DIFFERENTTYPES OF DISTANCES

A dissertation submitted to

MAHATMA GANDHI UNIVERSITY, KOTTAYAM

In partial fulfilment of the requirement for the award of

BACHELOR OF SCIENCE IN MATHEMATICS

Submitted By

Athulya T R	No: 220021030710
Lakshmi A J	No: 220021030713
Shifa Thasnim M A	No: 220021030717

DEPARTMENT OF MATHEMATICS

ST XAVIER'S COLLEGE FOR WOMEN (AUTONOMOUS),ALUVA

MARCH 2025

St. Xavier’s College for Women (AUTONOMOUS), Aluva

Re-accredited by NAAC with A++ CGPA 3.68, ISO 9001:2015

Affiliated to Mahatma Gandhi University, Kottayam

CERTIFICATE

This is to certify that the project entitled ***Different Types of Distances*** is anauthentic record of the work done by Athulya T R (220021030710), Lakshmi A J (220021030713),Shifa Thasnim M A (220021030717)under my supervision in the Department of Mathematics as a part of partial fulfilment for the award of degree of Bachelor of Science in Mathematics during 2024-2025.

Place: Aluva

Date: 25-03-2025

Project Guide	Head of theDepartment
Dr.Rakhimol Isaac	Dr. Resmi Varghese
M.Phil. B.Ed.,Ph.D.	M.Phil. NET, Ph.D.

Declaration

We hereby declare that this dissertation presented by us under the guidance of Dr.Rakhimol Isaac contains no material which has been accepted for any other Degree or diploma in any university and to the best of our knowledge. It contains no materials previously published by any other person except where due reference is made in the text.

Place: Aluva

Date: 25-03-2025

1. Athulya T R
2. Lakshmi A J
3. Shifa Thasnim M A

B.Sc. Mathematics

Department of Mathematics

St. Xavier’s College for Women

(Autonomous), Aluva

Acknowledgement

We are immensely thankful to Dr.Rakhimol Isaac, our highly esteemed and Invaluable mentor from the Department of Mathematics at St. Xavier's College for Women, Aluva. Her invaluable guidance, unwavering support, and encouragement have been pivotal to the success of our project. We also extend our deep appreciation to Dr. Resmi Varghese, our Department Head, whose assistance and encouragement played a significant role in our achievements.

We would like to express our sincere gratitude to all the faculty members in The Department of Mathematics for their constant guidance, supervision, and motivation throughout this project. Their blessings, support, and guidance have been indispensable, and we believe they will continue to illuminate our path in the journey of life.

Lastly, we want to express our deep gratitude to everyone involved, particularly our parents, friends, and all those who helped, whether directly or indirectly, in the successful execution of this project.

CONTENTS

Introduction

Distance is a fundamental concept in mathematics, used to measure how far apart two points or objects are. In various fields such as geometry, computer science, machine learning, and graph theory, different types of distances serve distinct purposes. The choice of distance metric can significantly impact problem-solving approaches, algorithm efficiency, and data interpretation.

This project explores various types of distances used in metric spaces, highlighting their mathematical definitions, properties, and applications. We begin with the Euclidean distance, the most familiar form of distance derived from the Pythagorean Theorem. The Manhattan distance, also known as the city block distance, offers an alternative measure by summing the absolute differences between coordinates. Minkowski distance generalizes both Euclidean and Manhattan distances, making it a versatile metric.

Beyond traditional metric spaces, we examine graph-based distances, such as shortest path distances, which are crucial in network analysis. The Hamming distance measures the dissimilarity between two binary strings, widely used in coding theory and error detection. The Jaccard distance, based on set similarity, is essential in text mining and clustering applications. Finally, we discuss the broader framework of metric space distances, which encompass all these measures while adhering to specific mathematical properties such as non-negativity, symmetry, and the triangle inequality.

By understanding these different types of distances, we gain deeper insight into their applications across disciplines, from spatial analysis to machine learning and beyond.

EUCLIDEAN DISTANCE

Definition:

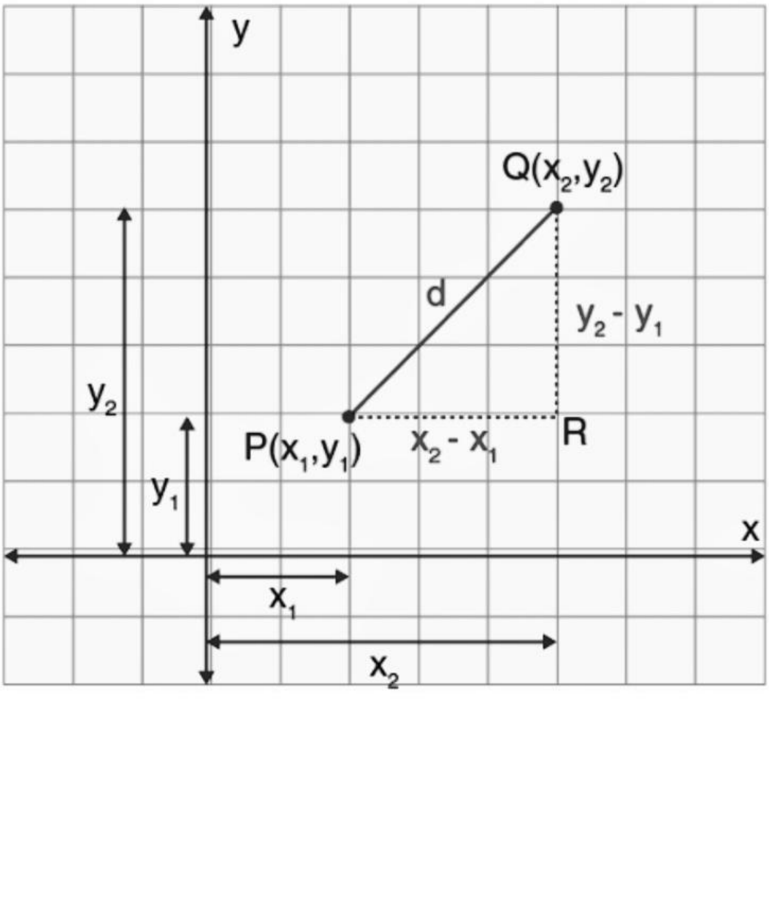
The Euclidean distance is defined as the distance between two points. In other words, the Euclidean Distance between two points in the Euclidean space is defined as the length of the line segment between two points. As the Euclidean distance can be found by using the coordinate points and The Pythagoras theorem, it is occasionally called the Pythagorean distance.

Euclidean Distance Formula:

The Euclidean distance formula helps to find the distance of a line segment. Let us assume two points, Such as $(x_{1,}y_1)$ and $(x_{2,}y_2)$in the two-dimensional coordinate plane. Thus, the Euclidean distance formula is given by:

Euclidean distance = $\sqrt{|(x_2 - x_1)^2 + (y_2 - y_1)^2|}$

Where, $(x_{1,}y_1)$ is the coordinate of the first point and $(x_{2,}y_2)$ is the coordinate of the second point.



Example:

Find the distance between points $p(3,2)$ and$Q(4,1)$

Solution: Given,

$p(3,2)$= $(x_{1,}y_1)$ and$Q(4,1)$=$(X_{2,}Y_2)$

Using Euclidean distance formula,

Distance =

PQ = $\sqrt{[(4-3)^2 + (1-2)^2]}$

PQ = $\sqrt{[(1)^2 + (-1)^2]}$

PQ = $\sqrt{2}$ units

Non-Euclidean Distance:

Non-Euclidean distance refers to distance measurements in curved or non-flat spaces, such as spheres, Cylinders, or hyperbolic spaces. There are two main types of non-Euclidean geometries:

1. Elliptical geometry (e.g., sphere): Angles and shapes are not preserved, and parallel lines can intersect.
2. Hyperbolic geometry (e.g., saddle-shaped surface): Angles and shapes are not preserved, and parallel Lines can diverge.

Advantages:

1. Simplicity and Ease of Interpretation: One of the biggest advantages of Euclidean distance is its simplicity. The calculation is straightforward and intuitive, representing the most direct path between two points. This also makes it easy to interpret and understand.
2. Applicability across Dimensions: Euclidean distance can be applied across different dimensions – be it two, three, or more. This makes it highly versatile for a wide range of multi-dimensional data analysis tasks across varied domains.

Limitations:

1. Sensitivity to Scaling: Euclidean distance can be significantly impacted by the scale or units of measurement of the data. This means that if variables are not in the same units or their scales are not normalized, the distance calculation may be skewed.
2. Inadequate for High-Dimensional Data: As the dimensionality of the data increases, Euclidean distance tends to become less effective, a phenomenon known as the "curse of dimensionality". In high-dimensions, distances between most pairs of points start to look similar, reducing the discriminatory power of Euclidean distance.

3. Ignorance of Correlation or Dependence: Euclidean distance treats each dimension separately and doesn’t account for any potential correlation or inter-dependencies that might exist among the dimensions.

Applications of Euclidean distance:

1. Google Map:

Euclidean distance is used to calculate the shortest distance between two points on a Map, providing users with the most efficient route.

Route calculation: When you enter a destination, Google Maps uses Euclidean distance to calculate the shortest distance between your current location and the destination. This helps to provide the most efficient route.

Location-Based Services: Euclidean distance is used in location-based services, such as finding nearby Restaurants, shops, or hotels. Google Maps calculates the distance between your current location and the nearest point of interest. By using Euclidean distance, Google Maps can provide users with

- More accurate distance estimates and travel times
- More efficient routes and turn-by-turn directions
- Better traffic avoidance and alternative route suggestions
- More accurate location-based services and geocoding results.

2. Computer Graphics and Animation:

Euclidean geometry forms the foundation of computer graphics and animation techniques used in video Games, movies, and virtual reality environments. Geometric transformations and spatial relationships enable the creation of realistic 3D models and simulations. For example, geometric transformations such as translation, rotation, and scaling are applied to virtual Objects to simulate movement and interaction in video games and animated films.

3. Traffic Management system:

Traffic management systems are designed to optimize traffic flow, reduce congestion, and improve Safety on roads and highways. These systems use various technologies, including sensors, cameras, and GPS, to monitor traffic conditions and make real-time decisions. It helps detect potential safety hazards, such as traffic accidents or road closures, and provides real-time information to drivers. It track the Movement of vehicles and predict their future location. This information is used to optimize traffic flow and reduce congestion.

MANHATTAN DISTANCE

Definition:

Manhattan distance is a metric used to determine the distance between two points in a grid-like path. Manhattan distance measures the sum of the absolute differences between the coordinates of the points. Mathematically, the Manhattan distance between two points in an n-dimensional space is the sum of the absolute differences of their Cartesian coordinates.

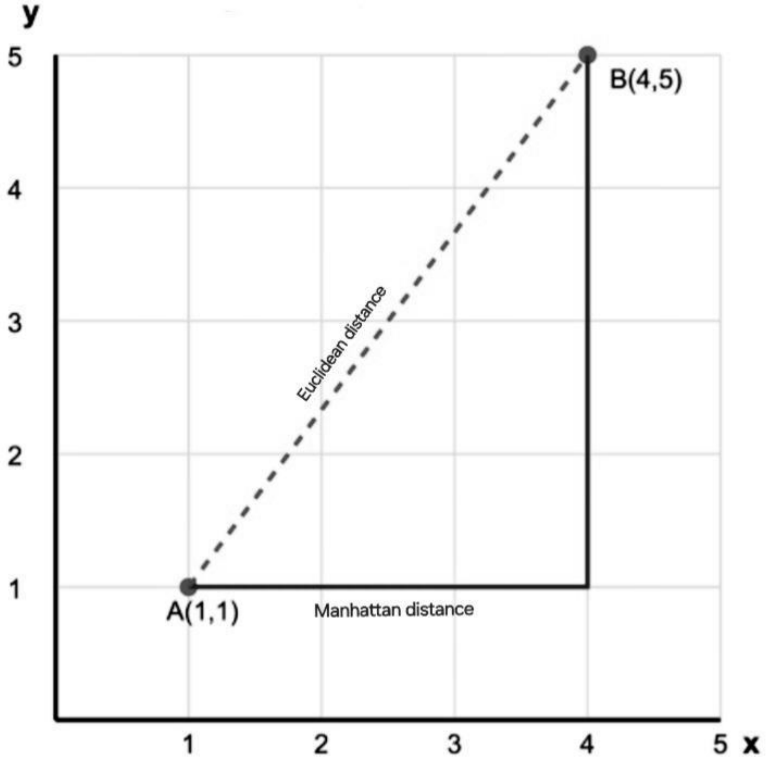
Manhattan formula:

Manhattan distance between two points $(X_{1,}Y_1)$ and $(X_{2,}Y_2)$ is

Manhattan Distance=$|X_1{-}X_2|$+$|Y_1{-}Y_2|$

Example:

Find the distance between the points $A(1,1)$ and $B(4,5)$



Solution:

Given $A(1,1)$ and $B(4,5)$

Manhattan distance= $|4-1|$+$|5-1|$

$= 3 + 4 = 7 units$

Where,

Euclidean distance $=\sqrt{[(4-1)+(5-1)]}$

$= \ 5\ units$

Properties of Manhattan distance:

1. Manhattan distance is a true metric, which means it satisfies all four conditions required for a distance function in a metric space such that

•Non-negativity: The distance between any two points is always non-negative. $D(X,Y) \geq 0$ for all x and y.

•Identity of indiscernible: The distance between a point and itself is zero, and if the distance between two points is zero, they are the same point. $D(X,Y) = 0$ if and only if $X = Y$

•Symmetry: The distance from point A to point B is the same as the distance from B to A. $d(x,y) = d(y,x)$ for all x and y.

•Triangle inequality: The distance between two points is always less than or equal to the sum of the distances between those points and a third point. $D(X,Z) \leq d(x,y) + d(y,z)$for all x, y, and z.

2. Translation invariance in Manhattan distance:

Translation invariance means that if you shift all points in a dataset by the same amount in any direction, the relative distances between those points remain unchanged. Manhattan distance is translation invariant, since the formula only involves absolute differences between coordinates, shifting all points by the same amount will simply add or subtract the same value from both coordinates in the absolute difference calculation, resulting in the same distance value.

Example:

Consider points $A(1,2)$ and $B(4,5)$

Manhattan distance between A and B: $|1-4|$+$|2-5|$=

If we shift both points by $(2,1)$ to $A(3,3)$ and B$(4,5)$ to $B(6,6)$:

Manhattan distance between A’ and B’:$|3-6|$+$|3-6| = 3 + 3 = 6$

Differences between Manhattan and Euclidean distance:

Manhattan Distance (L1 or City-Block Distance): We can only move horizontally and vertically, like a taxi driver. Euclidean Distance (L2 or Straight-Line Distance):A straight line connecting two points, like a bird flying directly.

1. Path:

Manhattan distance follows a grid-like path, while Euclidean distance measures the shortest straight-line path.

2. High Dimensionality:

Manhattan distance can be more robust in high-dimensional spaces, as it avoids the "curse of dimensionality" where Euclidean distances tend to become less informative.

3. Scalability:

Euclidean distance can be computationally expensive in high dimensions, while Manhattan distance is generally faster.

Advantages:

1. Computational Efficiency: Manhattan distance is computationally faster to calculate than Euclidean distance, especially in environments where movement is restricted to a grid-like path.
2. Suitable for High-Dimensional Data: It can be particularly effective in high-dimensional nearest neighbour search, as it is less affected by the "curse of dimensionality".
3. Applicable to Categorical/Binary Data:Manhattan distance is often preferred in distance-based clustering algorithms involving categorical or binary data.
4. Intuitive for Grid-Like Environments:It's a natural and intuitive measure for situations where movement is restricted to a grid-like path, such as city navigation.

Limitations:

1. Less Intuitive in High Dimensions:Manhattan distance can be less intuitive than Euclidean distance in representing distances in high-dimensional spaces.
2. Not the Shortest Path:It doesn't necessarily represent the shortest possible path between two points, as it only considers horizontal and vertical movements.
3. Directional Bias: Manhattan distance can be biased towards certain directions, as it only considers movements along the axes.
4. Not as flexible as Euclidean distance: Euclidean distance can be used in any space to calculate distance, whereas Manhattan distance is more limited to grid-like spaces.

Applications of Manhattan distance:

Manhattan distance finds applications in various fields of computer science, data analysis, and geospatial technology

1. Path finding algorithms (e.g., A* algorithm) :

In grid-based environments, Manhattan distance provides a quick and effective heuristic for estimating the distance between two points. It's particularly useful in the A* algorithm, where it can help guide the search towards the goal more efficiently in scenarios where movement is restricted to horizontal and vertical directions.

2. Geographic Information Systems (GIS) :

In GIS applications, Manhattan distance can model movement along a grid-like street network, making it useful for urban planning and logistics. It's used in location-allocation problems, such as determining optimal locations for facilities based on minimizing total travel distance in a city. Manhattan distance can also be applied in spatial analysis tasks, such as buffer zone creation around linear features like roads or rivers. Urban planners might use Manhattan distance to analyse the accessibility of public services, while logistics companies could employ it to optimize delivery routes in cities.

3. Image recognition:

Manhattan distance can be used to compare pixel values or feature vectors. It's particularly useful in template matching, where you're trying to find occurrences of a small image within a larger one. It is also valuable in facial recognition systems, object detection in video streams, or pattern matching in large image databases, where speed is crucial, and the slight loss in precision compared to Euclidean distance is often negligible.

MINKOWSKI DISTANCE

Definition:

Minkowski distance is a way to measure the distance between two points in a multi-dimensional space. It’s a generalization of other distance measures, including the Euclidean and Manhattan distances.

The Minkowski distance is calculated by adding the absolute differences between the coordinates of two points, raised to a power. The power, or parameter, “p”, determines the type of distance being measured. Different values of “p” result in different distance measures.

MinkowskiFormula:

$d(x, y) = (\sum_{i=1}^{n} |x_i - y_i|)^{1/p}$

Where,

x and y are two points in an n-dimensional space.

p is a parameter that determines the type of distance (p ≥ 1).

$|x_i - y_i|$ represents the absolute difference between the coordinates of x and y in each dimension.

Special Cases of Minkowski Distance:

1. When $:p = 1$

$$D_{Mnnhattan}(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

When p is set to 1, the Minkowski distance becomes Manhattan distance also known as city block distance or L1 norm, Manhattan Distance measures the sum of absolute differences.

2. When:$p = 2$:

$$DManhat\tan(x, y) = \left(\sum_{i=1}^{n} (x_i - y_i)^2\right)^{\frac{1}{2}}$$

When p is set to 2, Minkowski distance becomes Euclidean distance. Euclidean distance is the most common distance metric, representing the straight-line distance between two points.

3. When $p \to \infty$

$$D_{Chebyshev}(x, y) = max_{i|x_i - y_i|}$$

When p tends to infinity, Minkowski distance becomes Chebyshev distance, also known as chessboard distance, measures the maximum difference along any dimension.

Example:

Calculate the Minkowski distance between the points $A(2,3)\ and\ B(5,7)$ for different p values.

Solution:

When $p = 1$, (Manhattan distance)

$$D = |5 - 2| + |7 - 3| = 7units$$

When $p = 2$, (Euclidean distance)

$$D = [(5 - 2)^2 + (7 - 3)^2]^{\frac{1}{2}} = 5units$$

When p=3,

$$D = [(5 - 2)^3 + (7 - 3)^3]^{\frac{1}{3}} = 4.481\, units$$

Properties of Minkowski distance:

Minkowski distance satisfies the four essential properties required for a function to be considered a metric in a metric space:

•Non-negativity: The Minkowski distance between any two points is always non-negative, d$d(x, y) \geq 0$. This is evident as it is the p-th root of a sum of non-negative terms (absolute values raised to the power p).

• Identity of Indiscernible: The Minkowski distance between two points is zero if and only if the two points are identical. Mathematically $d(x, y) = 0$ and only if $x = y$ This follows because the absolute difference between identical components is zero.

• Symmetry: The Minkowski distance is symmetric, meaning $d(x, y) = d(y, x)$ This property holds because the order of subtraction in the absolute value terms does not affect the outcome.

• Triangle Inequality: The Minkowski distance satisfies the triangle inequality, which states that for any three points x, y, and z, the distance from x to z is at most the sum of the distance from x to y and from y to z; formally, $d(x, y) \leq d(x, y) + d(y, z)$

Advantages:

1. Flexibility: Minkowski distance allows for a wide range of distance measures by adjusting the parameter ‘p’. When p=1, it becomes Manhattan distance, and when p=2, it becomes Euclidean distance.
2. Adaptability: It can be used for different use cases by choosing the most suitable distance measure based on the nature of the data and the problem at hand.
3. Generalization: It generalizes both Euclidean and Manhattan distances, providing a unified framework for calculating distances in different scenarios.

Limitations:

1. Computational Cost: Finding the optimal value of ‘p’ can be computationally inefficient, especially for large datasets.
2. Scale Dependency: Like Euclidean distance, Minkowski distance can be sensitive to the scale of the features, meaning that distances computed might be skewed depending on the units of the features.
3. Noise Sensitivity: Minkowski distance can be sensitive to outliers and noise in the data, which can negatively impact the results.
4. Inherited Disadvantages: It inherits the same disadvantages as the distance measures it represents (Manhattan, Euclidean, etc.), such as problems in high-dimensional spaces.

Applications of Minkowski distance:

1. Image Similarity Measurement:

Minkowski distance can be used to measure the similarity between images. For example, in image retrieval systems, Minkowski distance can be used to rank images based on their similarity to a query image.

2. Video Object Tracking:

Minkowski distance can be used to track objects in videos. By calculating the Minkowski distance between the object’s positions in consecutive frames, the object’s trajectory can be estimated.

3. Anomaly Detection:

Minkowski distance can be used to detect anomalies in data. By calculating the Minkowski distance between a data point and the centroid of a cluster, anomalies can be identified.

4. Text Similarity Measurement:

Minkowski distance can be used to measure the similarity between text documents. This can help in information retrieval systems, such as search engines.

5. Text Classification:

Minkowski distance can be used as a distance metric in text classification algorithms.

6. Time Series Analysis:

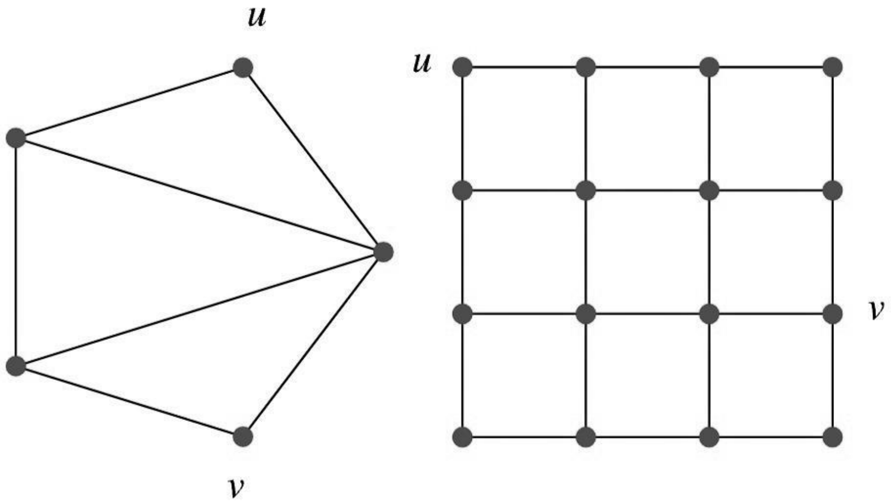
Minkowski distance can be used to analyse time series data and identify patterns or anomalies.

GRAPH DISTANCE

Definition:

The distance $d(u, v)$ between two vertices u and v of a finite graph is the minimum length of the paths connecting them (i.e., The minimum number of edges between two vertices u & v)

Example:



Graph: AGraph:B

Consider graph A and B, the distance between u and v of the;

i) $Graph\ A = d_A(u,v) = 2$

ii) $Graph\ b = d_B(u,v) = 5$

Properties of Distance in Graph Theory:

- Non-negativity: The distance between two vertices is always non-negative.
- Identity of Indiscernible: The distance between a vertex and itself is 0.
- Symmetry: The distance between two vertices is symmetric, i.e., $d(u,v) = d(v,u)$
- Triangle Inequality: The distance between two vertices satisfies the triangle inequality, i.e., $d(u,v) \leq d(u,w) + d(w,v)$

Different types of Distance in Graph Theory:

1. Eccentricity of graph:

It is defined as the maximum distance of one vertex from other vertex. The maximum Distance between a vertex to all other vertices is considered as the eccentricity of the vertex. It is denoted by $e(V)$.

2. Diameter of graph:

The diameter of graph is the maximum distance between the pair of vertices. It can also be defined as the maximal distance between the pair of vertices. Way to solve it is to find all the paths and then find the maximum of all. It can also be found by finding the maximum value of eccentricity from all the vertices.

3. Radius of graph:

A radius of the graph exists only if it has the diameter. The minimum among all the maximum distances between a vertex to all other vertices is considered as the radius of the Graph G. It is denoted as $r(G)$. It can also be found by finding the minimum value of eccentricity from all the vertices.

4. Detour Distance:

If u, v are two vertices in the graph G, the detour distance between these vertices denoted by $D(u, v)$is the length of a longest u – v- path in G.

Advantages:

1. Modelling Complex Systems: Graph theory excels at representing and analysing complex systems and relationships, making it useful for fields like social networks, transportation, and logistics. They can convey information quickly and clearly, making it accessible to a wide audience.
2. Data Visualization: Graphs provide a clear and intuitive way to visualize data, making complex relationships easier to understand.
3. Problem Solving: Graph theory algorithms are widely used to solve various problems, including finding shortest paths, identifying network bottlenecks, and optimizing routes.
4. Versatility: Graphs can represent a wide range of data types and structures, including social networks, road networks, and the internet.
5. Scalability: Graph data structures can handle large amounts of data and scale well with the size of the input, making them suitable for big data applications.
6. Real-World Applications: Graph theory has practical applications in diverse fields like social network analysis, logistics, transportation, computer networks, and even in understanding biological systems.

Limitations:

1. Complexity: Graph theory can be complex, especially when dealing with large or intricate networks, making it challenging to understand and implement.
2. Beginner Difficulty: The concepts and terminology of graph theory can be challenging for beginners, requiring a certain level of mathematical understanding.
3. Abstraction: Graph theory relies on abstraction, which can sometimes obscure the underlying details of the real-world problem being modelled.

4. Computational Cost:Analysing and manipulating large graphs can be computationally expensive, requiring efficient algorithms and resources.
5. Data Representation: While graphs are versatile, they might not always be the most suitable representation for all types of data or problems.
6. Potential for Misinterpretation:Graphs can be easily misinterpreted if not properly labelled or understood, leading to incorrect conclusions.

Applications of Graph Theory distance:

1. Navigation and route planning:

GPS systems and mapping applications leverage graph theory to calculate the shortest driving route between two locations by representing roads as edges and intersections as vertices, using algorithms like Dijkstra's algorithm to find the minimum distance path.

2. Transportation network optimization:

Designing efficient transportation networks for goods delivery, considering factors like distance, traffic congestion, and capacity constraints, can be done by modelling the network as a graph and finding the shortest paths between nodes.

3. Social network analysis:

Studying connections between individuals in social networks can be done by representing people as vertices and relationships as edges, allowing analysis of "distance" between individuals based on their social ties.

4. Biological network analysis:

In protein-protein interaction networks, the distance between two proteins represents the number of intermediary interactions, helping

HAMMINGDISTANCE

Definition:

Hamming distance is a measure of the number of positions at which two strings of equal length are different.

Hamming Distance Formula:

The Hamming distance between two strings of equal length is calculated using the following formula:

$$\boldsymbol{H(X,Y) = \sum |x_i \neq y_i|}$$

Where:

$H(x, y)$ is the Hamming distance between strings x and y.

x_i is the i-th character of string x.

$y_i$is the i-th character of string y.

position	String 1	String 2	Different?

∑ denotes the sum of the number of positions at which the characters are different.

Alternatively, the Hamming distance can be calculated using the following formula:

$$\boldsymbol{H(x,y) = |x \oplus y|}$$

Where:

⊕ denotes the bitwise XOR operation

|$|x|$ denotes the number of 1-bits in the binary representation of x

This formula is commonly used for binary strings, but can be generalized to other types of strings as well.

Calculating Hamming Distance:

To calculate the Hamming distance between two strings,

Follow these steps:

1. Ensure the strings are of equal length.
2. Compare the strings character by character.
3. Count the number of positions at which the characters are different.

Examples:

Calculate the Hamming distance between the following two binary strings

$$Strings\ 1{:}\,101010\ \&\ String\ 2{:}\,110011$$

Solution:

To calculate the Hamming distance, we compare the strings character by character:

position	String 1	String 2	Different?
1	**1**	**1**	**No**
2	**0**	**1**	**Yes**
3	**1**	**0**	**Yes**
4	**0**	**0**	**No**
5	**1**	**1**	**No**
6	**0**	**1**	**Yes**

The Hamming distance is 3, because the strings differ in three positions

(Positions 2, 3 and 6).

Advantages:

1. Simple and Easy to Compute: Hamming distance is straightforward to calculate, involving a simple comparison of corresponding bits.
2. Effective for Single-Bit Error Detection: It's particularly useful in error-detecting codes, where it can identify single-bit errors.
3. Applicable to Binary Data: Hamming distance is well-suited for analysing binary data, such as DNA sequences or error-correction codes.

Limitations:

1. Limited to Equal-Length Strings: Hamming distance can only compare strings of the same length.
2. Doesn't Account for Magnitude of Differences: It only indicates the number of differing bits, not the significance of those differences.
3. Not Suitable for All Error Scenarios: While effective for single-bit errors, it may not detect or correct multiple-bit errors.
4. Primarily for Binary Data: Hamming distance is primarily designed for binary data and may not be as suitable for other types of data.
5. Not Flexible:It is less flexible than other similarity metrics like Levenshtein distance which can handle strings of different lengths and different types of edits.

Applicationsof Hamming distance:

Hamming distance has numerous applications in various real-world fields, including

1. Data Transmission:

Hamming distance is used to detect and correct errors in digital data transmission, ensuring reliable communication.

2. DNA Sequencing:

Hamming distance is used to compare DNA sequences and identify similarities and differences.

3. Image and Video Compression:

It’s used in image and Video compression algorithms, such as JPEG and MPEG.

4. Language Modelling:

It’s used in language modellingtasks, such as language translation and text generation.

5. Iris Recognition:

Hamming distance can be used to compare iris patterns for authentication purposes.

6. Data Privacy:

Hamming distance can be used in privacy-preserving protocols.

JACCARD DISTANCE

Definition:

Jaccard distance is a measure of dissimilarity between two sets.

It quantifies how different two sets are based on the overlap (or lack thereof) of their elements.

Jaccard distance formula:

$$\boldsymbol{Jaccard\ Distance = 1 - (Jaccard\ Similarity)}$$

Where,

$$\boldsymbol{Jaccard\ Similarity} = |\boldsymbol{A}\cap\boldsymbol{B}|/|\boldsymbol{A}\cup\boldsymbol{B}|$$

Where,

$|A \cap B|$ is the number of elements common to both sets A and B (intersection).

$|A \cup B|$ is the total number of unique elements in both sets A and B (union).

Jaccard similarity:

Definition:The Jaccard similarity coefficient (or Jaccard index) quantifies the similarity between two sets by calculating the ratio of the size of their intersection to the size of their union.

$$Formula: J(A, B) = |A \cap B|/|A \cup B|$$

Example:

Sets: $A = \{1,2,3\}, B = \{2,3,4\}$

Intersection: $\{2,3\}$ (elements common to both sets)

Union: $\{1,2,3,4\}$ (all unique elements)

Jaccard Similarity= |Intersection| / |Union|

$$= \frac{2}{4} = 0.5$$

Jaccard Distance = 1 – Jaccard Similarity

= $1 - 0.5 = 0.5$

Remark:

Jaccard similarity measures the overlap between two sets, while Jaccard distance measures the dissimilarity, calculated as 1 minus the Jaccard similarity.

Advantages:

1. Simplicity and Intuition:The Jaccard distance is easy to understand and calculate, making it accessible for both data scientists and non-technical users.

2. Effective for Binary Data:It’s particularly well-suited for comparing sets where elements are either present or absent (binary or categorical data), as it focuses on the overlap of elements.
3. Versatility:Can be applied to various data types and domains, including text, biological data, and social network analysis.
4. Scale-Invariance:The Jaccard distance is not affected by the size of the sets being compared, making it robust for comparing sets of varying sizes.

Limitations:

1. Ignores Term Frequency:It doesn’t consider how many times an element appears in a set, which can be a limitation when dealing with data where frequency is important.
2. Less Effective for High-Dimensional Data:In high-dimensional spaces, the Jaccard distance might not be as informative as other similarity measures.
3. Doesn’t Capture Magnitude Information:The Jaccard distance focuses solely on the presence or absence of elements, and doesn’t consider the magnitude or values associated with those elements.
4. Sensitivity to Set Size:While Jaccard Similarity is scale-invariant, the Jaccard Distance can be sensitive to set size, as the denominator is the union of the two sets.

Applications of Jaccard distance:

1. Natural Language Processing (NLP):

Jaccard similarity is used to compare the similarity of two text documents in tasks like document clustering, plagiarism detection, and recommendation systems.

2. E-commerce:

It helps identify similar customers based on their purchase history, allowing for targeted recommendations and personalized experiences.

3. Data Deduplication:

Jaccard similarity can be used to detect and remove duplicate records by comparing the sets of attributes within records.

4. Social Network Analysis:

It helps analyse social networks by comparing the sets of friends or connections between individuals, identifying users with similar social circles.

5. Web Data Mining:

Jaccard distance can be used to estimate language discrepancies between travellers and destination marketers.

METRIC SPACE DISTANCE

Definition:

A metric space is a $d(X, d)$,

Where:

X is a set.

$d: X \times X \rightarrow R$ is a function (called a metric or distance function) that satisfies the following properties for all $x, y, z \in X$::

1. Non-negativity :$d(x, y) \geq 0$ (Distances are always non-negative.)
2. Identity of Indiscernibles :$d(x, y) = 0 \Leftrightarrow x = y$(The distance is zero if and only if the two points are the same.)
3. Symmetry: $d(x, y) = d(y, x)$(The distance from x to y is the same as from y to x.)
4. Triangle Inequality: $d(x, z) \leq d(x, y) + (y, z)$

Example:

If the function d defined on the set of Real Numbers as $d(x,y) = |x - y|$,for all $x, y \in R$ then (R, D) is a metric space .[USUAL METRIC]

1. Non-negativity :$d(x, y) = |x - y| \geq 0$

(Since absolute values are always non-negative.)

2. Identity of Indiscernibles: $d(x, y) = 0 \Leftrightarrow |x - y| = 0 \Leftrightarrow x = y$

(Distance is zero if and only if the two points are the same.)

3. Symmetry $(x, y) = |x - y| = |y - x| = d(y, x)$

(The distance remains the same when swapping x and y.)

4. Triangle Inequality:. $d(x, z) = |x - z| \leq |x - y| + |y - z| = d(x, y) + d(y, z)$

(This follows from the triangle inequality property of absolute values.)

Since all four properties are satisfied, $d(x, y) = |x - y|$ is a valid metric on R.

(R, d) a Metric space.

Advantages:

1. It provides a well-defined and structured way to measure distances.Ensures logical consistency due to metric properties (non-negativity, symmetry, triangle inequality, etc.).
2. Wide Applicability: Used in diverse fields like machine learning, physics, optimization, bioinformatics, and network analysis.
3. Flexibility in Choosing Metrics: Different distance metrics (Euclidean, Manhattan, Jaccard, etc.) can be selected based on the nature of the data.
4. Computational Efficiency: Simple metrics like Euclidean distance are easy to compute for small datasets.

5. Useful for Similarity and Classification: Distance-based approaches help in pattern recognition, recommendation systems, and anomaly detection.

Limitations:

1. Sensitivity to High Dimensions: Many metrics (especially Euclidean distance) become less meaningful as the number of dimensions increases.
2. Computational Cost for Large Datasets: Some distance calculations (e.g., pairwise distances) can be slow when dealing with big data so advanced optimization techniques are required for efficiency.
3. Not Always Intuitive for Certain Data Types: Traditional metrics may not work well for non-numeric data, requiring specialized distance measures (e.g., Jaccard for categorical data).
4. Not All Similarity Measures Are Metrics: Some useful similarity measures do not satisfy all metric properties (e.g., triangle inequality).

.Applications of Metric Space distance:

1. GPS Navigation:

Route planning: GPS navigation systems use metric space distance to plan routes and provide turn-by-turn directions.

Distance measurement: GPS systems measure distances between locations to provide estimated arrival times.

2. Banking & Fraud Detection:

Credit Card Fraud Detection: Banks use metric distances to identify unusual spending patterns that differ from a customer’s normal behaviour.

Loan Approval: Banks assess similarity between new applicants and past approved customers based on financial data.

3. . Food Delivery:

Distance measurement: Food delivery services like Grub Hub and UberEatsuse metric space distance to measure distances between restaurants and customers.

Route optimization: Food delivery services use metric space distance to optimize routes and reduce delivery times.

CONCLUSION

Distance metrics are fundamental in mathematics and have diverse applications across various fields, from geometry and computer science to data analysis and artificial intelligence. Through this project, we explored several key distance measures, including Euclidean, Manhattan, Minkowski, Graph-based, Hamming, Jaccard, and metric space distances. Each of these distances has unique properties and is suited for different types of data and problem domains.

Understanding these distances allows for more informed decision-making when applying mathematical models to real-world problems. As data science and artificial intelligence continue to evolve, selecting the appropriate distance metric will remain a critical factor in achieving accurate and efficient results.

REFERENCES

1. Herbert Busemann,Paul J. Kelly, *Projective Geometry and Projective Metrics*, Academic Press (1953)
2. Paul E. Black, *Dictionary of Algorithms and Data Structures,* National Institute of Standards and Technology (1998)
3. Dr.Dankan,VGowda,K.S. Shashidhara, Ramesha M, Sridhara S B, *Advances in Mathematics Scientific Journal 10(3):1407-1412* (2021)
4. Leigh Metcalf, William Casey, *Cyber security and Applied Mathematics*, Syngress (2016)
5. Martire,P. N. da Silva, A. Plastino, F. Fabris, A. A. Freitas,*A novel probabilistic Jaccard distance measure for classification of sparse and uncertain data*(2017)