Mathematical Foundations of Data Science

Introduction to Data Science

Satishkumar L. Varma

Professor, Department of Information Technology
PCE, New Panvel
Scopus | Web of Science | Google Scholar | Google Site | Website

Introduction to Data Science: Outline

- Machine Learning
  - Supervised and Unsupervised Learning,
  - Concepts of Classification,
  - Clustering and prediction
- Introduction to Data Science;
  - Importance of linear algebra from a data science perspective;
  - Importance of statistics and optimization from a data science perspective;
  - Structured thinking for solving data science problems;
  - Probability, Statistics and Random Processes:
    - Probability theory and axioms;
    - Random variables.

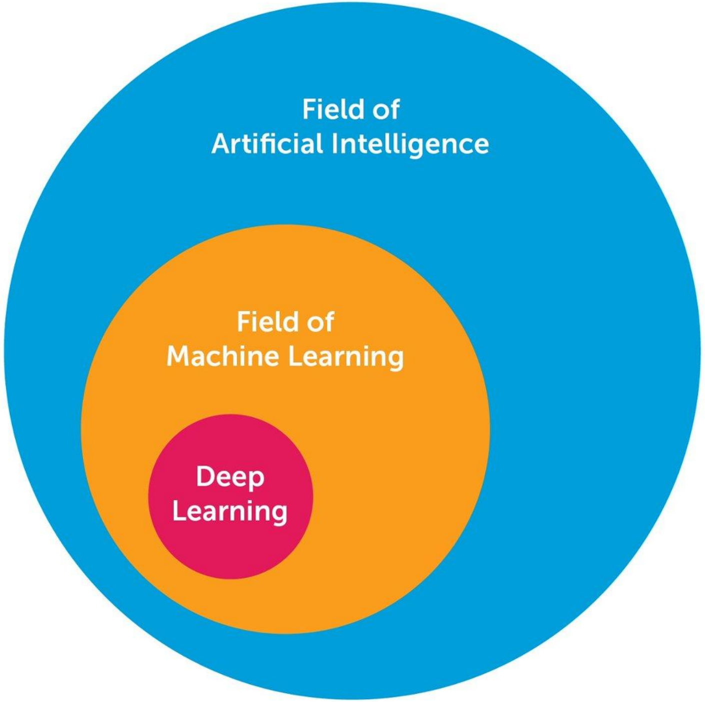
"Well planned milestone is a key to success."

Introduction to Data Science: Learning Objectives & Outcomes

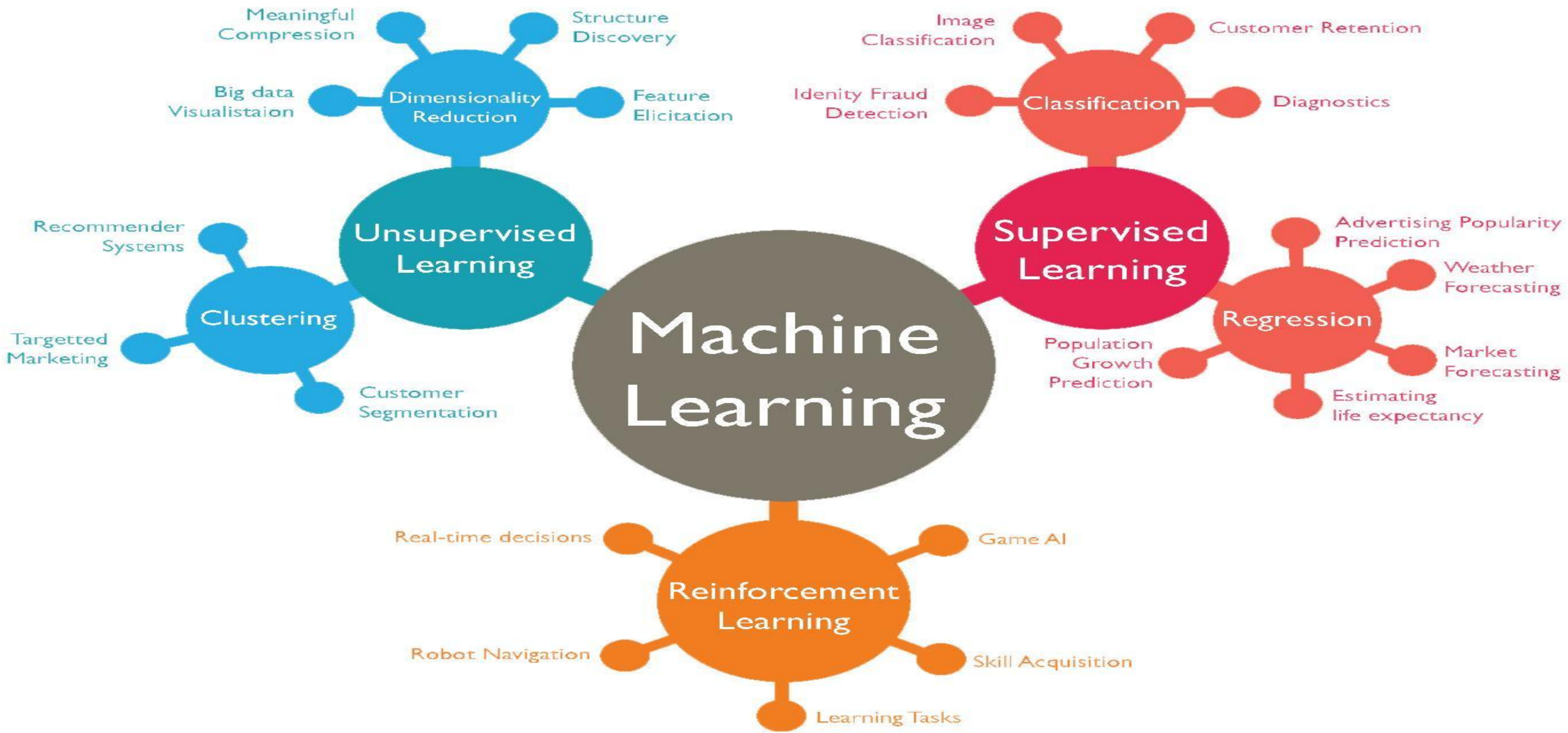
- **Learning Objectives:** Course Instructor or Faculty aims
  - To

Introduction to Data Science: Introduction

- Data Science: Interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data
  - Artificial intelligence
  - Machine learning
  - Deep learning



Introduction to Data Science: Introduction



https://sites.google.com/view/vsat2k/courses/AI-and-Data-Analytics

Introduction to Data Science: Introduction

- Supervised and Unsupervised learning are the two techniques of machine learning.
- But both the techniques are used in different scenarios and with different datasets.
- Supervised learning is classified into two categories of algorithms:
  - **Classification**:
    - A classification problem is when the output variable is a category, such as “Red” or “blue” , “disease” or “no disease”.
  - **Regression**:
    - A regression problem is when the output variable is a real value, such as “dollars” or “weight”.
- Supervised learning deals with or learns with “labeled” data.
- This implies that some data is already tagged with the correct answer.

Introduction to Data Science: Supervised Learning

- Unsupervised learning is the training of a machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance.
- Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.
- Unlike supervised learning, no teacher is provided that means no training will be given to the machine.
- Therefore the machine is restricted to find the hidden structure in unlabeled data by itself.
- Supervised learning Types:
    - Regression
    - Logistic Regression
    - Classification
    - Naive Bayes Classifiers
    - K-NN (k nearest neighbors)
    - Decision Trees
    - Support Vector Machine

Introduction to Data Science: Supervised Learning

- Advantages:
  - Supervised learning allows collecting data and produces data output from previous experiences.
  - Helps to optimize performance criteria with the help of experience.
  - Supervised machine learning helps to solve various types of real-world computation problems.
  - It performs classification and regression tasks.
  - It allows estimating or mapping the result to a new sample.
  - We have complete control over choosing the number of classes we want in the training data.
- Disadvantages:
  - Classifying big data can be challenging.
  - Training for supervised learning needs a lot of computation time. So, it requires a lot of time.
  - Supervised learning cannot handle all complex tasks in Machine Learning.
  - Computation time is vast for supervised learning.
  - It requires a labelled data set.
  - It requires a training process.

Introduction to Data Science: Unsupervised Learning

- **Advantages of unsupervised learning:**
- It does not require training data to be labeled.
- Dimensionality reduction can be easily accomplished using unsupervised learning.
- Capable of finding previously unknown patterns in data.
- **Flexibility**: Unsupervised learning is flexible in that it can be applied to a wide variety of problems, including clustering, anomaly detection, and association rule mining.
- **Exploration**: Unsupervised learning allows for the exploration of data and the discovery of novel and potentially useful patterns that may not be apparent from the outset.
- **Low cost**: Unsupervised learning is often less expensive than supervised learning because it doesn’t require labeled data, which can be time-consuming and costly to obtain.

Introduction to Data Science: Unsupervised Learning

- **Disadvantages of unsupervised learning :**
- Difficult to measure accuracy or effectiveness due to lack of predefined answers during training.
- The results often have lesser accuracy.
- The user needs to spend time interpreting and label the classes which follow that classification.
- **Lack of guidance:** Unsupervised learning lacks the guidance and feedback provided by labeled data, which can make it difficult to know whether the discovered patterns are relevant or useful.
- **Sensitivity to data quality**: Unsupervised learning can be sensitive to data quality, including missing values, outliers, and noisy data.
- **Scalability**: Unsupervised learning can be computationally expensive, particularly for large datasets or complex algorithms, which can limit its scalability.

Introduction to Data Science: Unsupervised Learning

- Unsupervised learning is classified into two categories of algorithms:
- **Clustering**: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
- **Association**: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.
- Types of Unsupervised Learning:
- Clustering
  - Exclusive (partitioning)
  - Agglomerative
  - Overlapping
  - Probabilistic
- Clustering Types:
  - Hierarchical clustering
  - K-means clustering
  - Principal Component Analysis
  - Singular Value Decomposition
  - Independent Component Analysis

Introduction to Data Science: Unsupervised Learning

- Supervised vs. Unsupervised Machine Learning

Parameters	Supervised machine learning	Unsupervised machine learning
Input Data	Algorithms are trained using labeled data.	Algorithms are used against data that is not labeled
Computational Complexity	Simpler method	Computationally complex
Accuracy	Highly accurate	Less accurate
No. of classes	No. of classes is known	No. of classes is not known
Data Analysis	Uses offline analysis	Uses real-time analysis of data
Algorithms used	Linear and Logistics regression, Random forest, Support Vector Machine, Neural Network, etc.	K-Means clustering, Hierarchical clustering, Apriori algorithm, etc.

Introduction to Data Science: Unsupervised Learning

- Supervised vs. Unsupervised Machine Learning

Parameters	**Supervised machine learning**	**Unsupervised machine learning**
Output	Desired output is given.	Desired output is not given.
Training data	Use training data to infer model.	No training data is used.
Complex model	It is not possible to learn larger and more complex models than with supervised learning.	It is possible to learn larger and more complex models with unsupervised learning.
Model	We can test our model.	We can not test our model.
Called as	Supervised learning is also called classification.	Unsupervised learning is also called clustering.
Example	Example: Optical character recognition.	Example: Find a face in an image.

Introduction to Data Science: Introduction

- Importance of Linear Algebra in Data Science
- Understanding linear algebra is key to becoming a skilled data scientist.
- Linear algebra is important in data science because of the following reasons:
  - It helps in organizing and manipulating large data sets with efficiency.
  - Many data science algorithms rely on linear algebra to work fast and accurately.
  - It supports major machine learning techniques, like regression and classification.
  - Techniques like Principal Component Analysis for reducing data dimensionality depend on it.
  - Linear algebra is used to alter and analyze images and signals.
  - It solves optimization problems, helping find the best solutions in complex data scenarios.

Introduction to Data Science: Introduction

- Key Concepts in Linear Algebra
- Linear algebra is a branch of mathematics useful for understanding and working with arrays of numbers known as matrices and vectors.

Key Concepts in Linear Algebra	Description
Vectors	Fundamental entities in linear algebra representing quantities with both magnitude and direction, used extensively to model data in data science.
Matrices	Rectangular arrays of numbers, which are essential for representing and manipulating data sets.
Matrix Operations	Operations such as addition, subtraction, multiplication, and inversion that are crucial for various data transformations and algorithms.
Eigenvalues and Eigenvectors	These are used to understand data distributions and are crucial in methods such as Principal Component Analysis (PCA) which reduces dimensionality.
Singular Value Decomposition (SVD)	A method for decomposing a matrix into singular values and vectors, useful for noise reduction and data compression in data science.
Principal Component Analysis (PCA)	A statistical technique that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables.

Introduction to Data Science: Introduction

- Applications of Linear Algebra in Data Science
- Linear Algebra turns complex problems into manageable solutions. Here are some of the most common applications of linear algebra in data science:
- Machine Learning Algorithms
- Linear algebra is vital for machine learning. It helps in creating and training models. For instance, in regression analysis, matrices represent data sets. This simplifies calculations across vast numbers of data points.
- Image Processing
- In image processing, linear algebra streamlines tasks like scaling and rotating images. Matrices represent images as arrays of pixel values. This representation helps in transforming the images efficiently.
- Natural Language Processing (NLP)
- NLP uses vectors to represent words. This technique is known as word embedding. Vectors help in modeling word relationships and meanings. For example, vector space models can determine synonyms based on proximity.

Introduction to Data Science: Introduction

- Applications of Linear Algebra in Data Science
- Data Fitting and Predictions
- Linear algebra is used to fit data into models. This process predicts future trends from past data. Least squares, a method that minimizes the difference between observed and predicted values, relies heavily on matrix operations.
- Network Analysis
- In network analysis, matrices store and manage data about connections. For instance, adjacency matrices can represent social networks. They show connections between persons or items, aiding in understanding network structures.
- Optimization Problems
- Linear algebra solves optimization problems in data science. It helps find values that minimize or maximize some function. Linear programming problems often use matrix notations for constraints and objectives, streamlining the solution process.

Introduction to Data Science: Introduction

- Advanced Techniques in Linear Algebra for Data Science
- Some techniques in linear algebra can be applied to solve complex and high-dimensional data problems effectively in data science. Some of the advanced Techniques in Linear Algebra for Data Science are :
    - Singular Value Decomposition (SVD)
    - Principal Component Analysis (PCA)
    - Tensor Decompositions
    - Conjugate Gradient Method

Introduction to Data Science: Introduction

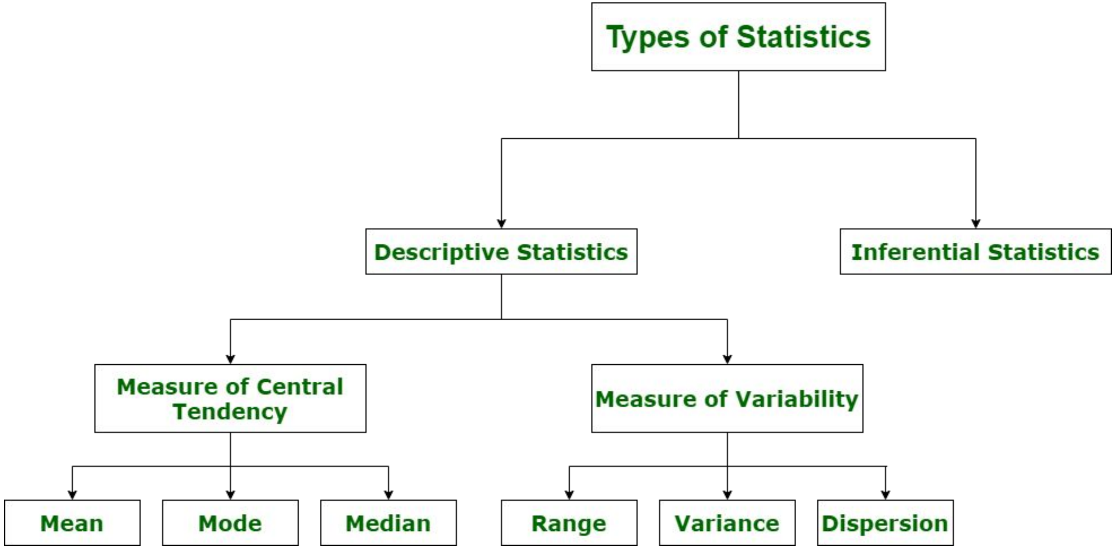
- What are Statistics?
- Statistics in Mathematics is the study and manipulation of data. It involves the analysis of numerical data, enabling the extraction of meaningful conclusions from the collected and analyzed data sets.
- According to Merriam-Webster: Statistics is the science of collecting, analyzing, interpreting, and presenting masses of numerical data.
- According to Oxford English Dictionary: Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, presentation, and organization of data.
- Statistics
- The practice or science of collecting and analysing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.

Introduction to Data Science: Introduction

- Statistics Terminologies: Some of the most common terms you might come across in statistics are:
- Population: It is actually a collection of a set of individual objects or events whose properties are to be analyzed.
- Sample: It is the subset of a population.
- Variable: It is a characteristic that can have different values.
- Parameter: It is numerical characteristic of population.
- Statistics Examples: Some real-life examples of statistics that you might have seen:
  - Example 1:
    - In a class of 45 students, we calculate their mean marks to evaluate performance of that class.
  - Example 2:
    - Before elections, you might have seen exit polls. Exit polls are opinion of population sample, that are used to predict election results.

Introduction to Data Science: Introduction

- Types of Statistics
- There are 2 types of statistics:
- Descriptive Statistics
- Inferential Statistics



Introduction to Data Science: Introduction

- Descriptive Statistics:
    - It uses data that provides a description of the population either through numerical calculated graphs or tables.
    - It provides a graphical summary of data.
    - It is simply used for summarizing objects, etc. There are two categories in this as follows.
        - Measure of Central Tendency
        - Measure of Variability
- Measure of Central Tendency:
    - It is also known as summary statistics that are used to represent the center point or a particular value of a data set or sample set.
    - In statistics, there are three common measures of central tendency that are:
    - Mean
    - Median
    - Mode

Introduction to Data Science: Introduction

- Measure of Variability: It is also known as the measure of dispersion and is used to describe variability in a sample or population.
- In statistics, there are three common measures of variability as shown below:
- 1. Range of Data: It is a given measure of how to spread apart values in a sample set or data set.
- Range = Maximum value – Minimum value
- 2. Variance: In probability theory and statistics, variance measures a data set’s spread or dispersion.
  - It is calculated by averaging the squared deviations from the mean. Variance is usually represented by the symbol $\sigma 2$.
  - $S2= \sum ni=1 \; [(xi - \overline{x})2 \;/\; n]$
  - n represents total data points
  - $\overline{x}$ represents the mean of data points
- xi represents individual data points

Introduction to Data Science: Introduction

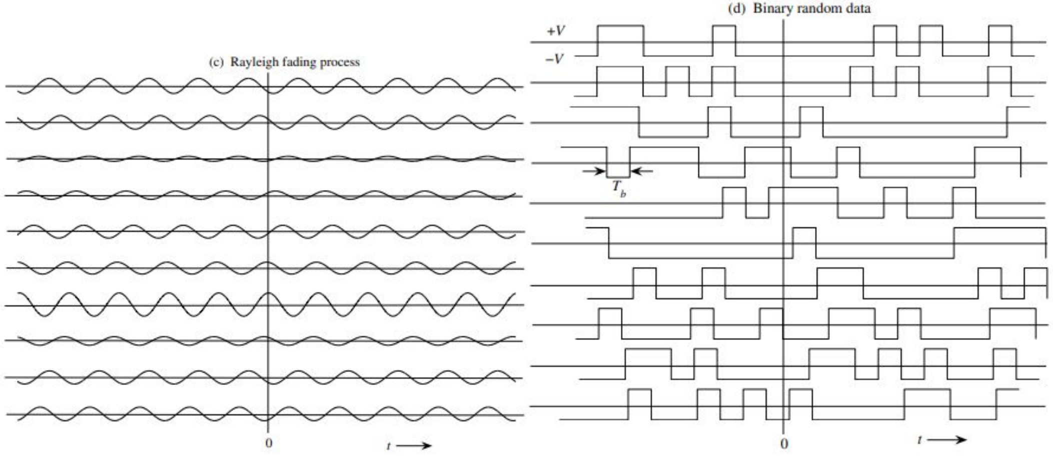
- Measure of Variability
- Variance measures variability.
  - The more spread out the data, the greater the variance compared to the average.
  - There are two types of variance:
  - Population variance: Often represented as σ^2
  - Sample variance: Often represented as s^2.
- Note: The standard deviation is the square root of the variance.
- Dispersion: It is the measure of the dispersion of a set of data from its mean.
  - $\sigma= \surd\ (1/n) \sum ni=1\ (xi - \mu)2$

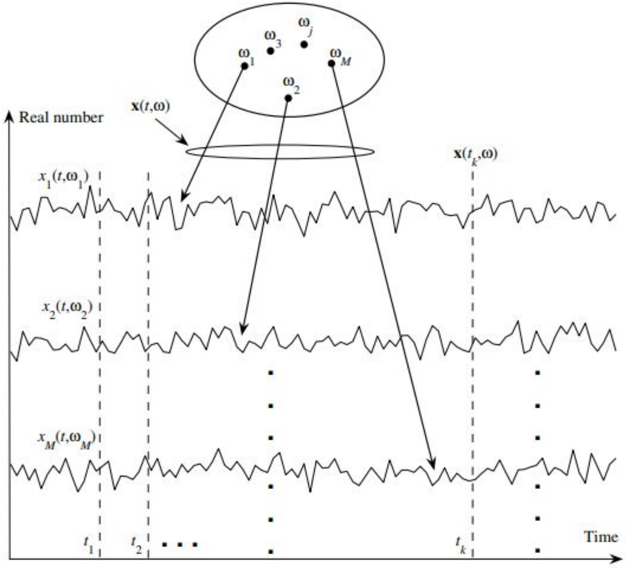
Introduction to Data Science: Introduction

- The main objective of a communication system is the transfer of information over a channel.
- Message signal is best modeled by a random signal Two types of imperfections in a communication channel:
- **Deterministic imperfection:**
  - such as linear and nonlinear distortions, inter-symbol interference, etc.
- **Nondeterministic imperfection:**
  - such as addition of noise, interference, multipath fading, etc.
- We are concerned with the methods used to describe and characterize a random signal, generally referred to as a random process (also commonly called stochastic process).
- In essence, a random process is a random variable evolving in time.

Introduction to Data Science: Introduction

- Random process (also commonly called stochastic process).
  - In essence, a random process is a random variable evolving in time.
- Classification of Random Processes Based on whether its statistics change with time:
  - the process is non-stationary or stationery.





Introduction to Data Science: Introduction

- Sample Space and Probability
- **Random experiment:** its outcome, for some reason, cannot be predicted with certainty.
  - **Examples:** throwing a die, flipping a coin and drawing a card from a deck.
- Sample space: the set of all possible outcomes, denoted by **Ω**.
  - Outcomes are denoted by ω's and each ω lies in **Ω**, i.e., $\omega \in \Omega$.
  - A sample space can be discrete or continuous.
- Events are subsets of the sample space for which measures of their occurrences, called probabilities, can be defined or determined.
- The events E1, E2, E3,. . . are mutually exclusive if $Ei \cap Ej = \varnothing$ for all $i \neq j$, where null ($\varnothing$) is the null set.

Introduction to Data Science: Introduction

- Three Axioms of Probability
- For a discrete sample space Ω, define a probability measure P on Ω as a set function that assigns nonnegative values to all events, denoted by E, in Ω such that the following conditions are satisfied
- Axiom 1:
  - ○ $0 \leq P(E) \leq 1$ for all $E \in \Omega$ (on a % scale probability ranges from 0 to 100%.
  - ○ Despite popular sports lore, it is impossible to give more than 100%).
- Axiom 2:
  - ○ $P(\Omega) = 1$
  - ○ when an experiment is conducted there has to be an outcome.
- Axiom 3:
  - ○ For mutually exclusive events E1, E2, E3,. . . we have
  - ○ P(∪ ∞ i=1 Ei) = Σ∞ i=1 P(Ei).

Introduction to Data Science: Introduction

- Probability theory and random variables form the foundation of statistical analysis.
- These concepts help us understand and quantify uncertainty in various scenarios, from coin tosses to complex scientific experiments. They're essential tools for making sense of random phenomena in our world.
- Probability theory is a branch of mathematics that deals with the analysis of random phenomena and the likelihood of events occurring
- A sample space is the set of all possible outcomes of a random experiment or process, typically denoted by the symbol Ω (omega)
- An event is a subset of the sample space, representing a collection of outcomes that satisfy a specific condition or property, typically denoted by capital letters (A, B, C)
- Complement of an event A, denoted as A^c or A', is the set of all outcomes in the sample space that are not in A
- The probability of the complement is given by
- P(Ac) = 1 − P(A)

Introduction to Data Science: Probability axioms and rules

- Probability axioms are the fundamental rules that govern the assignment of probabilities to events in a sample space
  - Axiom 1 (Non-negativity): The probability of any event A is always non-negative, i.e., $P(A) \geq 0$
  - Axiom 2 (Normalization): The probability of the entire sample space Ω is equal to 1, i.e., $P(\Omega)=1$
  - Axiom 3 (Additivity): For any two mutually exclusive events A and B, the probability of their union is equal to the sum of their individual probabilities, i.e., $P(A \cup B)=P(A)+P(B)$
- The addition rule states that for any two events A and B, the probability of their union (A or B occurring) is given by
- $P(A \cup B)=P(A)+P(B)-P(A \cap B)$, where
- $P(A \cap B)$ is the probability of the intersection (both A and B occurring)
  - For mutually exclusive events A and B, the probability of their union simplifies to
  - $P(A \cup B)=P(A)+P(B)$ since $P(A \cap B)=0$
- The multiplication rule states that for any two events A and B, the probability of their intersection (both A and B occurring) is given by
- $P(A \cap B)=P(A) \times P(B|A)$, where
- $P(B|A)$ is the **conditional probability** of B given that A has occurred
  - For **independent events** A and B, the probability of their intersection simplifies to
  - $P(A \cap B)=P(A) \times P(B)$ since
  - $P(B|A)=P(B)$ and $P(A|B)=P(A)$

Introduction to Data Science: Random variable fundamentals

- A **random variable** is a function that assigns a numerical value to each **outcome** in a sample space, serving as a mathematical abstraction used to quantify and analyze random phenomena
- Random variables can be classified into two main types: discrete random variables and continuous random variables
- The **probability distribution** of a random variable describes the likelihood of the random variable taking on different values, assigning probabilities to the possible values or ranges of values that the random variable can assume
- Random variables are used to model and analyze various real-world phenomena (outcome of a coin toss, number of defective items in a production process, time between arrivals in a queueing system)

Properties and measures of random variables

The **expected value** (or mean) of a random variable is a measure of its central tendency, representing the average value of the random variable over a large number of trials or observations, denoted by

- $E(X)$ for a random variable X
- The **variance** and **standard deviation** of a random variable measure the dispersion or spread of its values around the mean
  - Variance is denoted by $Var(X)$ and standard deviation by $\sigma(X)$ for a random variable X
- Other important properties of random variables include moments (measures of the shape and characteristics of the probability distribution) and moment-generating functions (used to uniquely characterize a probability distribution)

Introduction to Data Science: Discrete vs continuous distributions

Discrete random variables and probability mass functions

- Discrete random variables have a countable set of possible values, typically integers or a finite set of numbers (number of heads in a fixed number of coin tosses, number of customers arriving at a store in a given hour)
- The **probability mass function** (PMF) is used to describe the probability distribution of a **discrete random variable**, assigning probabilities to each possible value of the random variable
  - The sum of all probabilities in a PMF is equal to 1
- Examples of discrete probability distributions include the **Bernoulli distribution** (binary outcomes), **binomial distribution** (number of successes in a fixed number of trials), and **Poisson distribution** (number of events occurring in a fixed interval of time or space)

Continuous random variables and probability density functions

- Continuous random variables can take on any value within a specified range or interval, having an uncountable and infinite set of possible values (height of a randomly selected person, time until a radioactive particle decays)
- The **probability density function** (PDF) is used to describe the probability distribution of a **continuous random variable**, representing the relative likelihood of the random variable taking on a specific value within a given range
  - The area under the PDF curve over a specific range gives the probability of the random variable falling within that range
- The **cumulative distribution function** (CDF) is defined for both discrete and continuous random variables, giving the probability that the random variable takes on a value less than or equal to a specific value
  - The CDF is non-decreasing and ranges from 0 to 1
- Examples of continuous probability distributions include the **uniform distribution** (equal probability over a specified range), **normal distribution** (bell-shaped curve), and **exponential distribution** (time between events in a Poisson process)

Introduction to Data Science: Probability calculation rules

Conditional probability and independence

- Conditional probability is the probability of an event occurring given that another event has already occurred, denoted by
- $P(A|B)$ and calculated using the formula
- $P(A|B)=P(A∩B)/P(B)$, where
- $P(B)>0$
- Two events A and B are considered independent if the occurrence of one event does not affect the probability of the other event occurring
  - For independent events,
  - $P(A|B)=P(A)$ and $P(B|A)=P(B)$
  - The probability of the intersection of independent events is the product of their individual probabilities, i.e.,
  - $P(A∩B)=P(A)\times P(B)$

Introduction to Data Science: Probability calculation rules

Law of total probability and Bayes' theorem

- The law of total probability states that for a partition of the sample space into mutually exclusive and exhaustive events
- $B1,B2,\ldots,Bn$, the probability of an event A can be calculated as
- P(A)=P(A∣B1)P(B1)+P(A∣B2)P(B2)+...+P(A∣Bn)P(Bn)
  - This law is useful when calculating probabilities in situations where the sample space can be divided into smaller, more manageable events
- **Bayes' theorem** is used to calculate the conditional probability of an event A given event B, using the conditional probability of B given A and the individual probabilities of A and B
  - It is expressed as
  - $P(A \mid B)=P(B \mid A)P(A)/P(B)$, where
  - $P(B)>0$
  - Bayes' theorem is particularly useful in updating probabilities based on new information or evidence (diagnostic testing, machine learning, spam email filtering)

Introduction to Data Science: Applications

- Real-world Applications of Data Science
    - 1. In Search Engines
    - 2. In Transport
    - 3. In Finance
    - 4. In E-Commerce
    - 5. In Health Care
    - 6. Image Recognition
    - 7. Targeting Recommendation
    - 8. Airline Routing Planning
    - 9. Data Science in Gaming
    - 10. Medicine and Drug Development
    - 11. In Delivery Logistics
    - 12. Autocomplete

Summary

- The supervised and unsupervised learning both are the machine learning methods, and selection of any of these learning depends on the factors related to the structure and volume of your dataset and the use cases of the problem.
- Data science is used for a wide range of applications, including predictive analytics, machine learning, data visualization, recommendation systems, fraud detection, sentiment analysis, and decision-making in various industries like healthcare, finance, marketing, and technology.

References

A. Text Books:

1. Stuart J. Russell and Peter Norvig, "AI A Modern Approach — 2ed Pearson Education.
2. Elaine Rich and Kevin Knight —AI. Third Edition, TMH Pvt. Ltd., 2008.
3. George F Luger “AI” Low Price Edition, Pearson Education, Fourth edition.
4. Alex Holmes “Hadoop in Practice”, Manning Press, DreamTech Press.

B. References:

5. Anil Sharma, Introduction to AI and Expert Systems. LPU, Excel Books Pvt. Ltd.
6. Peter J.F. Lucas & Linda C. van der Gaag, Principles of Expert Systems. Addison-Wesley, 2014.
7. Chuck Lam, “Hadoop in Action”, Dreamtech Press
8. Phil Simon, “Too Big To Ignore: The Business Case For Big Data”, Wiley India

Thank You.