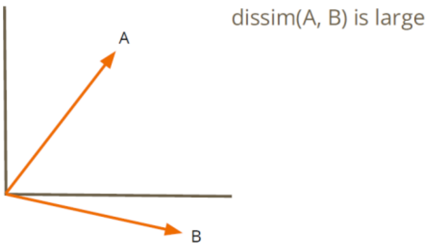
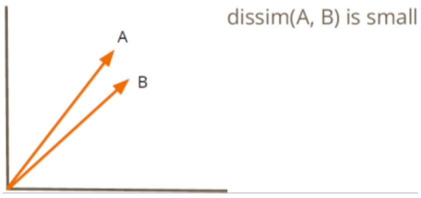
Lecture 03

Dissimilarity

A dissimilarity function is a function that takes two data points and returns a large value if objects are dissimilar.





A special type of dissimilarity function is a **distance** function.

Minkowski Distance

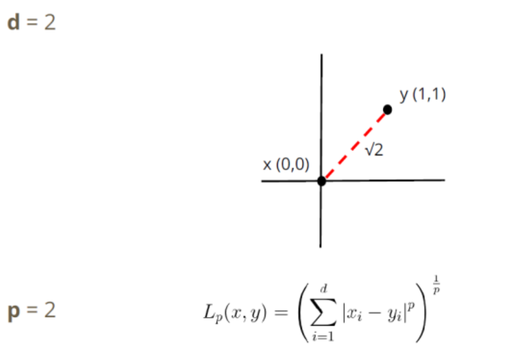
For x, y points in d-dimensional real space I.e. $x = [x_1, \ldots, x_d]$ and $y = [y_1, \ldots, y_d]$ for $p \geq 1$

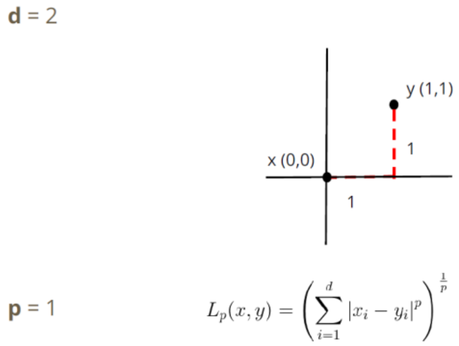
$$L_p(x,y) = \left(\sum_{i=1}^{d} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

Where,

When $p = 2$ is **Euclidean Distance**

When $p = 1$ is **Manhattan Distance**





Example 1: 2D Points

Let:

- $X = (2, 4)$
- $Y = (5, 8)$

Manhattan Distance ($p = 1$):

$$D = |2-5| + |4-8| = 3 + 4 = 7$$

Euclidean Distance ($p = 2$):

$$D = \sqrt{(2-5)^2 + (4-8)^2} = \sqrt{9+16} = \sqrt{25} = 5$$

Chebyshev Distance ($p \to \infty$):

$$D = \max(|2-5|, |4-8|) = \max(3, 4) = 4$$

Example 2: Word Frequency Vectors

Suppose we have two documents represented as word frequency vectors:

- Doc1 = $[1, 2, 3]$
- Doc2 = $[2, 4, 6]$

Manhattan Distance ($p = 1$):

$$D = |1-2| + |2-4| + |3-6| = 1 + 2 + 3 = 6$$

Euclidean Distance ($p = 2$):

$$D = \sqrt{(1-2)^2 + (2-4)^2 + (3-6)^2} = \sqrt{1+4+9} = \sqrt{14} \approx 3.74$$

Chebyshev Distance ($p \to \infty$):

$$D = \max(|1-2|, |2-4|, |3-6|) = \max(1, 2, 3) = 3$$

Jaccard Similarity

Is the measure of similarity between two sets.

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

Where,

A and B are two sets

$|A \cap B|$ is the number of elements in the intersection (common elements),

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

<u>*Example:*</u>

$A = \{1,2,3,4\}$

$B = \{3,4,5,6\}$

$$A \cap B = \{3,4\}; \ |A \cap B| = 2$$
$$A \cup B = \{1,2,3,4,5,6\}; \ |A \cup B| = 6$$
$$J(A,B) = \frac{2}{6} = 0.33$$

The Jaccard Distance is the **complement** of the Jaccard Similarity.

$$D(A,B) = 1 - J(A,B)$$

$D(A,B) = 1 - 0.33 = 0.67$ (Above Example)

- Doc1: "*The cat sat on the mat*"
- Doc2: "*The cat lay on the rug*"

Convert to sets of words:

- $D_1 = \{the, cat, sat, on, mat\}$
- $D_2 = \{the, cat, lay, on, rug\}$

📌 Step 2: Compute Jaccard Similarity

$$J(D_1, D_2) = \frac{|D_1 \cap D_2|}{|D_1 \cup D_2|}$$

- Intersection: $\{the, cat, on\}$ → size = 3
- Union: $\{the, cat, sat, on, mat, lay, rug\}$ → size = 7

$$J(D_1, D_2) = \frac{3}{7} \approx 0.43$$

📌 Cosine Distance

Sometimes we also use **cosine distance:**

$$\text{Cosine Distance} = 1 - \text{Cosine Similarity}$$

Here:

$$1 - 0.6 = 0.4$$

Cosine Similarity

- Cosine similarity measures the cosine of the angle between two vectors in a high-dimensional space.
- It captures **orientation** (direction), not magnitude.
- Often used for text represented as **word frequency vectors** (Bag of Words, TF-IDF).

$$Cosine\ similarity\ (A, B) = \frac{A \cdot B}{\|A\|\|B\|}$$

Where,
$A \cdot B\ =\ dot\ product\ of\ vectors\ A\ and\ B$
$\|\ A\ \|\ and\ \ \|\ B\ \| =\ Euclidean\ norms\ (lengths\ of\ vectors)$

Example:

Let:

- $A = (1, 2)$
- $B = (2, 3)$

Step 1: Dot Product

$$A \cdot B = (1)(2) + (2)(3) = 2 + 6 = 8$$

Step 2: Magnitudes

$$\|A\| = \sqrt{1^2 + 2^2} = \sqrt{1 + 4} = \sqrt{5}$$

$$\|B\| = \sqrt{2^2 + 3^2} = \sqrt{4 + 9} = \sqrt{13}$$

Step 3: Cosine Similarity

$$\frac{8}{\sqrt{5}\sqrt{13}} = \frac{8}{\sqrt{65}} \approx 0.99$$

👉 These two vectors are **almost pointing in the same direction** (very high similarity).

- Doc1: "*The cat sat on the mat*"
- Doc2: "*The cat lay on the rug*"

Step 1: Vocabulary

Unique words = $\{the, cat, sat, on, mat, lay, rug\}$

Step 2: Word Count Vectors

- Doc1 → $[1, 1, 1, 1, 1, 0, 0]$
- Doc2 → $[1, 1, 0, 1, 0, 1, 1]$

(Each position corresponds to a word in the vocabulary.)

Step 3: Compute Cosine Similarity

- Dot product:

$$A \cdot B = (1)(1) + (1)(1) + (1)(0) + (1)(1) + (1)(0) + (0)(1) + (0)(1) = 3$$

- Norms:

$$\|A\| = \sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2} = \sqrt{5}$$

$$\|B\| = \sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2} = \sqrt{5}$$

- Cosine similarity:

$$\frac{3}{\sqrt{5}\sqrt{5}} = \frac{3}{5} = 0.6$$

📌 **Cosine Distance**

Sometimes we also use **cosine distance:**

$$\text{Cosine Distance} = 1 - \text{Cosine Similarity}$$

Here:

$$1 - 0.6 = 0.4$$

Correlation Coefficient

The correlation coefficient (most commonly the Pearson correlation coefficient) measures the strength and direction of the linear relationship between two variables.

For two variables X and Y with n data points:

$$r = \frac{\sum_{i=1}^{n}(X_i - \hat{X})(Y_i - \hat{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \hat{X})^2 \sum_{i=i}^{n}(Y_i - \hat{Y})^2}}$$

When you plot data points (X, Y):

- If the points lie **close to a straight upward-sloping line**, r will be **positive** (close to +1).
- If they lie **close to a downward-sloping line**, r will be **negative** (close to −1).
- If they are **widely scattered** or show **no clear direction**, r will be **near 0**.

Change on Correlation Co-efficient

Transformation	Effect on r
Multiply X by positive constant	Unchanged
Multiply X by negative constant	Sign flips
Add or subtract constant from XXX	Unchanged