Finding Similar Items

Ravi Kumar Gupta

https://kravigupta.in

Finding Similar Items

Similarity

- Measure of how alike two samples/objects are
- Range: 0 to 1
  - 0 => Completely dissimilar
  - 1 => Completely similar

Correlation

- Measures the linear relationship between two variables
- Instead of comparing the data points directly, correlation compares how two variables change relative to each other
- Range: -1 to 1
  - -1 => Perfectly negative, if x changes by 1, y will change by -1
  - 0 => no linear relationship, if x changes , no change in y
  - 1 => Perfectly positive relation, if x changes by 1, y will change by 1

Finding Similar Items

Similarity & Correlation

- Basic building blocks for activities such as –
  - Clustering
  - Classification
  - Anomaly detection
- Key applications
  - Advertiser keyword suggestions
  - Collaborative filtering
  - Web search

Advertiser keyword suggestions

- Expand the manually input keyword set for better ad targeting.
- Use similarity measures to identify keywords with like meanings.
- **Ex –**
  - For **running shoes,** suggestions like "jogging footwear" or "athletic shoes" might be generated
  - **Organic Coffee -** "Natural coffee beans," "Eco-friendly coffee," "Pesticide-free coffee," "Fair trade coffee beans.”
  - **Yoga Mat -** "Exercise mats," "Eco yoga pads," "Non-slip yoga surfaces," "Pilates mats."

Collaborative Filtering

- Identify users with similar interests to make tailored recommendations.
- Compare user profiles or behavior to find those with interests that align beyond a set threshold.
- Example - Movies –
  - If User A and User B both enjoyed movies "Inception" and "Interstellar," they might have similar tastes.
  - Thus, if User A liked "Blade Runner 2049," it could be recommended to User B
- Books –
  - User A - Purchased books "Harry Potter," "Percy Jackson," "Hunger Games," "Divergent.”
  - User B – bought – Percy Jackson, recommendation - "Hunger Games," "Maze Runner” etc.

Web Search

- Enhance user query results using similarity measures.
- Expand the user's initial query by adding clusters of similar queries for more comprehensive results.
- Example – For a search query – apple pie recipe
  - The search engine might also consider results for "homemade apple pie" or "best apple pie ingredients"
  - to provide richer, more relevant results.
- Another example - Tourist attractions in Paris
  - Expanded queries - "Must-visit places in Paris," "Historical sites in Paris," "Best views in Paris," "Popular parks in Paris."

More applications ..

Similarity

- **Document Retrieval** - Finding documents similar to a given document in large databases.
- **Image Recognition** - Comparing features of an input image with features of labelled images in a database.
- **Voice recognition** - Comparing voice command input with known voice patterns.
- **Plagiarism Detection** - Comparing documents to identify potential copying.
- **Product Recommendation** - Recommending products similar to a user's past purchases or preferences.

More applications ..

Correlation

- **Financial Markets**: Analyzing how different stock prices or commodities move in relation to each other.
- **Healthcare:** Investigating relationships between different health factors.
- **Environmental Studies**: Determining relationships between different environmental factors.
- **Marketing :** Analyzing the relationship between advertising spend and sales volume.

Nearest Neighbor (NN) Search

Nearest Neighbor Search

- Nearest Neighbour Search (NN Search)
- Also known as – Proximity search, Similarity Search, Closest Point Search
- NN Search – is an optimisation problem for finding closest or most similar points
- Formally – NN Search problem is defined as -
    - Given a set S of points in Space M
    - A query point $q \in \boldsymbol{M}$
    - Find the set of closest points in S to q

Nearest Neighbor Search

Domain of application

- **Multimedia**: Find similar images in vast databases.
- **Biology**: Match DNA sequences.
- **Finance**: Compare stock trends with historical data.
- **Social Networks**: Identify users with similar behaviors.
- **Sensors**: Spot events that match a particular sensor reading.

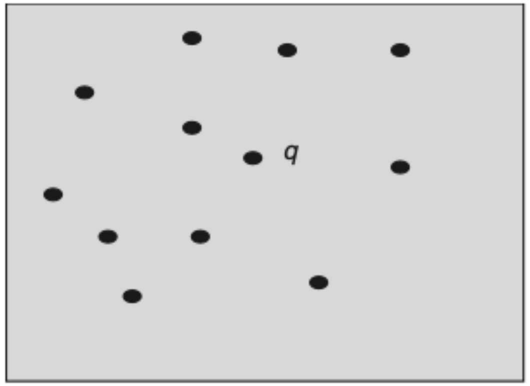
Importance in machine learning

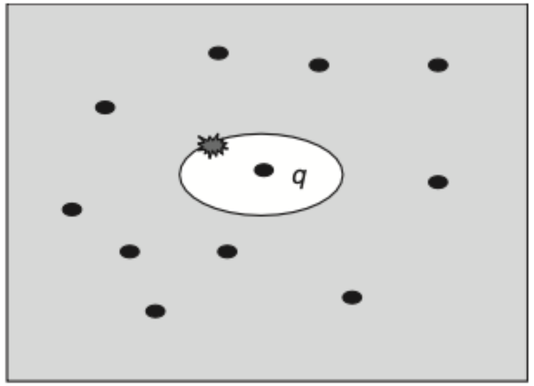
- Techniques like k-NN and pattern-based classification are essentially forms of NN search.
- Used in recommendation systems to find similar users/items.

Relevance in Big Data

- As data grow for Data mining problems e.g., classification, clustering, pattern identification etc. , approximate NN search becomes vital to efficiently compute similarities.

NN Search - Formulation





Dataset X with n points

Query Point q

Distance Metric D(s,t)

Nearest Neighbor Xnn to q in X

Formally,

Suppose there is a dataset X with n points $X = \{X_i, i = 1, \ldots, n\}$.

Given a query point q and a distance metric $D(s, t)$,

find q's nearest neighbor X_{nn} in X, that is

$\mathbf{D(X_{nn}, q) \leq D(X_i, q)}$, $i = 1, \ldots, n$

Jaccard Similarity

- Alternate formulation of NN – in the realm of Set Theory
- Answers to the query –
  - Given a set, find similar sets from a large dataset.
  - Or Given a large dataset, find all similar sets of items
- Basically, amounts to finding the size of intersection of two sets to evaluate similarity.

- A similarity measure s(A,B) – indicates the closeness between A and B.
- Properties of Good measure –
  - It has a large value if the objects *A* and *B* are close to each other.
  - It has a small value if they are different from each other.
  - It is (usually) 1 if they are same sets.
  - It is in the range [0, 1].

Jaccard Similarity

Example

Given two sets

- $A = \{0, 1, 2, 4, 6\}$
- $B = \{0, 2, 3, 5, 7, 9\}$

Jaccard Similarity => $JS(A,B) = | A \cap B | \, / \, |A \cup B|$

$|X|$ => cardinality of a set X => number of items in a set; $|A|$ => 5, $|B|$ => 6

Find Jaccard Similarity of the sets above

- $JS(A,B) = |\{0,2\}| \, / \, |\{0,1,2,3,5,6,7,9\}| = 2/8 = 0.25$

Applications of NN Search

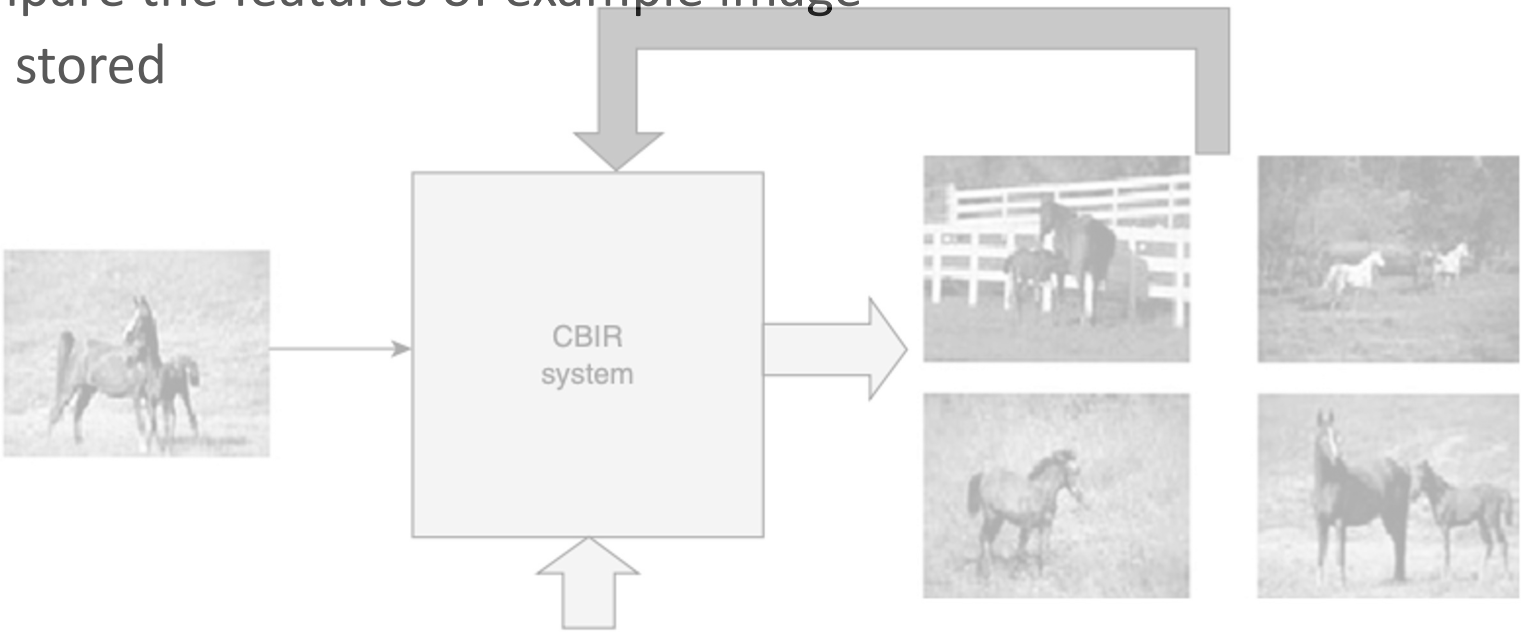
Applications of NN Search

- Optical Character Recognition (OCR)
  - Converts different types of documents into editable and searchable data
  - Documents like scanned paper documents, PDF files, handwritten document, image captured by digital camera etc.
  - OCR software use NN classifier e.g. k-NN algorithms; It compares image features of characters with stored glyph features.
- Content-based image retrieval
- Collaborative filtering
- Document Similarity

Applications of NN Search

Content-based image retrieval

- Process in which images are retrieved from databases based on the content or features present in the images rather than metadata such as keywords, tags, description etc
- Use of NN approach
  - NN approach is used to compare the features of example image
  - with the features of images stored
- This method is used in
  - Medical imaging
  - Digital libraries etc.



Applications of NN Search

- Collaborative filtering
- Document Similarity

We will learn these in detail

Similarity of Documents

Similarity of Documents

- Automated organization and analysis of vast document repos
- Crucial and challenging for modern applications such as –
    - Enterprise storage, Hospital record system, web search engines, trending topics on twitter
- The immense volume, variety and velocity of documents creation makes assessment of document similarity a significant big data problem.

What are we looking for –

- A reliable and efficient similarity measure
- Which will help to answer the questions like
    - How similar are the two text documents?
    - Are two patient histories similar? Etc.

Similarity of Documents

- The similarity we are looking for is – character level and not the semantic meaning
- This requires us to examine the words in the documents and their uses
- Simple hash algorithms can detect identical duplicate documents
- However, finding near-duplicates, similar are more complex

- Useful applications for the text-based similarities in big data scenarios –

1. Near-duplicate detection in search engines
2. HR applications such as automated CV to job description matching, finding similar employees
3. Patent research – Matching new application against a vast database to ensure originality.
4. Document clustering and auto categorization using seed documents
5. Security scrubbing – finding documents with similar content but with different access control lists

Plagiarism Detection

- Process of locating instances of plagiarism within a work or document
- Most cases are found in academia – where documents are typically essays or reports
- Possible in virtually any field, including scientific papers, art designs and source code as well.
- Uses textual similarity measures

Plagiarism detection tools

- Turnitin
- iThenticate

Turnitin

- Web-based system for plagiarism and citation checks.
- Compares document content against a massive database.
- Produces a similarity report highlighting suspicious content.
- Database includes:
  - Academic databases and journals.
  - 200+ million student assignments.
  - 17+ billion web pages.

iThenticate

- Popular professional tool for plagiarism prevention.
- Targets faculty research, articles, grant proposals, course materials, and academic theses.
- Database includes:
  - Content from 90,000+ newspapers, magazines, scholarly journals, and books.
  - 14+ billion web pages (current and archived).
  - Materials from 50,000+ scholarly journals.
  - Content from 150+ STM (Scientific, Technical, and Medical) publishers.
- Outputs a similarity report shortly after document submission.
  - Displays an overall similarity index (percentage of matched content).
  - Lists all sources contributing to the similarity index.

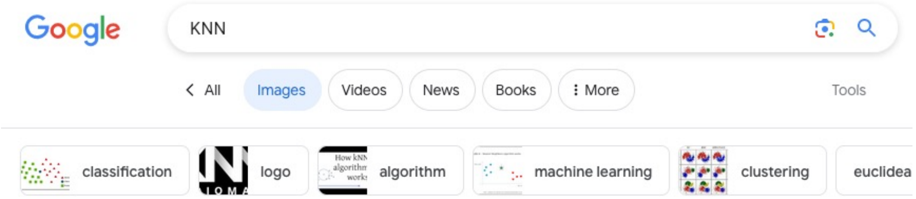
iThenticate

- Other reports include the following:

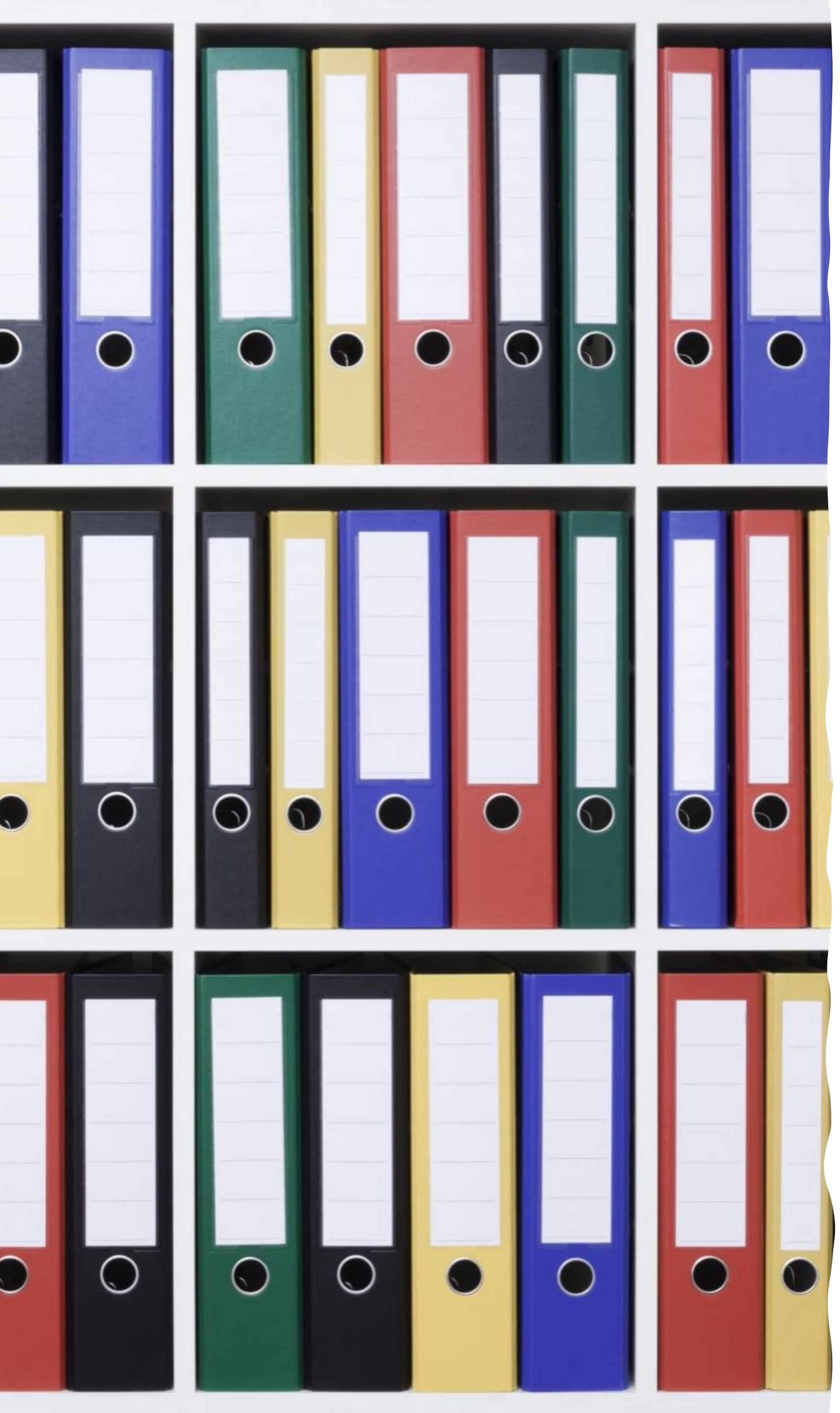
1. Side-by-side comparison of the document to matched sources in the database.
2. Largest matches showing where sources and text of the largest content match in the document.
3. Summary report that provides a high-level overview of matched content in the document.

Document Clustering

- Common in problems like clustering and cross-document co-reference resolution.
- **Cross-document Co-reference Resolution** refers to the process of determining when entities (e.g., names, pronouns, nouns) in different documents refer to the same underlying real-world entities
- Example - For example, in a single document, "Barack Obama" might be later referred to as "he", "the President", or "Obama".
- **Web Search Engines:**
  - Broad queries yield thousands of results.
  - Engines like Google offer a "Similar" link for each primary result.
  - This link leads to documents similar to the primary result.
  - Similarity is evaluated based on the user's query.
  - Primary link directs to the top-ranked page.



- **Document Clustering:**
  - Automatically groups retrieved documents into meaningful categories.
  - Yahoo! auto-generates a taxonomy of web documents.



Document Clustering

Content Management at Hewlett-Packard:

- HP has millions of technical support documents across collections.
- Periodic merging and grooming of these collections.
- Essential to identify and remove near-duplicate or outdated documents.
- **Aims**: Improve collection quality, refine search results, boost customer satisfaction.
- **Challenge**: Identify similar documents based on content, not potentially unreliable metadata.

News Aggregators

- Aggregate content from multiple sources like RSS feeds and online news agencies.
- Cluster similar articles covering the same events or stories.
- **Example**:
  - Multiple outlets reporting on "Germany winning the World Cup in 2014".
  - Goal is to recognize if these articles are discussing the same event, despite variations in reporting.
- **Unique Challenge**:
  - Each source may have an original take or perspective on the story.
- **Role of Aggregators like Google News**:
  - Identify when two articles are textually similar, but not exact copies.
  - Present them as different versions of the same core news story.

Collaborative Filtering as a Similar-Sets Problem

Collaborative Filtering

- **Digital Age Challenges**:
  - The Internet era provides us with numerous choices in our daily lives, from movies to shopping and more..
  - Decision domains are vast. E.g., Netflix offers over 17,000 movies; Amazon's Kindle store boasts over 410,000 titles. – illustrates the decision overload
  - Helping users navigate these vast domains is challenging.
- **Collaborative Filtering**:
  - A method to guide choices by analysing vast amounts of user behaviour and preference data.
  - Predicts preferences based on similarities between users.
  - Does not rely on understanding the content itself but rather on user interactions.
  - Operates on the principle: if users had similar preferences in the past, they're likely to have similar ones in the future.

Online retail

E-commerce recommendation algorithms operate in a challenging environment.

- **Volume of Data**: Major retailers deal with vast data from millions of customers and diverse catalog items.
- **Speed Requirement**: Recommendations often need to be near-instant, within half a second, while still being precise.
- **Customer Data Disparity**: New customers offer limited data from a few interactions, whereas repeat customers might provide extensive histories.
- **Dynamic Data**: With every customer interaction, fresh data gets generated, and algorithms must adapt promptly.

Recommendation Algorithms

- Two well-known types
- **User-Based Recommendation**:
  - Considers similarity between users.
  - Analyses user behaviours (e.g., clicks, purchases, ratings).
  - Recommends items liked by similar users.
  - Used by Platforms like Netflix, Youtube, Facebook, Twitter, Goodreads etc.
- **Item-Based Recommendation**:
  - Focuses on user interactions with items (e.g., books, movies).
  - Suggests items similar to those a user interacted with.
  - Used by platforms like Amazon.com and E-Bay.
  - E.g. Items bought together; customers who bought this also bought xyz.
  - An item's similarity is based on sets of purchasers; high Jaccard Similarity indicates similar items.
- Both rely on similarity functions e.g. Jaccard Measure
- **Jaccard Similarity**: Even a 20% similarity can indicate similar tastes due to the vast amount of data. Lower similarities can still be significant.

Recommendation Based on User Ratings

- Applications catalogue user ratings for every transaction. E.g. MovieLens, Netflix, TripAdvisor, Yelp etc
- They use rating similarities and customer similarities to suggest new products or experience.

MovieLens

- Service by: GroupLens Research at the University of Minnesota.
- **Functionality**: Users rate movies they like or dislike, and based on this input, the system offers movie recommendations
- **Technique**:
  - Employs collaborative filtering.
  - It pairs users with similar movie opinions, creating a “neighborhood” of like-minded users.
  - This neighborhood's ratings inform recommendations.

Recommendation Based on User Ratings

MovieLens

- Two classes of entities – users and items
- **Data Representation**: Utilizes a utility matrix where each cell represents a user's rating for a specific movie.
  - Ratings range from 1 to 5 stars.
  - Matrix is mostly Sparse – most of the entries are unknown
- Uses Jaccard Similarity
- Sets are made as –
  - If Bob has HP1 rating as 5 and Ann has 4 then
  - Bob => {HP1, HP1, HP1, HP1, HP1} → 5 stars
  - Ann => {HP1, HP1, HP1, HP1} → 4 stars
  - Intersection Ann ∩ Bob => 4 (minimum of 5 and 4)
  - Union Ann ∪ Bob => 5 + 4 => 9
  - Jaccard Similarity => 4/9 => 0.45

	HP1	HP2	HP3	SW1	SW2	SW3
Ann	4			1		
Bob	5	5	4			
Carl				4	5	
Doug		3				3

Recommendation Based on User Ratings

MovieLens

- Significance
  - Since the highest possible Jaccard Similarity is ½ = 0.5
  - So even 0.3 or 30% score is quite good.

Challenges

- It may be difficult to detect similarities between movies and users
  - Because we have little information about movie-item pairs in the sparse utility matrix
- Even when two movie belongs to same genre, there are likely few users who bought/rated both
- Similarly, if two users like a genre, they may not have bought any movie from the genre

Recommendation Based on User Ratings

MovieLens

To address this –

- Clustering of Movies
- Matrix revision as
  - Columns now represent – cluster of movies
  - Rating - average rating per cluster
  - There may be blank cells
  - Because a user may have not rated any of the movie from cluster
- Other similarity measures are used to identify similar items

	HP	SW
Ann	4	1
Bob	4.67	
Carl		4.5
Doug	3	3

Distance Measures

Distance Measures

- Similarity is crucial in determining how alike two data objects are.
- In Data mining this is represented as a Distance measure

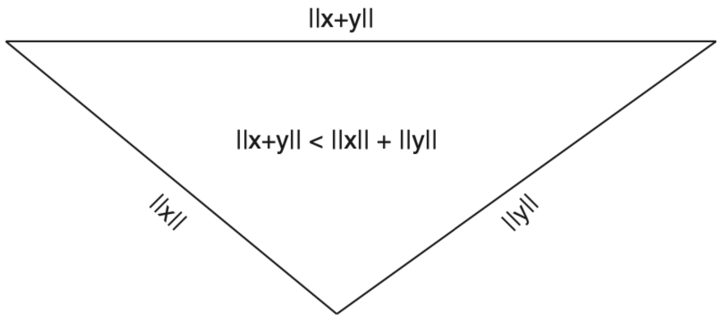
- Small Distance => High similarity between items
- Large Distance => Low similarity

- A distance measure indicates the degree of dissimilarity between two items
- Distance is a subjective measure and highly depends on domain and applications

Distance Metric

- Numerical Measure of how different two data objects are.
- It is a function which maps pairs of objects to real values
  - Lower when objects are more alike
  - Minimum distance is 0, when comparing an object with itself
  - Upper limit varies.

More formally, a distance d is a distance metric if it is a function from pairs of objects to real number such that -



1. d(x, y) > 0. (**Non-negativity**)

2. d(x, y) = 0 iff x = y. (**Identity**)

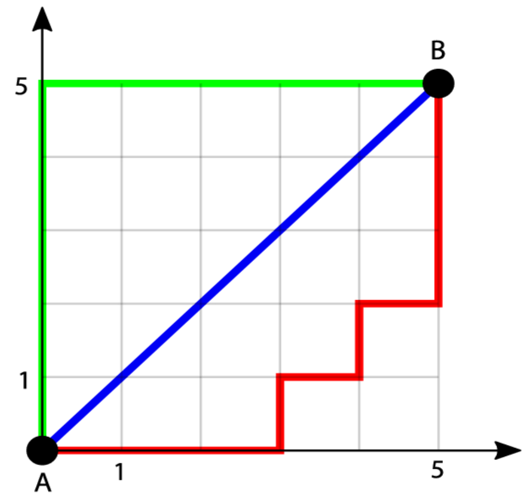
3. d(x, y) = d(y, x). (**Symmetry**)

4. d(x, y) < d(x, z) + d(z, y). (**Triangle inequality**)

The triangle inequality property guarantees that the distance function is well-behaved. It indicates that the direct connection is the shortest distance between two points.

Euclidean Distance

- Easiest measure is to compute Manhattan Distance
- Manhattan distance
  - Also known as cab-driver
  - Represents grid-line travel, like traveling through Manhattan's block-based street layout.
  - Consider two points –(x1,y1), (x2,y2)
    - Distance → Calculated in a 2D space as |x1−x2|+|y1−y2|
- **Euclidean Space**:
  - It's an n-dimensional space where each data point is a vector of n real numbers.
- Manhattan Distance measure is a special case of a distance measure in a "Euclidean Space"



Example -

- Consider Two points. A (x1,y1)=(2,3) and B (x2,y2)=(4,1)
- Manhattan Distance=|2−4| + |3−1| = 2+2 = 4 units
- This implies that from point A to B you would need to walk 4 units.

Euclidean Distance

- **Euclidean Space**:
  - It's an n-dimensional space where each data point is a vector of n real numbers.
- The general form of distance measure used for Euclidean spaces is called *Lr* norm.
- For any constant *r*, we can define the ***Lr***-norm to be the "distance measure" *d* defined by

$$d([x_1, x_2, \ldots, x_n], [y_1, y_2, \ldots, y_n]) = \left(\sum_{i=1}^{n} |x_i - y_i|^r\right)^{1/r}$$

- This general form of distance measure is called **Minkowski** measure
- Manhattan distance is a special case where r = 1. Also known as L1-Norm
- The conventional distance measure in this space is referred to as - L2-norm

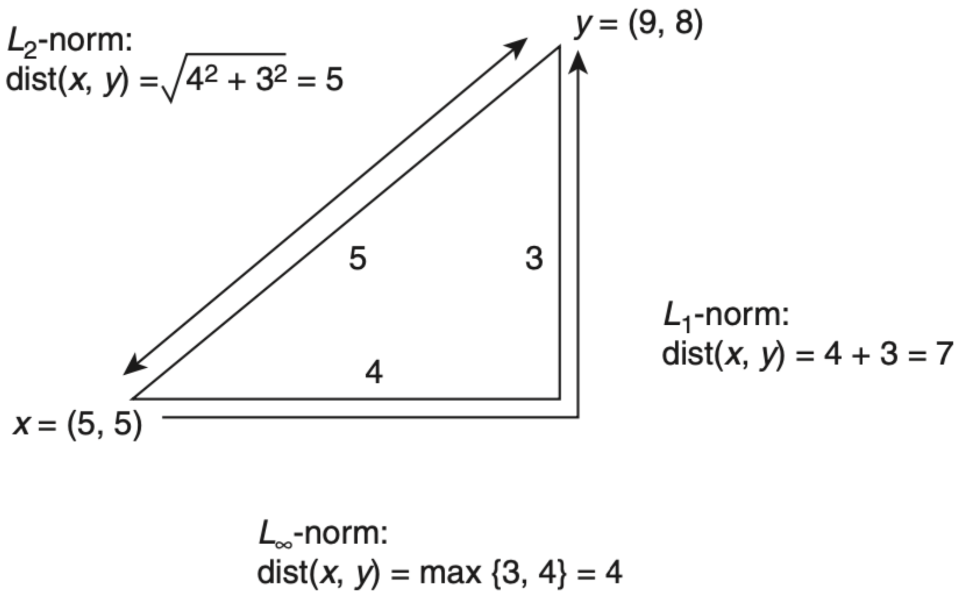
$$d([x_1, x_2, \ldots, x_n], [y_1, y_2, \ldots, y_n]) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- This is also called as Euclidean Distance

Euclidean Distance

- The Euclidean distance is derived from Pythagoras theorem of straight-line distance
- We are taking a positive root – meaning that distance can never be negative
- When two points are identical, distance becomes zero
- The measure is symmetric

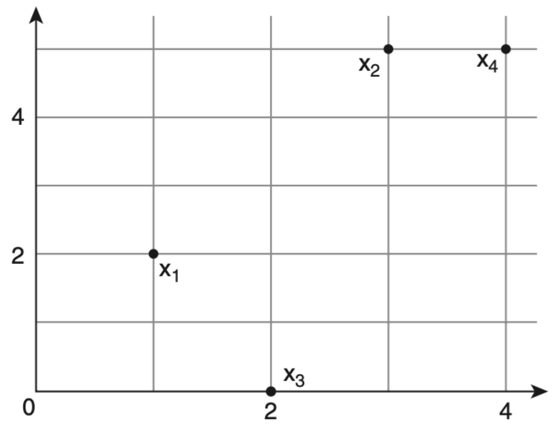
$$(x_i - y_i)^2 = (y_i - x_i)^2.$$



- ***L*∞**-norm
- Distance measure in which r approaches infinity
- As r gets larger, only the dimension with largest difference matters
- So, it is defined as maximum of $|x_i - y_i|$

Euclidean Distance

Data points



Data matrix

Point	Attribute1	Attribute2
x_1	1	2
x_2	3	5
x_3	2	0
x_4	4	5

Manhattan distance

	x_1	x_2	x_3	x_4
x_1	0			
x_2	5	0		
x_3	3	6	0	
x_4	6	1	7	0

Euclidean distance

	x_1	x_2	x_3	x_4
x_1	0			
x_2	3.61	0		
x_3	2.24	5.1	0	
x_4	4.24	1	5.39	0

Jaccard Distance

- Measures dissimilarity between the sample sets
- Complimentary to the Jaccard coefficient
- Obtained by subtracting Jaccard coefficient from 1

$$d_J(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

- It has all the constraints of a distance measure
- Non-negative => Size of intersection will always be less than or equal to union
- Size of union and intersection can never be same except when both sets are same
- Jaccard similarity is 1 only when same sets are used.
- Union and intersection of two sets are always symmetric. A∪B=B∪A and A∩B=B∩A.

Cosine Distance

- Cosine distance between two points is the angle formed between their vectors
- Angle always lies between 0 to 180 degrees. – Regardless of number of dimensions
- Smaller the angle – higher the cosine similarity
- Cosine Distance = 1 – Cosine Similarity

Cosine Similarity

The cosine similarity between two vectors $\vec{a}$ and $\vec{b}$ is calculated as follows:

$$\text{cosine\_similarity}(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{||\vec{a}||_2 \cdot ||\vec{b}||_2}$$

Where:

- $\vec{a} \cdot \vec{b}$ is the dot product of the two vectors.
- $||\vec{a}||_2$ and $||\vec{b}||_2$ are the L2-norms (Euclidean norms) of the vectors, calculated as the square root of the sum of the squares of the components.

Cosine Distance

Example

Let's consider two 2-dimensional vectors $\vec{a} = [1, 2]$ and $\vec{b} = [2, 3]$:

Calculating Cosine Similarity:

1. **Dot Product:** $\vec{a} \cdot \vec{b} = 1 * 2 + 2 * 3 = 8$
2. **L2-norms:** $||\vec{a}||_2 = \sqrt{1^2 + 2^2} = \sqrt{5}$ and $||\vec{b}||_2 = \sqrt{2^2 + 3^2} = \sqrt{13}$
3. **Cosine Similarity:** $\frac{8}{\sqrt{5} * \sqrt{13}} \approx 0.94$

Calculating Cosine Distance:

- **Cosine Distance:** $1 - 0.94 = 0.06$

Cosine Distance

It is a distance measure because

- Since the values are taken in range of (0, 180), there cannot be any negative distance
- Angle between two vectors is 0 only if they are in same direction
- Angle between two vectors satisfies symmetry. Angle between x and y is same as angle between y and x
- Cosine distance also satisfies triangle inequality.
  - One way to rotate from x → y is to rotate from x → z and then rotate from z → y
  - The sum of these two rotations cannot be less than rotation directly from x → y

Edit Distance

- Primarily used for comparing the similarity between two strings
- Determines the minimum number of single character edits required to change one string into the other

Examples:

- Between “Hello” and “Jello”, the Edit Distance is 1, because only one substitution is needed.
- Between “good” and “goodbye”, the Edit Distance is 3, as three insertions are needed.
- Between any string and itself, the Edit Distance is 0.

Edit Distance

Edit Distance Formula -

- $d(x,y) = |x| + |y| - 2 \cdot LCS(x,y)$

- **x, y**: Two strings being compared.
- **LCS(x, y)**: Longest Common Subsequence of x and y.
- **d(x, y)**: Edit Distance between strings x and y.

Let $x = abcde$ and $y = bcduve$. Turn x into y by deleting a; then insert u and v after d. **Edit-distance = 3.**
Now $\text{LCS}(x,y) = bcde$. So

$$|x| + |y| - 2|\text{LCS}(x,y)| = 5 + 6 - 2 * 4 = 3$$

Edit Distance

Properties:

1. **Non-negativity**: The number of insertions and deletions needed to convert string x into string y can never be negative.
2. Edit Distance between two identical strings is zero.
3. **Symmetry**: Edit Distance is symmetric. The number of edits to convert string x to y is the same as converting string y to x.
4. **Triangle Inequality**: The sum of the Edit Distances from string x to z and from z to y is always greater than or equal to the Edit Distance from string x to y.

Applications:

- Spell correction,
- DNA sequencing, and
- Other areas where understanding the similarity or difference between strings is crucial.

Hamming Distance

- Measure used to find the difference between two Boolean vectors of equal length
- The number of items in which two items differ is the Hamming Distance between them

It is a distance measure

1. **Non-negativity**: The Hamming Distance can never be negative.
2. The Hamming Distance is zero only if the vectors are identical.
3. **Symmetry**: The Hamming Distance is symmetrical, meaning the order of the vectors does not affect the distance.
4. **Triangle Inequality**: If x is the number of components in which p and r differ, and y is the number of components in which r and q differ, then p and q cannot differ in more than x+y components.

Applications

- Used in Error detection and error correction
  - To measure error rates
  - To correct errors in data transmission and storage

Hamming Distance

Let's consider two Boolean vectors, p and q:

- $p = [0, 1, 0, 1, 1]$
- $q = [1, 0, 0, 1, 1]$

Calculation:

1. Compare the first element: 0 (from p) is different from 1 (from q).
2. Compare the second element: 1 (from p) is different from 0 (from q).
3. The third and the last two elements are the same in both vectors.

Result:

So, the Hamming Distance between p and q is 2, as there are two positions at which the elements are different.