PROXIMITY MEASURES- NON METRIC METHODS

Dr. Umarani Jayaraman
Assistant Professor

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY, DESIGN AND MANUFACTURING, KANCHEEPURAM

Topic

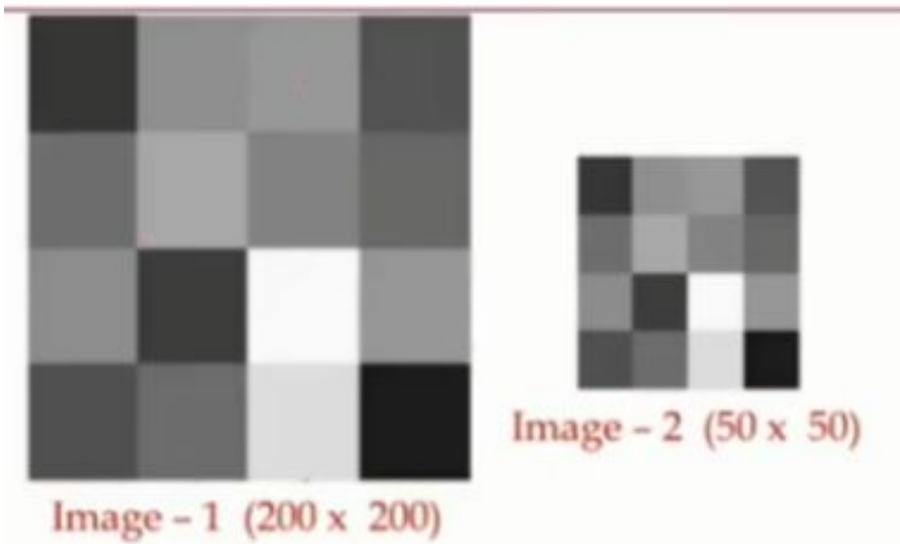
- Cosine distance

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?

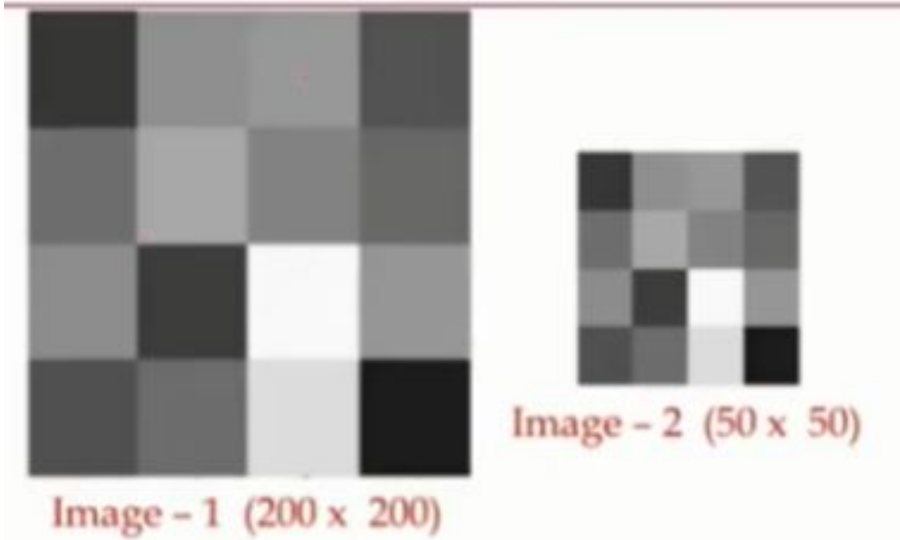


Motivation

- 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)



Motivation

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

Image 1 (200x200)

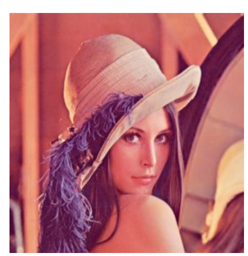
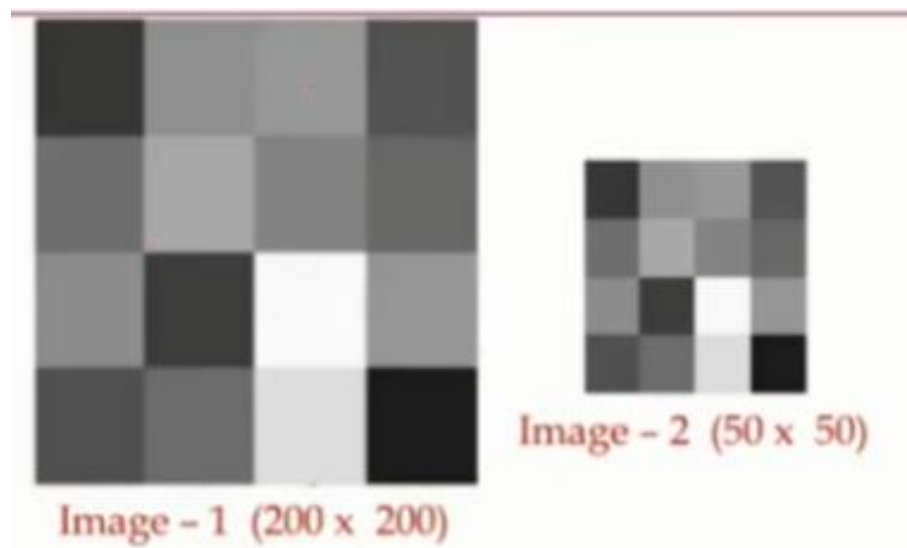
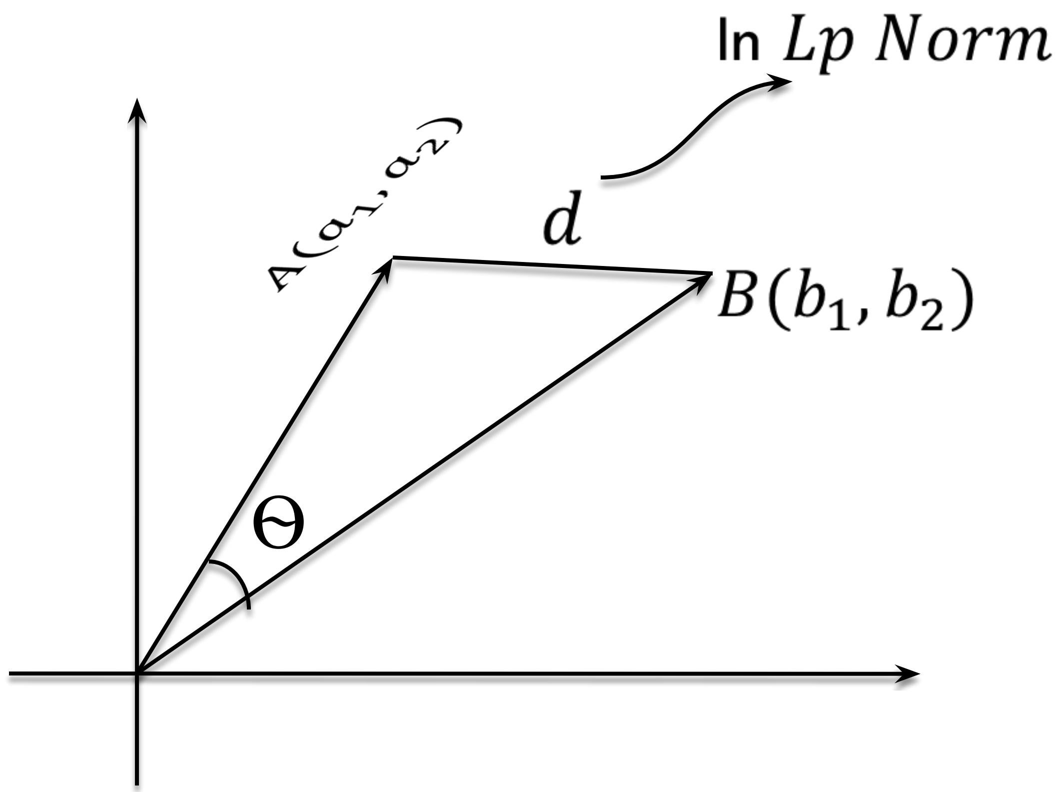


Image 2 (50x50)



Cosine Distance



- Cosine considers the angle between vectors (not taking magnitude into account).
- d= Euclidean Distance
- d_θ = 1- cos(θ), cosine distance

Cosine Distance: Definition

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

Cosine Distance: Definition

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

$$d_\theta = 1 - \cos\theta = 1 - \frac{A.B}{||A||\,||B||}$$

$$= 1 - \frac{\sum_{i=1}^{d} a_i b_i}{\sqrt{\sum_{i=1}^{d} a_i^2}\sqrt{\sum_{i=1}^{d} b_i^2}}$$

Only angle relevant, not vector lengths

Example: Cosine distance

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

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

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

$$\cos\theta = \frac{a.b}{||a||\ ||b||} = \frac{1.4+2.-5+3.6}{\sqrt{1^2+2^2+3^2}\sqrt{4^2+5^2+6^2}} = \frac{12}{\sqrt{14}\sqrt{77}}$$

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

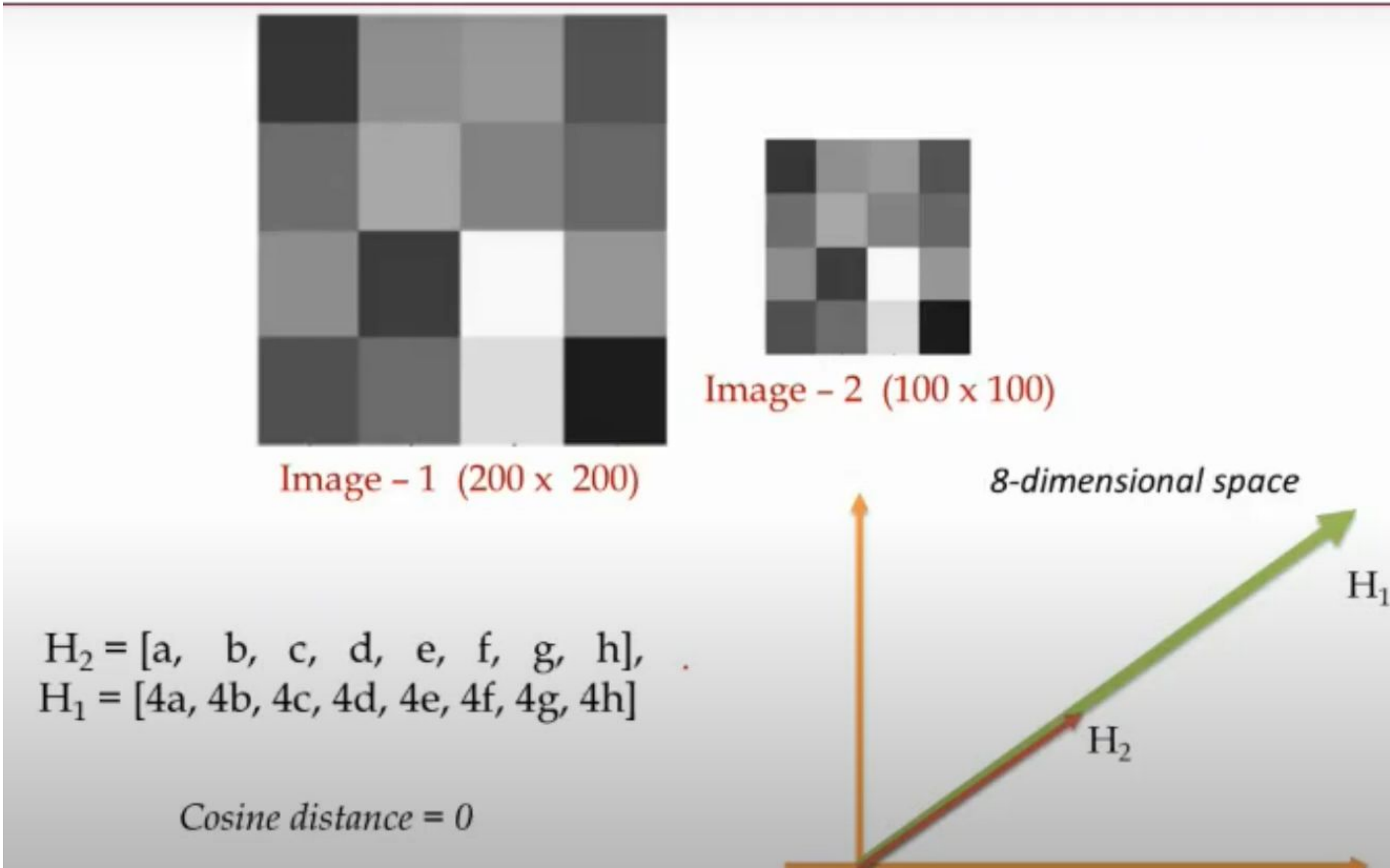
Cosine Distance

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

Cosine distance: Applications

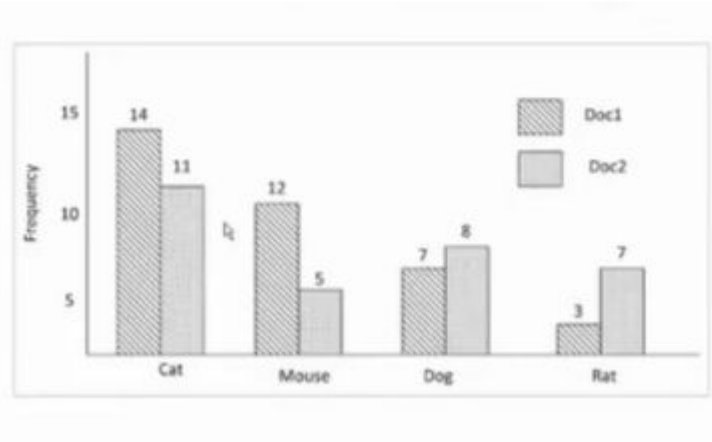
- Used when magnitude of the vectors does not have much significance
- Used wherever the directions are so important cosine distance is used
- Applications
  - Image Matching
  - Document Matching

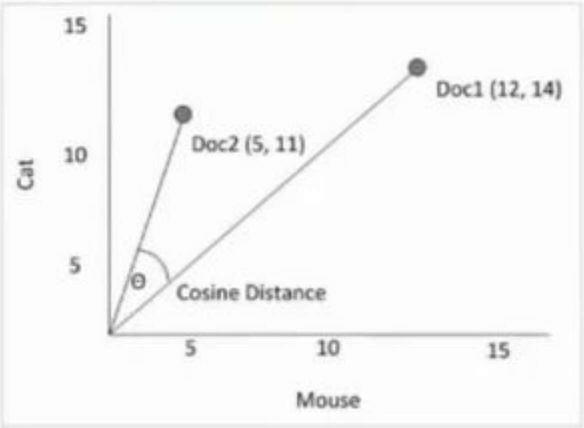
Cosine distance: Image Matching



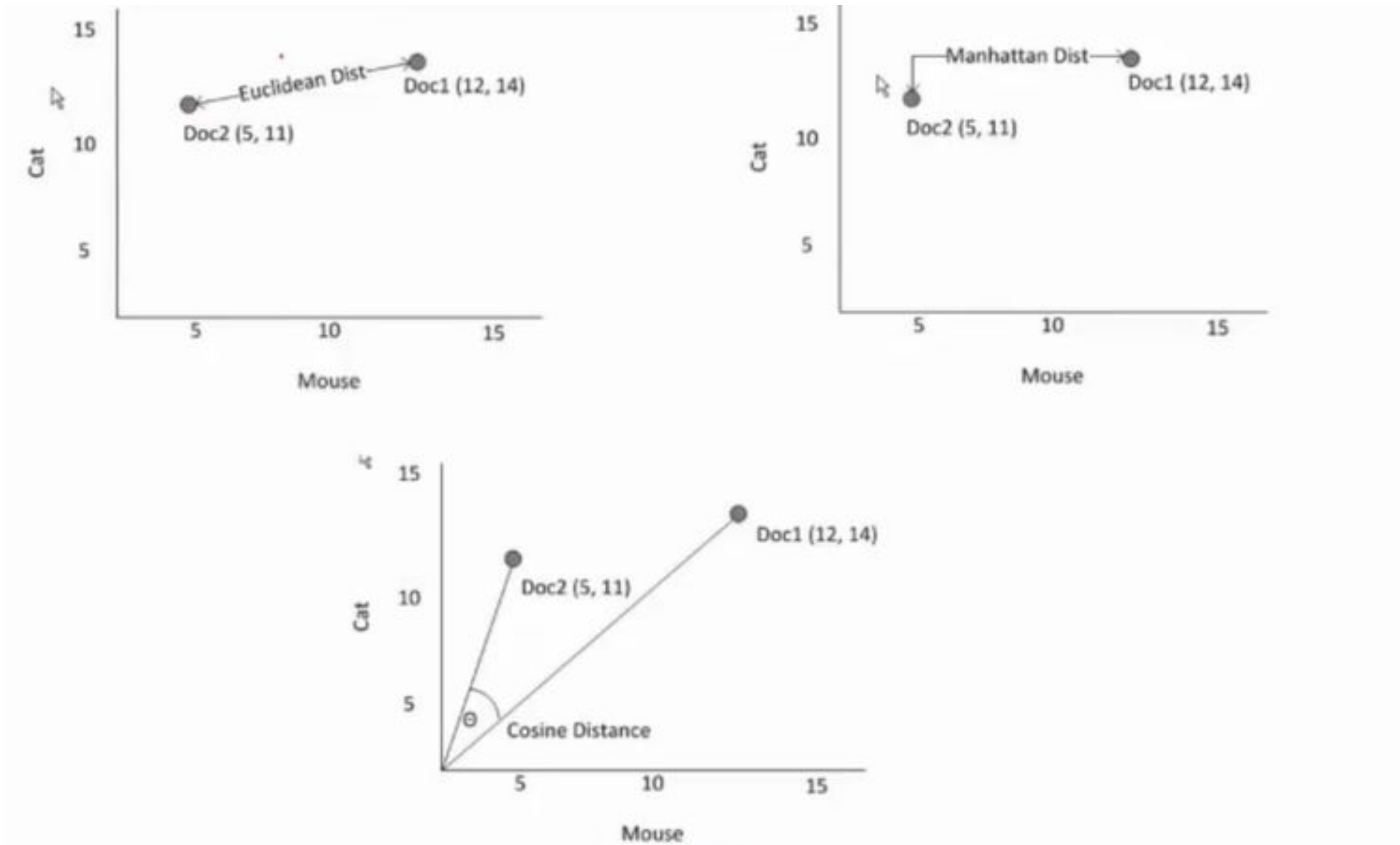
Cosine distance: Document Matching

- 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.

$$file_1 = (0, 3, 0, 0, 2, 0, 0, 2, 0, 5)$$
$$file_2 = (1, 2, 0, 0, 1, 1, 0, 1, 0, 3)$$
$$file_1 . file_2 = 0 \times 1 + 3 \times 2 + \cdots .. + 5 \times 3$$
$$= 25$$
$$||d_1|| = \sqrt{42} = 6.481$$
$$||d_2|| = \sqrt{17} = 4.12$$

Cosine distance: Document Matching

- $$\cos(d_1, d_2) = \frac{file_1 . file_2}{||file_1|| \, ||file_2||}$$

$$\cos(d_1, d_2) = \frac{25}{6.481x4.12}$$

$$D(d_1, d_2) = 1 - 0.94$$

$$= 0.06$$

Is it metric?

- Angular Similarity

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

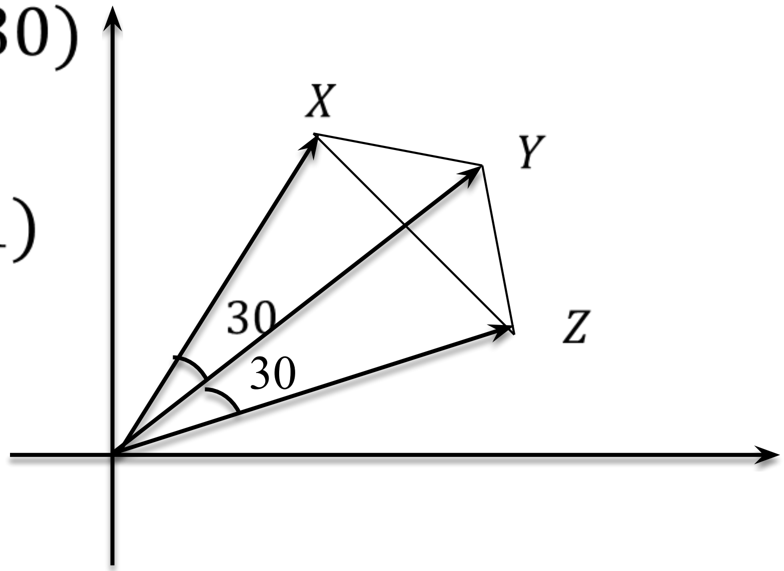
$$D(A,B) = 1 - S(A,B)$$

- 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 $30°$ and that between $Y\ and\ Z$ is $30°$, 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

- $D(X,Y) + D(Y,Z) \; = \; (1 - \cos 30) \; + \; (1 - cos30)$

$$= \left(1 - \frac{\sqrt{3}}{2}\right) + \left(1 - \frac{\sqrt{3}}{2}\right)$$

$$= 2\left(1 - \frac{\sqrt{3}}{2}\right)\text{------}Eqn(2)$$

- From $Eqn1$ and $Eqn2$;

$$\frac{1}{2} \nleq \; 2 - \sqrt{3}$$

$$D(X,Z) \; \nleq \; DX,Y) \; + D(Y,Z)$$

Cosine Distance: Triangular Inequality

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

Is it metric?

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

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

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

Summary

- Cosine distance
- Applications
- Why it is non metric

THANK YOU