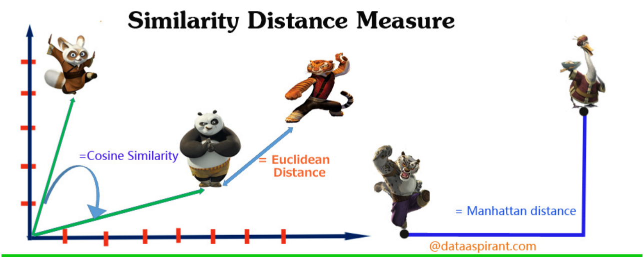
SIMILARITY METRICS

STARTER GUIDE



swipe right →

Euclidean Distance

Straight line distance between two points in euclidean space

Formula: For two points x and y with n dimensions,

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- Used in clustering algorithms, image processing, biometrics
- Computationally efficient
- versatile for wide range of applications

BUT

- Sensitive to scale and outliers
- Poor performance in high dimensions
- Assumes linearity

swipe right →

Manhattan Distance

Distance between two points in a grid based path

Formula: For two points x and y with n dimensions,

$$d_{\text{Manhattan}}(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

- Used in routing, robotics, and any grid-based scenarios
- It is more effective than euclidean in high-dimensional spaces
- versatile for wide range of applications

BUT

- Not suitable for non-grid worlds
- doesn't consider diognal relationships between two points
- Too much emphasis of axal distance

slide right

Minkowski Distance

Generalization of Euclidean and Manhattan distance

Formula: For two points x and y with n dimensions,

$$D(X, Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

- If p=1 it is manhattan distance; if p=2 it is euclidean distance
- p can be considered hyper-parameter in training
- Higher values of p can be used to emphasize larger distances between points

swipe right→

Cosine Similarity

measures the cosine of angle between two vectors

Formula: For two points x and y with n dimensions,

$$\mathrm{d}(x, y) = \frac{x \cdot y}{||x|| \cdot ||y||} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}$$

- Used extensively in NLP tasks like text clustering and classification
- Also used for image-image and image-text retrieval
- It captures semantic similarity, insensitive to magnitude and scaling, and works great in high-dimensional data

BUT

- Sensitive to noise
- May not work well for non-vector data

swipe right →

Hamming distance

measures the positional differences between two strings of equal length

Formula: For two strings s and t of length n,

$$d_H(s, t) = \sum_{i=1}^{n} \left| s_i - t_i \right|$$

- Used in error detection, DNA sequencing and networking
- Works well for categorical data with small no. of categories
- Computationally efficient

BUT

- Doesn't capture any context and semantics
- Doesn't work well for continuous data

swipe right ⟶

Jaccard Index

Size of intersection divided by size of union (IoU)

Formula:

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

- Used in image segmentation, information retreival, and bioinformatics
- Effective for sparse data
- Insensitive to size of A and B

BUT

- Doesn't consider the magnitude of elements in A and B
- Doesn't work well for continuous data
- Insensitive to small overlaps

swipe right→

THANK YOU FOR READING TILL THE END

SEE YOU IN THE NEXT ONE!

FOLLOW ME FOR MORE CONTENT LIKE THIS