to this, a driven student who is devoting more time to studying can also be attending classes on a regular basis.

Simpsons Paradox: It is an intriguing circumstance that occasionally results in erroneous interpretations.

Data Dredging: As the name suggests, data dredging is a thorough investigation of very big data sets. Many data associations are produced as a result of this research. Many of those relationships might not be coincidental, necessitating further investigation using different methods.

APPLICATIONS OF DATA SCIENCE

Large data sets can be analysed with the aid of data science to provide insightful results that support decision-making and business development. Some uses for data science are highlighted in this section.

Applications using Similarity analysis

These applications employ algorithms to analyse data similarity and classify or cluster it into categories. Examples of these applications include:

- **Spam detection system:** Classifies emails as spam or nonspam. The system evaluates IP addresses, word patterns, and frequency to determine if a mail is spam.
- **Financial Fraud Detection System:** A key application for online financial services. Classifying transactions as safe or hazardous based on numerous characteristics is the basic idea.
- E-commerce companies may use your buying patterns, search history, and account information to recommend products to you. This data can be categorised into buyer groups to suggest certain products.

Applications related to Web Searching

These apps mostly improve web content discovery.

Applications in this area include search algorithms employed by various search engines. These algorithms identify relevant webpages using search phrases. They may employ technologies for semantic analysis, indexing key websites and phrases, and link analysis. Additionally, browser predictive text usage is an example.

Applications related to Healthcare System

Healthcare applications can benefit greatly from data science. Applications include processing and interpreting photos for infant care and detecting tumours, abnormalities and organ issues. Additionally, public health data can be used to develop disease-factor connections and make recommendations for public health. This includes genomic analysis, medication development, and testing. Using streaming data for patient monitoring is a potential application of data science in healthcare.

Applications related to Transport sector

E-commerce enterprises may utilise these tools to determine the most cost-effective pathways for logistic assistance from warehouses to customers. Determine optimal dynamic routes from source to destination, considering road network loads.

DATA SCIENCE LIFE CYCLE

A data science-based application's life cycle. The following stages are typically included in the development of a data science application:

Data Science Project Requirements Analysis Phase: Finding the project's goals is the first and most important phase in any data science project. Along with the determination of objectives, a study of the project's advantages, resource needs, and cost is also conducted.

Data Collection and Preparation Phase: This phase begins with the identification of all the data sources and ends with the creation of the data gathering mechanism. It should be emphasized that gathering data might be an ongoing activity.

Descriptive Data Analysis: Both univariate and bivariate analytic methods are used to analyze the data. As a result, the data will produce descriptive information. This stage can also be used to determine whether the data is valid and appropriate for the purposes of data analysis.

Data Modeling and Model Testing: On the basis of the data, several data models are created. The validity of each of these data models is subsequently examined using test data.

Model Deployment and Refinement: The tried-and-true best model is utilized to tackle the data science challenge, but this model needs to be updated frequently because the decision-making environment is always shifting and new data sets and qualities could emerge over time. All the previous processes are repeated during the refinement phase.

CHECK YOUR PROGRESS

Q. 1. Define the term data science.

Ans. In the interdisciplinary subject of data science, knowledge and insights are derived from both organized and unstructured data. To analyze and comprehend complex data sets, it incorporates numerous methods and theories from statistics, mathematics, computer science, and information science. Finding patterns, making predictions, and producing actionable insights that can be used to decision-making and problem-solving in the real world are the major objectives of data science.

Data collecting, data cleaning, data analysis, data visualization, and the development of prediction models are essential elements of data science. It frequently involves the use of large data processing frameworks, machine learning, artificial intelligence, and programming languages including Python, R and SQL.

Data scientists are experts at using sophisticated programming and analytical techniques to glean valuable insights from data. They are essential in assisting companies, groups, and governments in making data-driven decisions and gaining a competitive edge in a variety of industries, including marketing, finance, healthcare and many others.

Q. 2. Differentiate between structured, semi-structured, unstructured and stream data.

Ans. Depending on its structure and properties, data can be categorized into several categories. Here are the differences between organized, semi-structured, unstructured, and stream data:

Structured Data: Data that has been organized in a way that is highly predictable and clearly defined is referred to as structured data. Typically, it is kept in a tabular format with a defined schema in relational databases or spreadsheets. The links between various data points are well stated, and each data element is easily searchable. Inventory lists, financial records, and transactional data are a few examples of structured data.