An Introduction to Mathematics and Machine Learning for Data Analysis

An Introduction to Mathematics and Machine Learning for Data Analysis

By

Junye Wang

and Mojtaba Aghajani Delavar

Cambridge Scholars Publishing

An Introduction to Mathematics and Machine Learning for Data Analysis

By Junye Wang and Mojtaba Aghajani Delavar

This book first published 2026

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-1-0364-6531-5
ISBN (Ebook): 978-1-0364-6532-2

CONTENTS

Preface x

1. Introduction 1
- 1.1. Data, big Data and information
- 1.2. Data Science
- 1.3. Databases and bata warehouse
- 1.4. Data analysis and data analytics
- 1.5. Problems and tasks of statistical analytics
- 1.6. Data scientist, big data professional and data analyst
- 1.7. Who should take this course?

Exercises

2. Data types, statistical properties, and statistical methods 9
- 2.1. Introduction
- 2.2. Data types and properties
  - 2.2.1. Qualitative and quantitative data
  - 2.2.2. Independent vs. dependent data
  - 2.2.3. Structured data, semi-structured data, and unstructured data
- 2.3. Data statistical properties
  - 2.3.1. Sample distributions
  - 2.3.2. Mean, median, mode and central tendency
  - 2.3.3. Sample variability and standard deviation
- 2.4. Statistical methods
  - 2.4.1. Descriptive statistical analysis
  - 2.4.2. Inferential statistical analysis
  - 2.4.3. Predictive analysis
  - 2.4.4. Prescriptive analysis
  - 2.4.5. Exploratory data analysis
  - 2.4.6. Causal analysis
  - 2.4.7. Mechanistic analysis

Exercises

3. Jupyter and Python 26
3.1. Introduction
3.2. Installing Jupyter using Anaconda
3.3. Running Jupyter Notebook and Python
3.4. Python packages or libraries and installation
3.4.1. Python libraries
3.4.2. Python libraries and installation
3.5. Summary
Exercises

4. Data mining and methods 85
4.1. Introduction
4.2. Data mining methods
4.3. Tasks of data mining
4.4. Data analysis process
4.4.1. General steps
4.5. Artificial Neural Network (ANN)
4.5.1. Neurons and neural network
4.5.2. Feed forward neural network
4.5.3. Back-propagation neural network
4.5.4. Example of ANN and ML
4.6. Summary
Exercises

5. Regression 104
5.1. Introduction
5.2. Single linear regression
5.2.1. Concept of single linear regression
5.2.2. Least squares estimation
5.2.3. Example 1
5.2.4. Example 2
5.3. Multiple linear regression
5.3.1. Least-squares estimation
5.3.2. Example 3
5.4. Nonlinear regression
5.4.1. Polynomial regression
5.4.2. Logistic regression
5.4.3. Quadratic regression
5.4.4. Cubic regression
5.5. Summary
Exercises

6. Classification 132
6.1. Introduction
6.2. Binary and multiclass classification
6.3. K-Nearest Neighbor (K-NN) algorithm
6.3.1. K values
6.3.2. k-NN model
6.3.3. k-Nearest neighbor classification and regression
6.3.4. Feature Weighting and distance-weighted k-NN
6.4. Probability-based algorithms
6.4.1. Naïve Bayes algorithms
6.4.2. Mathematical model of naïve Bayes
6.4.3. Gaussian naive Bayes
6.5. Summary
Exercises

7. Clustering 158
7.1. Introduction
7.2. Hierarchy clustering
7.2.1. Single linkage clustering
7.2.2. Complete linkage clustering
7.2.3. Average linkage clustering
7.2.4. Centroid linkage
7.3. K-mean algorithm
7.3.1. Voronoi diagram
7.3.2. k-means clustering
7.4. Density-based clustering
7.4.1. DBSCAN clustering algorithm
7.4.2. Example of DBSCAN clustering
7.5. Summary
Exercises

8. Data fitting, approximation and interpolation 178
8.1. Introduction
8.2. Data fitting and approximation
8.2.1. Linear fitting
8.2.2. Nonlinear curve fitting
8.2.3. Fitting surface
8.2.4. Evaluation of the model performances
8.3. Data interpolation and extrapolation
8.3.1. Interpolation
8.3.2. Extrapolation

8.4. Summary
Exercises

9. Gradient and heuristic Algorithms 202
9.1. Introduction
9.2. Gradient Descent
9.2.1. Adaptive optimization algorithm (mentum-based optimization)
9.2.2. Hill climbing
9.2.3. Example 1
9.2.4. Example 2: Rosenbrock's function
9.3. Simulated Annealing Algorithm
9.3.1. SA implementation
9.3.2. Example 3
9.4. Genetic algorithm (GA)
9.4.1. GA implementation
9.4.2. GA advantages and limitations
9.4.3. Example 4
9.5. Summary
Exercises

10. Time series analysis 232
10.1. Introduction
10.2. Time series data
10.3. Autoregressive Integrated Moving Average (ARIMA) model
10.4. Long Short-Term Memory
10.5. Summary
Exercises

11. Visualize & interpret data 252
11.1. Introduction
11.2. 2D Plots
11.1.1. Scatter plotting
11.1.2. Line plots
11.1.3. Curve plots
11.1.4. Add labels, legend, and title
11.1.5. Multiple plots using subplot function
11.1.6. Density and contour plots
11.1.7. 3D frame, axes and scatter plots
11.1.8. 3D line or curve plots
11.1.9. 3D wireframes and surface plots

11.2. Bar and pie plots
11.3. Maps
11.4. Summary
Exercises

PREFACE

Data is everywhere and digital life is revolutionizing everything from their work, play and home lives every day. The amount of data we create every day is growing exponentially. A very large amount of digital data of 2.5 quintillion bytes is produced each day at our current pace. However, raw data refers to unprocessed facts, figures, images, and values that, on their own, lack context or meaning. It is essentially the raw material from which information can be derived once it is processed, organized, and interpreted. Therefore, data needs to be processed, organized, structured or presented to extract meaningful information. With growing data exponentially, businesses need to interpret data and extract useful information, particularly complex data, and provide insights with visualization tools. Data analysis skills are necessary in current business management. Fast-growing data poses challenges in data handling using traditional statistical approach. Data handling plays a key role in making the data useful to improve public services, optimize resource allocation, urban planning, health care and enhance citizen engagement. Thus, artificial intelligence (AI) and machine learning (ML) are revolutionizing data analysis in improved statistical and computational methods and create opportunities for deeper analysis, research, and the development of more advanced tools and solutions in data sciences. As an interdisciplinary field, data analysts require to be trained in interdisciplinary environments rather than only mathematics or computers.

There are many books of data analysis in markets. However, these existing books were written for students of computer machine learning or mathematical statistics. The authors have worked on multidisciplinary modelling and data analysis and understand well different needs for data analysis in different sectors (engineering, mathematics, biogeochemistry, biology, computer and physics). First, the textbook is self-contained for senior undergraduates and first-year graduates that introduces many basic principles and techniques of mathematics and statistics needed for modern data analysis. Many students or data analysts may want to learn these subjects in a simpler way in many different areas, such as environment, biology, social sciences and engineering, but know a little bit about mathematics, statistics, computing or programming. Therefore, there is a need of textbooks in data analysis to bridge mathematical concepts, AI and

practical applications that would be more suitable for the practical applications in various sectors. Thus, our textbook aims to avoid jargon, such as advanced mathematical operators in advanced mathematics, to enable students to quickly grasp concepts which are understandable by any university student whichever their background is and contribute to various practical needs and applications of data analysis. It is challenging in data analysis how students develop their interdisciplinary capacities and skills beyond computer and mathematics. This textbook intends to provide a new, more friendly course not only for computing and mathematical students, but also for various other disciplines. Second, the textbook has no pre-requisites and avoids all but the simplest mathematics. Today, applying ML or AI does not require a high degree. The textbook covers all the important aspects of implementing data analysis using ML in practice, without requiring you to spend years' worth of advanced math courses, such as linear algebra, and probability theory. The students that came from different backgrounds all acquired an understanding of the basic algorithm and its theory of operation. This textbook will explore the use of non-traditional means to credential learning and blended learning across educational settings and learners' lifespans. Learning science and data analytics entails physical intuition and understanding, which can be gained through practice and experience. This textbook derives theories from physical intuition and understanding linked to real applications, so that students can better understand the context and questions of science. This textbook will help students to use their intuition along with mathematics to analyze and evaluate data problems. To help them gain enough intuition on how learning can be nurtured, the book will include selected sample data problems using alternative approaches to get the students to illustrate each problem from different perspectives. The textbook will not go into unnecessary theoretical details, as unnecessary details overshadow the concept. Anyone with moderate computer experience should be able to master the materials in this course. Third, the textbook will be kept concise but self-pace and self-study courses. This textbook is unique to write for both students and data analysts in mathematical statistics, computer, and other disciplines to bridge knowledge gaps for those students and data analysts in various sectors. Thus, we will include a chapter of statistics of data properties and a chapter of Jupyter and Python with associated libraries. These basic mathematics and programming will allow the learners who are not from computers and mathematics to solve problems of data analysis using machine learning. This knowledge is important for data analysis and is appealing for students and analysts. For students in computer or mathematics, the textbook will provide more background

knowledge of data applications. However, some students may want to learn these subjects in a simpler way in many different areas, such as environment, biology, social sciences and engineering, but know a little bit about mathematics, computing or programming. It aims to be understandable by any university student, whichever their background is and contribute to various practical needs and applications of data analysis. Therefore, students do not need to be students of mathematics, computer sciences, or even a degree student. It is expected that after learning this textbook, the students will be able to develop a project on data analytics, assuming that they are already familiar with the business area of the project and are able to analyze complex data. However, mathematics and programming knowledge are indeed helpful. Those who have already studied subjects, such as statistics and machine learning, will recognize some of the content described in this book. As an interdisciplinary science, integration of machine learning and data analysis is an emerging, hot subject. Our textbook has main differences from the above books.

This textbook is organized into two main parts: The first part (Chapters 1, 2, 3, and 11) presents the fundamental concepts of data science, data types, statistical properties, statistical methods, and visualization; and the second part centers on advanced methods of data analysis and their applications (Chapters 4, 5, 6, and 7) and a few specific topics and techniques of data analysis (Chapters 8, 9, and 10). Chapters 1 introduce the background concepts and problems of data science and Chapter 2 presents fundamental of data types, statistical properties, and statistical methods. Moreover, the basics of Python are included for self-contained and learning. Chapter 3 introduces Jupyter, Python basic and Python libraries and describes their setup we will be using. The Jupyter Notebook offers an interactive environment for coding, editing, visualizing and running python programs. Although the Jupyter supports many programming languages, we illustrate the Python only. Chapter 11 offers an overview of visualization and mapping, including references to more advanced visualization topics. It is important that using charts, graphs, and maps presents the results of data analysis using python libraries. These chapters introduce key conceptual tools which are often used for data analysis and are for those students who are not from computers and mathematics.

The advanced methods of data analysis and their applications comprise the backbone of modern data analysis, including Chapters 4, 5, 6, and 7. Chapter 4 describe the actual datamining problems and processes and concepts of machine learning and artificial neural network algorithms that are most widely used in practice and discuss their advantages and shortcomings. Chapters 5, 6 and 7 introduce regression, classification and

clustering, respectively and how to apply the methods described in rigorous ML and AI courses. In this textbook we illustrate the big picture of the algorithm or method themself to solve a data problem and in its various physical and other applications and not spend a lot of time in the rigorous mathematics of the algorithm or method that may lose sight of the ultimate goals. For this reason, the mathematical background includes only brief proofs, and no attempt is made to elaborate on many details of the method. Thus, we still encourage you to keep in mind all the problems that you might be making, explicitly or implicitly, when you start building ML or AI models. Chapters 8, 9, and 10 dig deeper in a few more specific topics and techniques of data analysis, including data fitting and interpolation, optimization algorithms and time series analysis. The use of optimization algorithms is a quickly evolving field of data analysis, and there is much new to recommend. Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time. We do, however, give readers a flavor for how algorithms are being used and how to fiddle with them to get the best performance. It should be useful to incorporate these methods into data analysis. The task is facilitated if students have already some backgrounds in data science and Python, the background material in Chapters 1, 2, 3, and 11 can be assigned for independent review.

Our textbook aims to be understandable by any university student, whichever their background is and contributes to various practical needs and applications of data analysis beyond computer and mathematics. Many data analysts may want to learn these subjects in a simpler way in many different domains, such as environment, biology, social sciences and engineering, but know a little bit about mathematics, computing or programming. This textbook is appealing to them. Our textbook aims to be understandable by any university student, whichever their background is and contributes to various practical needs and applications of data analysis beyond computer and mathematics. The textbook provides exercises at the end of every chapter, and students will therefore have plenty of opportunities to test their own data analysis skills and expertise. Example codes and data are provided according to each chapter in https://github.com/JunyeWang888/An-Introduction-to-Mathematics-and-Machine-Learning-for-Data-Analysis

CHAPTER 1

INTRODUCTION

We utilize data constantly for anything from environmental monitoring to business analysis. The capacity to generate, store, and process data generated by humans and machines is expanding exponentially. At a 50× rate, machine-generated data is expanding even more quickly. Since the invention of technology, enormous amounts of text, video, image, and audio data are produced every year, especially on smartphones and the internet. This data is mostly unstructured. Every day, an estimated 328.77 million terabytes of data are produced.

The amount of data being created and stored appears to be rapidly increasing, as indicated by the compound annual growth rate of 21.2%. Global data generation is expected to increase by 180 zettabytes by 2025. Overcoming 221,000 exabytes by 2026 would represent a remarkable turning point in the digital landscape, illustrating the exponential growth in both technology and data production (Burgener and Rydning, 2022). On the other hand, data are unprocessed, raw facts, statistics, or figures that contain a limited amount of factual information but are not considered information. In order to obtain meaningful knowledge, data must be processed, analyzed, organized, or structured in a specific context. Data analysis aims to transform data into information that can be used to make decisions. Businesses must understand data and give decision-makers with visual patterns and insights as a result of the daily increase in data generation. Data has been essential to practically every activity we do today in a variety of industries. The change from a product-focused to a data-focused paradigm is a reflection of the realization that there are priceless insights in the massive data streams that are essential for strategy development, innovation, and decision-making.

Data is now a crucial indicator for assessing the performance of an organization. It enables companies to understand their clients better, modify their current strategies, and enhance their client experience. Organizations hoping to succeed in the digital sphere must foster a data-driven culture. Due to massive data generation, there is currently a need for qualified experts who can extract insightful information from complex datasets. Data

is like fuel to get a firm moving in the correct direction. It offers data that can be utilized to improve the strategy of ongoing promotions and upcoming goods. For companies looking to maintain their competitiveness in the fast-paced, information-driven economy of nowadays, data scientists, analysts, and engineers are essential. Beginners to intermediate data analysis foundations are provided by the special book.

1.1. Data, Big Data, and Information

In the digital age, data is a raw ingredient and is made up of elements like numbers, words, photos, sounds, and movies. After those ingredients are analyzed and interpreted, knowledge is the final product. Data have no meaning until we turn them into knowledge. To extract useful ideas and turn them into knowledge, we must arrange, examine, and evaluate this material. Information is, in fact, structured, processed data that has practical value for the user and can be used to make decisions. Accuracy, completeness, consistency, accessibility, understandability, relevance, and timeliness in decision-making are critical requirements for processed data to be of any meaningful use. Therefore, the ultimate goal of data analysis is to transform data into knowledge and information that can be useful in making decisions.

In modern computer applications, the term "data" has taken on a far broader definition. Massive amounts of data must now be stored and analyzed using both conventional database administration techniques and their strategic value in fostering innovation, informing choices, and defining our digital world. Big data is the fashionable term for the incredibly enormous and complicated datasets—both organized and unstructured—that businesses deal with on a daily basis. It is not feasible to handle big data using conventional methods, including storing all of the data in a single computer's memory or processing it using typical databases, due to its enormous scale or volume, velocity, and variety. Consequently, the five Vs—Variety, Volume, Value, Veracity, and Velocity—are used to define big data. Businesses can find patterns, trends, correlations, and other useful insights to assist with strategic decisions by analyzing big data. Cost-effective data processing that improves understanding and decision-making necessitates innovative approaches.

1.2. Data Science

Data science is a multidisciplinary field that processes structured, unstructured, and semi-structured data. It combines expertise in mathematics,

statistics, computer science, domain expertise, and algorithms to analyze and explain complex datasets. The field encompasses various activities, including data collection, preprocessing, synthesis, and predictive modeling. The primary goal of data science is to uncover patterns, trends, and correlations within data, providing valuable insights that support decision-making, problem-solving, and innovation across industries and businesses.

1.3. Databases and Data Warehouse

A database is a structured collection of information designed for efficient storage, retrieval, manipulation, and management of data. It typically organizes data into tables, rows, and columns, with each piece of information linked to a specific attribute or category. A well-designed database ensures data integrity, supports aggregation, enables concurrency control, enhances security, and offers scalability for efficient and reliable data management. While data can be stored digitally on a computer system or physically in non-digital formats like filing cabinets, this book will focus exclusively on digital databases.

A Data Warehouse is a specialized database that stores and manages large volumes of structured and unstructured historical data from multiple sources. Unlike traditional databases, which are optimized for transactional processing in daily operations, data warehouses are built for the analytical processing of vast datasets. They support decision-making processes by enabling business intelligence (BI) and data analytics. The main differences between databases and data warehouses lie in their purpose, data processing methods, and data structures—while databases handle real-time transactions, data warehouses facilitate complex queries and trend analysis over time.

Databases and data warehouses have distinct workload types, data architectures, and query patterns. Traditional databases often utilize normalized schemas to reduce redundancy and ensure data consistency. They are designed for efficient transaction processing and optimized for Online Transaction Processing (OLTP), which prioritizes speed and accuracy when handling large numbers of short transactions. OLTP systems enable real-time processes such as sales recording, hotel reservations, inventory updates, and customer order processing by performing insert, delete, replace, and update actions quickly. Data warehouses, on the other hand, are intended for analytical workloads and frequently employ denormalized, star, or snowflake schemas to improve query performance. They are capable of handling organized, semi-structured, and unstructured data, making them ideal for online analytical processing (OLAP). Unlike OLTP, OLAP is designed to handle complicated queries, aggregations, and

multidimensional analysis, allowing business analysts and data scientists to extract significant insights quickly. OLAP systems may perform analytical queries up to 1000 times faster than OLTP, allowing real-time data exploration from different perspectives for business information and strategic decision-making.

1.4. Data Analysis and Data Analytics

Data in its unprocessed state (raw data) is useless to any business. To transform raw data into relevant insights, various procedures are used, including gathering, cleaning, filtering, sorting, analyzing, and storing. Each phase is critical in ensuring that the data is usable, readable, or visualizable in decision-making formats such as graphs, charts, and texts. The science of data analysis involves analyzing raw data to gain insight, knowledge, and information. As a result, data analysis is the procedure of collecting, organizing, manipulating, pre-processing, handling, transforming, modeling, interpreting, and examining data in order to gain a comprehensive understanding of it or convert it into usable information. Data analytics and data analysis are frequently used interchangeably; however, their meanings might vary based on context. While data analysis is primarily concerned with reviewing and interpreting data, data analytics includes a larger variety of operations, such as data analysis, in order to obtain actionable insights and inform decision-making processes.

Data analytics involves the process of analyzing raw data to identify insights, trends, and patterns that may be used to make decisions. It includes the complete process of gathering, cleaning, processing, analyzing, interpreting, and presenting data in order to provide actionable insights. It focuses on the application of data analysis tools to generate insights and meaningful correlations for decision-making. Data analytics focuses on using current theories or models to draw conclusions that allow businesses to make more informed decisions and drive commercial outcomes.

1.5. Problems and goals of statistical analytics

Data analytics involves extracting information from raw data. Data analytics uses models to find patterns and relationships in data and extract useful information that will benefit the data's users. Thus, data analytics is the process of extracting meaningful information from raw data in order to make decisions (Figure 3.1).

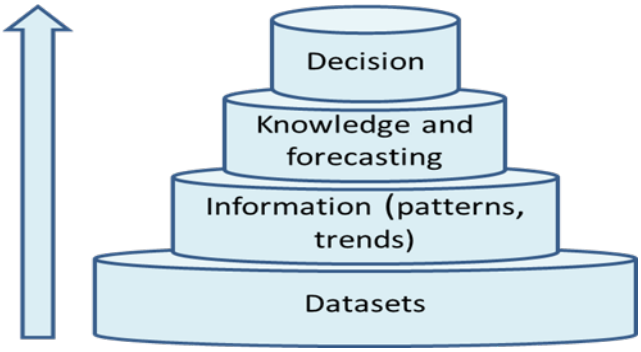

Figure 1.1. Process of extracting useful information from raw data for decision-making.

1.6. Data scientist, data analyst, and data engineer

A data scientist is in charge of collecting, storing, manipulating, and analyzing data, as well as developing novel ways to ask and answer questions. A data scientist's primary responsibility is to extract insights and information from data. They are typically proficient in statistics, mathematics, and programming. Data scientists frequently collaborate with stakeholders to understand business requirements and objectives before generating data-driven queries. A data analyst, on the other hand, is responsible for collecting, cleaning, storing, and organizing data in order to evaluate it and answer business-related issues.

Data analysts examine data, analyze results, and provide insights to help organizations make decisions. They employ statistical methods to interpret patterns, trends, and relationships in data sets. Data analysts frequently work with structured data, conducting tasks such as querying databases, preparing reports, and producing visualizations with tools such as SQL, Excel, and Tableau. Data analysts work with organizations to understand their requirements and translate them into actionable insights. A data engineer not only transfers and processes data into "pipelines," but also designs, creates, and maintains the data infrastructure and architecture of data systems used by team members. They are in charge of the step-by-step design and implementation of pipelines that extract, transform, and load (ETL) data from diverse data sources and store it in data warehouses or other systems. As a knowledge domain, data analytics requires input from a variety of sources.

There is a substantial overlap between the responsibilities of data analyst, data engineer, and data scientist, particularly in firms with a small data team or a requirement for flexibility and versatility. In some smaller businesses, data scientists may wear numerous hats and take on duties spanning data analysis, data engineering, and data science. This adaptability enables them to better meet the demands and constraints of their environment while also delivering value more effectively. As the data team grows and projects get more complicated, organizations can decide to further specialize these roles, but teamwork and cross-functional skills remain critical for success in the data area.

1.7. Who should read this book?

Currently, as the volume of data collected in numerous industries grows, students from various backgrounds are becoming more interested in data analysis techniques. This book was created to present the basic concepts and methodologies for data analysis in an understandable manner. It attempts to be understandable to any university student, regardless of their background. Many real-world examples of data applications and images are provided to help explain the concepts. These examples are intended to help students acquire a better knowledge of data analysis and to inspire them to experiment on their own. Since most applications now need programming, this book chose to use the Python programming language. Jupyter and the Python programming language are introduced in broad strokes. This book was intended. Students studying computer information systems and statistics are especially interested in these topics. It seeks to deliver a new, more user-friendly book on data analysis, though you are not required to have a degree in mathematics or computer science. Mathematics and programming knowledge are certainly useful. Those who have previously studied subjects such as statistics will recognize some of the information presented in this book. Students in computer science will be conversant with pseudocode and Python.

This book additionally provides a simple but comprehensive introduction to students who want to explore this subject in a variety of fields, including biology, social sciences, and engineering, but only have a basic understanding of mathematics, computing, or programming. Students of economics and management can utilize financial data to investigate market patterns, consumer behaviour, and economic indicators. In biology and medicine, data analysis is critical for assessing genomic data, clinical trials, patient records, and medical imaging in order to detect disease patterns, develop individualized treatments, and enhance healthcare results. Engineers utilize

data analysis to optimize processes, track performance, and enhance product design. Environmentalists can use sensor data, monitoring data, and field data to forecast environmental threats and allocate resources. Socialists require examining trends and patterns using survey data, census data, or social media data to better comprehend social dynamics, inequality, and public opinion.

It is believed that after reading this book, you will be well-equipped to handle and carry out analytics projects efficiently, from data preparation, handling, pre-processing, and analysis to solution implementation and maintenance. You should be able to select appropriate approaches for the project and interpret the results in relation to business objectives and the effective implementation of an analytics solution. However, to have a deeper understanding of a topic, you must integrate your knowledge and experience. It sometimes takes several years to obtain expertise on this particular issue, which necessitates learning from mistakes, effort, and patience. Ongoing learning and adaptability to new technologies, methodologies, and business challenges can strengthen your ability to cope with data analytics issues and advance your career.

Exercises

1. Explain in your own words how data interacts with hardware and software components within an information system. Discuss the role of data in processing, storage, and decision-making.
2. Define and differentiate data, information, and knowledge. Provide examples to illustrate how raw data is transformed into meaningful information and ultimately into actionable knowledge.
3. Explain the differences between quantitative and qualitative data, including their characteristics and uses.
4. Provide examples of situations where qualitative data is applicable and discuss why it is important in those contexts.
5. Why is data analysis useful for your work?

References

- https://www.simplilearn.com/data-science-vs-big-data-vs-data-analytics-article
- Burgener, E., Rydning, J., 2022. High Data Growth and Modern Applications Drive New Storage Requirements in Digitally Transformed Enterprises.

- https://www.delltechnologies.com/asset/en-my/products/storage/industry-market/h19267-wp-idc-storage-reqs-digital-enterprise.pdf

CHAPTER 2

DATA TYPES, DATA TOPOLOGIES, AND STATISTICAL PROPERTIES AND METHODS

2.1. Introduction

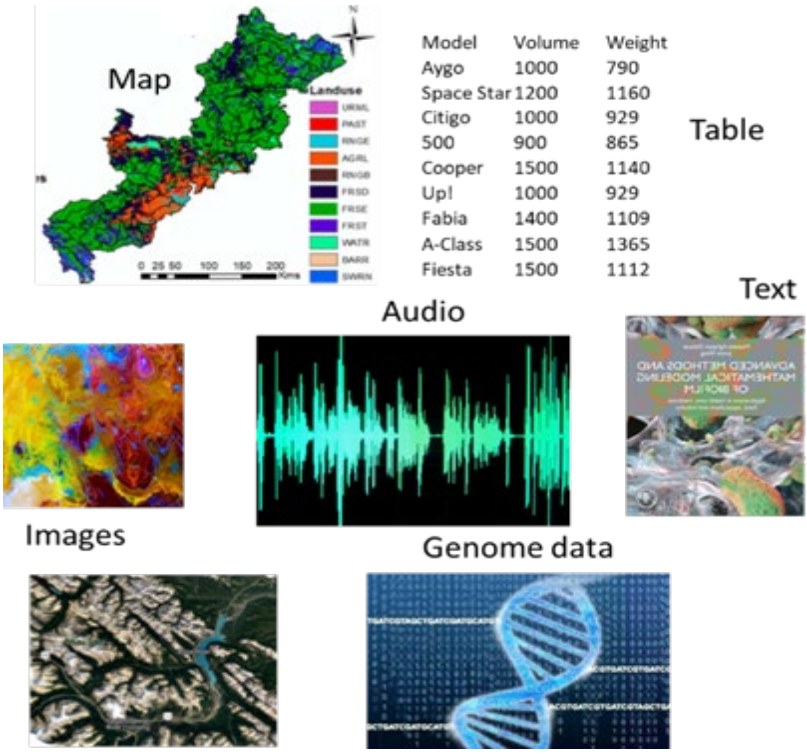

Figure 2.1. Various data types

A data type is a classification of data that describes what kind of value a variable can have and what operations can be applied to it. In the real world, various types of data exist, including numbers, tables, text, photos, video, and audio. Data can be structured, yet much of it may be unstructured. Data types are critical for organizing data into classes for retrieval, sorting, and storage, as well as assisting organizations in manipulating, tracking, and analyzing data. Understanding the types of data and their classification enables data to be collected most appropriately for a given aim. The goal determines the approach they use and the classification levels they employ.

Data is a set of measurements or observations. Data can be collected through a variety of methods, including measurement, monitoring, questionnaires, and sampling. Nowadays, the definition of data gathering extends beyond the data sources used in computer applications, such as web, files, photos, and videos.

2.2. Data types

When data analysis is used, the data must represent how a variable's values are measured as well as its characteristics. Understanding the mechanism of the underlying measurement scale is critical for selecting appropriate statistical studies and interpreting the results correctly. Different statistical methods apply to different types of data, and applying the wrong method can result in inaccurate results.

Data types are important in programming languages because they define what sort of data can be stored in a variable and what actions can be performed on that data. They ensure that activities are completed appropriately and avoid errors caused by mismatched types or processes. Each computer language has a unique set of data types, as well as specific rules and procedures that instruct the compiler or interpreter on how to use the data. Data types can be categorized via a variety of approaches, including qualitative and quantitative data, dependent and independent data, structured, semi-structured, and unstructured data (Figure 2.2).

Most programming languages do offer common data types, including integers, floating-point numbers (real), characters, strings, and Booleans. While the particular syntax or implementation details may differ between languages, the way things work and the use of certain common data types is generally constant across computer languages. Python's data types include integer, floating point, text, Boolean, and bytes.

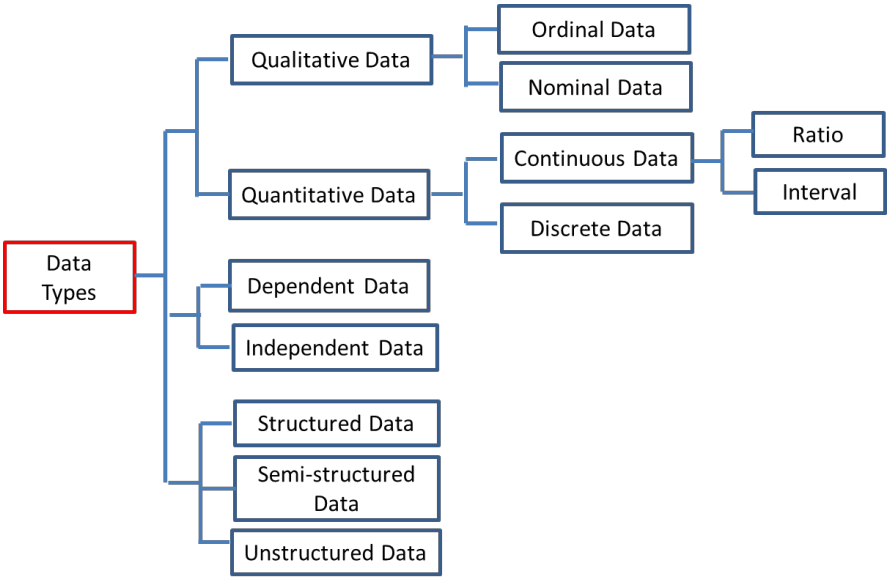

Figure 2.2. Data or variable types

2.2.1. Qualitative and quantitative data

Qualitative data describes information and concepts that cannot be expressed numerically. It is used to describe the qualities or characteristics of objects, people, events, or phenomena. It provides explanations for the "why" and "how" of a specific issue and is frequently utilized to investigate meanings, attitudes, beliefs, and personal experiences. Qualitative data is information on qualities or information that cannot be measured and can be expressed through a name, symbol, or number code as measures of 'types'. Qualitative data is typically obtained using methods such as interviews, observations, or surveys.

There are two forms of qualitative data: nominal and ordinal. Nominal data, also known as a nominal scale, refers to data used to classify variables or represent different and mutually exclusive groups or labels. Nominal data cannot be organized or measured without a numeric value. They lack inherent order or ranking. Nominal data is frequently used to divide items into groups or categories based on a specific trait or attribute. For example, you can be married or unmarried, masculine or female, pregnant or not pregnant. You can identify as an ethnicity (e.g., black, Hispanic, Asian,

white, other), a political party (e.g., liberty, conservative, other), or an animal kind (e.g., cat, dog, cow).

Ordinal data classifies data according to an order or ranking. Ordinal data, as opposed to nominal data, categorizes and ranks the categories or labels in terms of magnitude or degree. While the intervals between the categories may not be equal or quantitative, they do follow a distinct sequence or ranking. This provides for a more precise degree of measurement than nominal data. You can have both larger and smaller amounts. You can use the hierarchy to assess economic status, for example, rich, middle-income, or poor. You can also rank cancer stages (e.g., stage I, stage II, stage III) and pain levels as mild, moderate, or severe. As a result, by capturing the concept of higher or lower levels within a particular feature, ordinal variables enable a more comprehensive understanding of category relationships. Nevertheless, should be emphasized that the exact difference or distance between categories is not necessarily uniform or precisely quantifiable, distinguishing ordinal data from interval or ratio data.

Quantitative data consists of numbers that may be measured or compared on a numerical scale. It is often utilized in mathematical operations such as addition and subtraction, and statistical analysis. One can add and split them. Quantitative data includes measurements like length, weight, temperature, time, and numerical counts. Quantitative data is divided into discrete and continuous categories (Figure 2.2). Discrete data is made up of unique and separate values that are typically collected by counting. These values are usually numerical and have defined, fixed values. Continuous data has an infinite number of possible values within a defined range and can be any value, such as height, weight, temperature, or length.

Complex numbers, as well as any value inside a given range, including fractions and decimals, are examples of continuous data. Time, distance, temperature, and weight are common metrics for continuous data. Continuous data includes interval and ratio data. Interval data is data in which variations between values are meaningful and constant across the whole scale. Equal intervals on the scale indicate equal differences in the underlying variable. However, there is no actual zero point in interval data. A value of zero represents a point on the scale rather than the absence of the variable. In contrast, ratio data exhibits all of the features of interval data. It also has a genuine zero point, which symbolizes the lack of a variable. This enables the calculation of ratio data, resulting in proper expressions like "twice as much" or "half as much." Some examples of ratio variables are lengths measured in meters or inches, blood pressure measured in millimetres of mercury, time expressed in seconds or hours, and common mass and object counts.

2.2.2. Independent vs. Dependent data

Independent data is one that do not rely on one another. When the values of one variable reveal nothing about the values of the other variables, the variables are said to be independent. For example, the dataset 'diet of students' is not dependent on the variable 'diet of teachers'. On the contrary, the dependent variable's value is determined by the independent variable. The dependent variable is measured in reaction to changes in the independent variable. The dependent variables' responses to changes in the independent variables reveal what happens to the independent variables. For example, animal growth is dependent on nutrition. The dependent variable would be animal weight gain, which is determined by nutrition absorption (the independent variable). The link between the variables is frequently investigated using experiments, observations, or statistical studies to determine how one variable affects the other.

2.2.3. Structured data, Semi-structured data, and unstructured data

Structured data is well-organized and can be simply entered into a prepared database. It is data that is quantitative and understood by both machines and people. They have a clear structure and format, making them ideal for storing in designated fields. Their architectural data is specific and well-organized, making it simple to manipulate, extract, and query for accurate analysis. This systematic approach to data organization enables efficient data storage and manipulation, which is critical in many applications, particularly database management, data analysis, and software development. Structured query language (SQL) is used in SQL databases to modify data. Organizations may quickly enter, delete, sort, and alter structured data in the form of rows and columns in a table.

Semi-structured data combines elements of both structured and unstructured data. Semi-structured data does not follow the format of a tabular data model or another rigorous, predetermined data model, but structured data does. Instead, it contains structure while still allowing for flexibility and unpredictability in how the data is structured. This flexibility is frequently derived from the use of tags, markers, or other identifiers that provide some level of structure to the data, making it easier to comprehend and process than unstructured data. Flexibility provides features that make it easier to distinguish semantic parts and impose data ranking within a dataset when compared to structured data. Semi-structured data, on the other hand, frequently incorporates qualitative features that enable it to be machine-

readable or human-interpretable. It serves as a link between structured and unstructured data by retaining some order while allowing for flexibility and diversity in structure.

Unstructured data is data that is not organized in a preset trend or model, such as written documents, photos, audio and video, social media postings, emails, or sensor data. This trait complicates typical analysis approaches. Organizations have both obstacles and possibilities when dealing with unstructured data. On the one hand, the absence of order in unstructured data makes it difficult to store, manage, and analyze. On the other side, studying unstructured data can provide significant insights and patterns that structured data alone may not reveal, giving firms a competitive edge. Structured data fits neatly into predetermined categories and is readily managed and processed by machines, whereas unstructured data lacks a definite structure or format. As a result, qualitative data is commonly used to describe unstructured data. It cannot be handled or evaluated traditionally.

As a result, unstructured data is increasing quickly as the internet and media create massive volumes of information from a variety of sources. These forms of data frequently include significant insights, but retrieving useful information from them necessitates more advanced techniques, such as natural language processing, picture recognition, or artificial intelligence. As a result, there is an increasing interest in technology and methodologies that can efficiently manage and extract value from unstructured data. Table 2.1 illustrates the advantages and disadvantages of organized, semi-structured, and unstructured data.

Table 2.1. Advantages and disadvantages of structured, semi-structured and unstructured data

	Advantages	Disadvantages	Examples
Structured data	• Easily used, crawled, and manipulated by machine learning (ML) due to simplified querying algorithms • Easy to access and by more tools and users • Easy to analyze and interpret	• Limited usability due to a predefined structure for its intended purpose • Limited storage options due to their rigid schemas (e.g., "data warehouses"). • Low scalability and flexibility due to difficulty in changing format	Tables, SQL database
Semi-structured data	• Flexible because the schema can be easily changed • Highly storable and portable to collate, query, and analyze heterogeneity of sources. • Scalable due to its easy fitting into structured data tools and the tools that manage unstructured data. It is possible to view structured data as semi-structured data.	• More challenging to analyze, interpret, or categorize • Higher cost of storage • More difficult to query • Lower reliability because it is not as well-organized.	XML data, HTML code, graphs and tables, e-mails, CSV documents.
Unstructured data	• Undefined or native data format until needed • Massive storage and eases scalability • Adaptability to enable data scientists to analyze only the data they need • Quick and easy data collection or accumulation • Better insights and more opportunities to turn your data into a competitive advantage	• Require expertise in data science due to its undefined/non-formatted nature • Require specialized tools to manipulate unstructured data • Time-consuming and expensive to identify, extract, and analyze. • Require huge storage capacity due to its enormous size and greater number of formats	Word, PDF, Text, audio, videos, and media logs

2.3. Data statistical properties

In data analysis, many essential statistical features are critical for understanding and interpreting information. These features aid in data summarization, pattern recognition, and informed decision-making. Some of the most essential statistical qualities are distribution, mean and median, variability, minimal mean square error, central tendency, and computational ease.

2.3.1. Sample distributions

A probability distribution applies to a certain population. The sampling distribution is a frequency distribution formed from all possible random samples of a population across a certain range. Figure 2.3 shows a normal or Gaussian distribution. You can see that 60% of the population spans -1 to 1, with just 2.1% beyond -2 and 2.

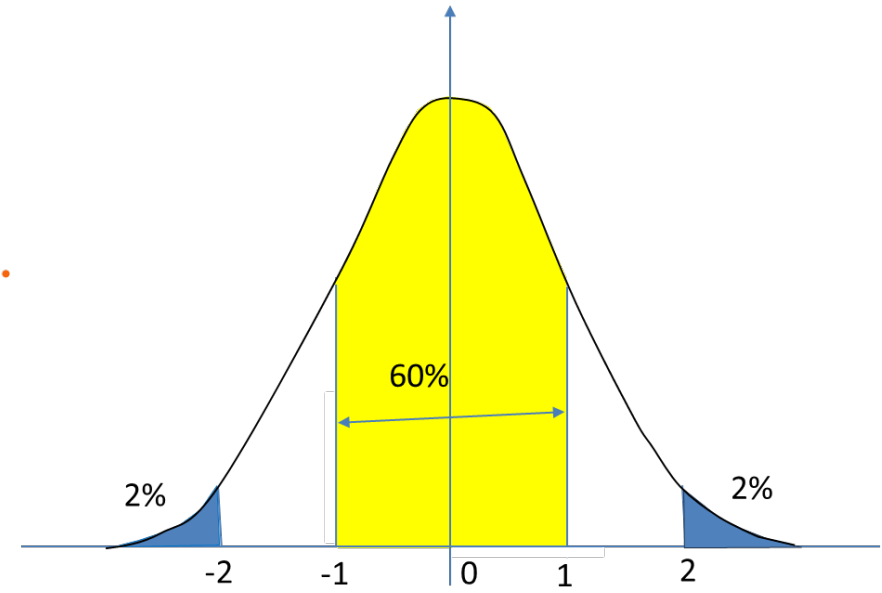

Figure 2.3. Normal or Gaussian Distribution of a population