

# PROXIMITY MEASURES- NON METRIC METHODS

Dr. Umarani Jayaraman  
Assistant Professor



INDIAN INSTITUTE OF INFORMATION TECHNOLOGY,  
DESIGN AND MANUFACTURING,  
KANCHEEPURAM

# Topic

---

- Cosine distance

# Non-metric Similarity Function

- Similarity functions which **do not obey** either the properties such as positive reflexivity, symmetry or triangle inequality come under this category.
- Usually these similarity functions are useful for comparing **images or text documents**.
- They are robust to outliers or to extremely noisy data.

# Motivation

- Image 1 is 4 times of Image 2
- Say there are 8 different gray levels in these images
  - ▣  $H_2 = [a, b, c, d, e, f, g, h]$
  - ▣  $H_1 = [4a, 4b, 4c, 4d, 4e, 4f, 4g, 4h]$
- Based on histograms comparison
  - ▣ Are these two images same?



Image - 1 (200 x 200)

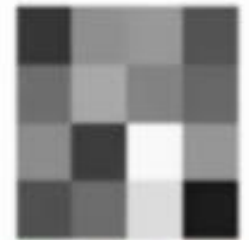


Image - 2 (50 x 50)

# Motivation

- Image 1 is 4 times of Image 2
- Say there are 8 different gray values in these images
  - ▣  $H_2 = [a, b, c, d, e, f, g, h]$
  - ▣  $H_1 = [4a, 4b, 4c, 4d, 4e, 4f, 4g, 4h]$
- Based on histogram comparison
- Are these two images similar?



Image 1 (200x200)

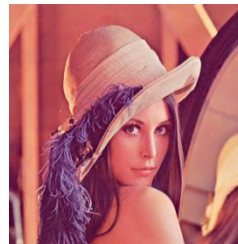


Image 2 (50x50)



Image - 1 (200 x 200)

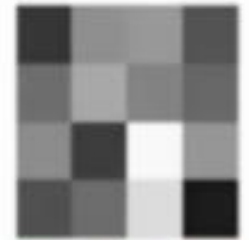


Image - 2 (50 x 50)

# Motivation

- Use of Euclidean distance-No
- Use of Manhattan distance-No
- Use of Cosine distance- Yes



Image 1 (200x200)

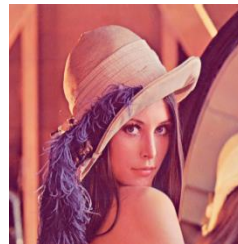


Image 2 (50x50)

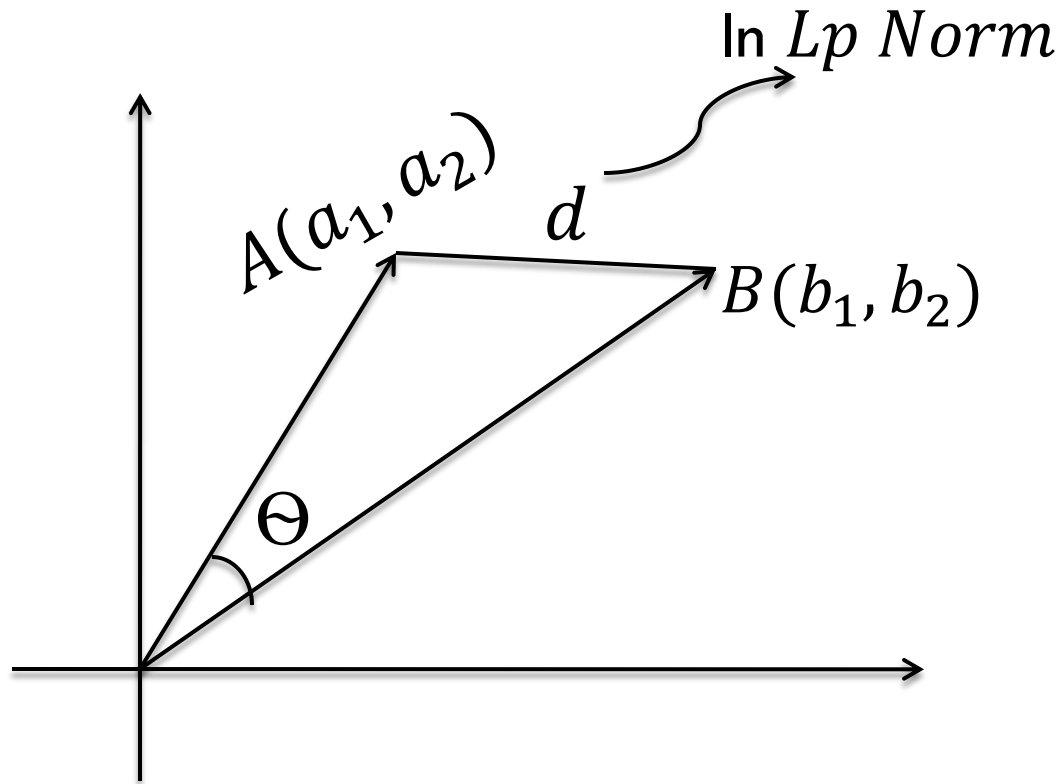


Image - 1 (200 x 200)



Image - 2 (50 x 50)

# Cosine Distance



- ❑ Cosine considers the angle between vectors (not taking magnitude into account).
- ❑  $d$  = Euclidean Distance
- ❑  $d_\theta = 1 - \cos(\theta)$ , cosine distance

# Cosine Distance: Definition

- The Cosine of two non-zero vectors can be derived by using the Euclidean dot product formula:
- $A \cdot B = ||A|| ||B|| \cos \theta$
- $A = (a_1, a_2, \dots, a_d)$
- $B = (b_1, b_2, \dots, b_d)$



# Cosine Distance: Definition

- $A \cdot B = ||A|| ||B|| \cos \theta$
- Cosine similarity is given by  $\cos \theta$
- The dissimilarity between the two vectors 'A' and 'B' is given by

$$\begin{aligned} d_{\theta} = 1 - \cos \theta &= 1 - \frac{A \cdot B}{||A|| ||B||} \\ &= 1 - \frac{\sum_{i=1}^d a_i b_i}{\sqrt{\sum_{i=1}^d a_i^2} \sqrt{\sum_{i=1}^d b_i^2}} \end{aligned}$$

Only **angle** relevant, **not vector lengths**

# Example: Cosine distance

- Euclidean distance is similar to using a ruler to actually measure the distance.
- E.g.

$$a = [1, 2, 3]$$

$$b = [4, -5, 6]$$

$$\cos \theta = \frac{a \cdot b}{||a|| ||b||} = \frac{1 \cdot 4 + 2 \cdot (-5) + 3 \cdot 6}{\sqrt{1^2 + 2^2 + 3^2} \sqrt{4^2 + (-5)^2 + 6^2}} = \frac{12}{\sqrt{14} \sqrt{77}}$$

$$d_{\theta} = 1 - \frac{12}{\sqrt{14} \sqrt{77}}$$

# Cosine Distance

- In cosine distance
- Angle 0 degree (cosine distance = 0)
  - ▣ for identical vectors
- Angle 90 degrees (cosine distance = 1)
  - ▣ for dissimilar vectors

# Cosine distance: Applications

- Used when **magnitude** of the vectors does not have much significance
- Used wherever the **directions are so important** cosine distance is used
- Applications
  - ▣ Image Matching
  - ▣ Document Matching

# Cosine distance: Image Matching



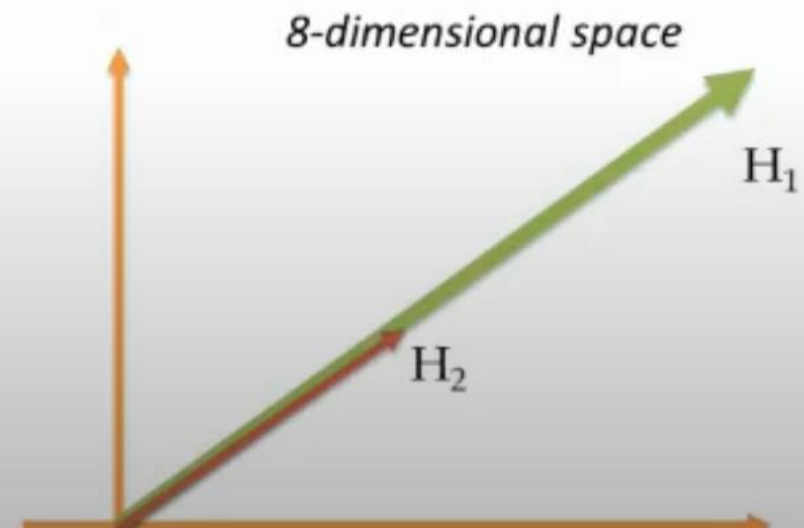
Image - 1 (200 x 200)



Image - 2 (100 x 100)

$$H_2 = [a, b, c, d, e, f, g, h]$$
$$H_1 = [4a, 4b, 4c, 4d, 4e, 4f, 4g, 4h]$$

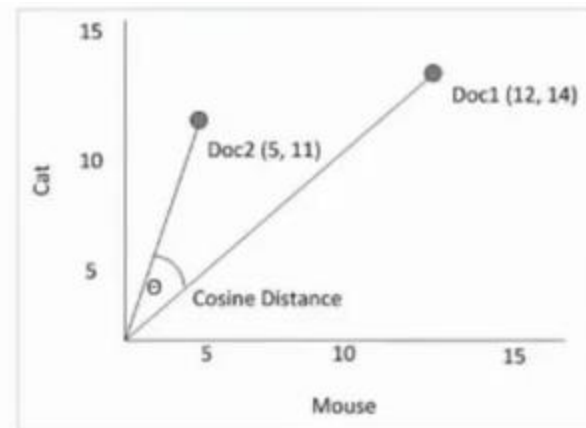
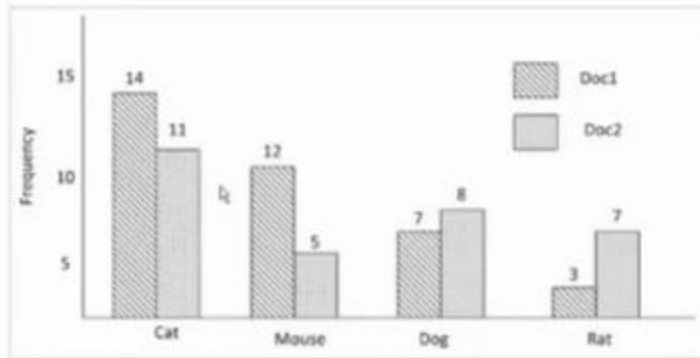
Cosine distance = 0



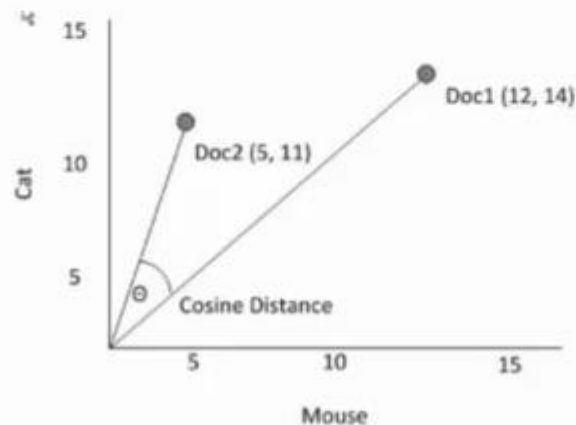
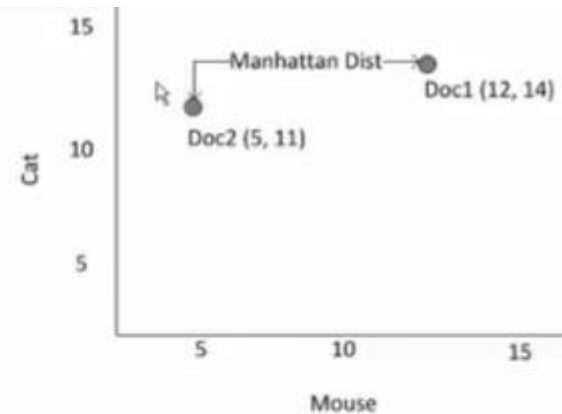
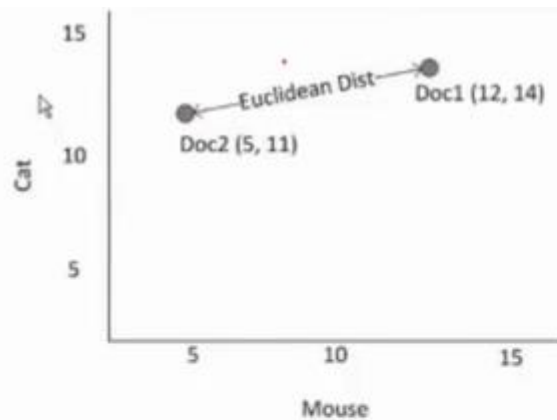
# Cosine distance: Document Matching

## □ Document Matching

- ▣ Two documents may be considered **same** if they **have certain important words with same frequency**.
- ▣ However, a document with, say, **twice** as many occurrences **of all words compared to another document** will be regarded **as identical**.



# Comparison of three distances



# Cosine distance: Document Matching

- Cosine similarity is a measure to find the similarity between two files/documents.

$$file_1 = (0, 3, 0, 0, 2, 0, 0, 2, 0, 5)$$

$$file_2 = (1, 2, 0, 0, 1, 1, 0, 1, 0, 3)$$

$$file_1 \cdot file_2 = 0 \times 1 + 3 \times 2 + \dots + 5 \times 3 \\ = 25$$

$$||d_1|| = \sqrt{42} = 6.481$$

$$||d_2|| = \sqrt{17} = 4.12$$



# Cosine distance: Document Matching

$$\cos(d_1, d_2) = \frac{file_1 \cdot file_2}{||file_1|| ||file_2||}$$
$$\cos(d_1, d_2) = \frac{25}{6.481 \times 4.12}$$
$$D(d_1, d_2) = 1 - 0.94$$
$$= 0.06$$

# Is it metric?

- Angular Similarity

$$\cos \theta = S(A, B) = \frac{A \cdot B}{||A|| ||B||}$$

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

- This  $D(A, B)$  does not satisfy the *triangular inequality*. So, **It is not a metric.**

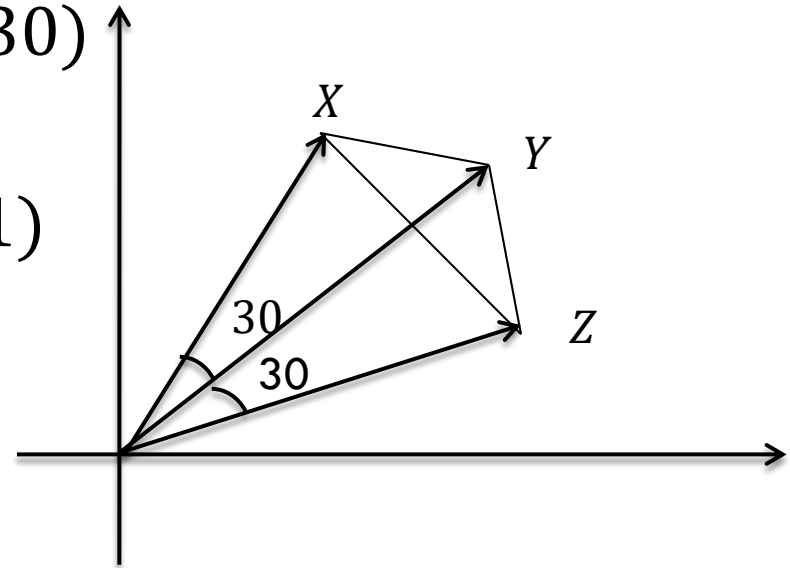
- However, it is symmetric, because

$$\cos \theta = \cos(-\theta)$$

# Cosine Distance: Triangular Inequality

- If  $X, Y$  and  $Z$  are the three vectors in a  $2 - d$  space such that the angle between  $X$  and  $Y$  is  $30^\circ$  and that between  $Y$  and  $Z$  is  $30^\circ$ , then :

- $$\begin{aligned} D(X, Z) &= 1 - S(X, Z) \\ &= 1 - \cos(30 + 30) \\ &= 1 - \cos 60 \\ &= \frac{1}{2} \text{ -----Eqn(1)} \end{aligned}$$



# Cosine Distance: Triangular Inequality

$$\begin{aligned}\square D(X, Y) + D(Y, Z) &= (1 - \cos 30) + (1 - \cos 30) \\ &= \left(1 - \frac{\sqrt{3}}{2}\right) + \left(1 - \frac{\sqrt{3}}{2}\right) \\ &= 2 \left(1 - \frac{\sqrt{3}}{2}\right) \text{-----Eqn(2)}\end{aligned}$$

□ From *Eqn1* and *Eqn2*;

$$\frac{1}{2} \not\leq 2 - \sqrt{3}$$

$$D(X, Z) \not\leq D(X, Y) + D(Y, Z)$$

# Cosine Distance: Triangular Inequality

- Hence, cosine distance is not a metric, as it does not satisfy triangular inequality.

# Is it metric?

- There is a way to convert into a metric.
- If the vectors are **always positive**:

$$\textit{Angular Distance} = \frac{2 \cos^{-1}(\textit{cosine similarity})}{\pi}$$

$$D(A, B) = \frac{2 \cos^{-1} S(A, B)}{\pi}$$

# Summary

---

- Cosine distance
- Applications
- Why it is non metric

THANK YOU