ONLINE MASTERS IN **DATA SCIENCE** 

DSC 255 - MACHINE LEARNING FUNDAMENTALS

# **DISTANCE METRICS**

SANJOY DASGUPTA, PROFESSOR



COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE



## **Metric Spaces**

Let  $\mathcal{X}$  be the space in which data lie.

A distance function  $d: \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$  is a **metric** if it satisfies these properties:

- $d(x, y) \ge 0$  (nonnegativity)
- d(x, y) = 0 if and only if x = y
- d(x, y) = d(y, x) (symmetry)
- $d(x, z) \le d(x, y) + d(y, z)$  (triangle inequality)

## Example 1

$$\mathcal{X} = \mathbb{R}^{\mathbf{m}}$$
 and  $d(x, y) = \|x - y\|_{p}$ 

# **Check:**

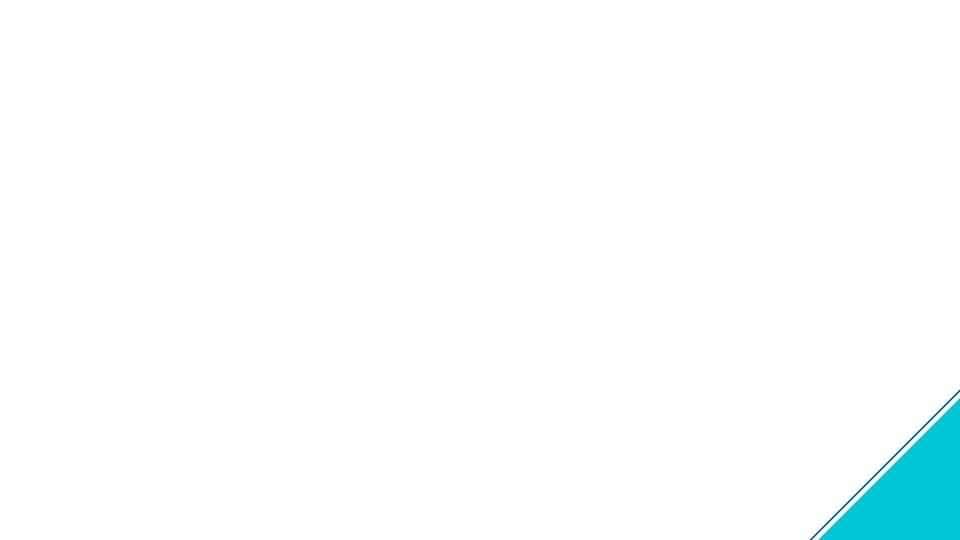
- $d(x, y) \ge 0$  (nonnegativity)
- d(x, y) = 0 if and only if x = y
- d(x, y) = d(y, x) (symmetry)
- $d(x, z) \le d(x, y) + d(y, z)$  (triangle inequality)

#### Example 2

 $\mathcal{X}$  = {strings over some alphabet} and d=edit distance

# Check:

- $d(x, y) \ge 0$  (nonnegativity)
- d(x, y) = 0 if and only if x = y
- d(x, y) = d(y, x) (symmetry)
- $d(x, z) \le d(x, y) + d(y, z)$  (triangle inequality)



#### A non-metric distance function

Let p, q be probability distributions on some set  $\mathcal{X}$ .

The Kullback-Leibler divergence or relative entropy between p, q is:

$$d(p,q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$