ONLINE MASTERS IN **DATA SCIENCE**

DSC 257R - UNSUPERVISED LEARNING

# SIMILARITY FUNCTIONS

SANJOY DASGUPTA, PROFESSOR

UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

# Jaccard Similarity

A notion of similarity between sets:

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

Widely used in information retrieval (e.g., web search).

- In what range does this lie?
- For what $B$ is $s(A, B)$ maximized?

# Cosine Similarity

A notion of similarity between sets:

$$s(x, z) = \frac{x \cdot z}{\|x\|\|z\|}.$$

- In what range does this lie?

- How is it related to the angle between the vectors?

- For what $z$ is $s(x, z)$ maximized?

# Dot Product

Even simpler than the cosine distance:

$$s(x, z) = x \cdot z\,.$$

- In what range does this lie?
- Can $s(x, z)$ ever be larger than $s(x, x)$?

# Kernel Functions

Generalization of dot products:

- Let $\mathcal{X}$ be any instance space
- We say $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ is a **kernel function** if

$$k(x, z) = \emptyset(x) \cdot \emptyset(z)$$

for some mapping $\emptyset : \mathcal{X} \longrightarrow \mathbb{R}^d$, where $1 \leq d \leq \infty$.

Examples:

$$k(x, z) = (x \cdot z)^2$$
$$k(x, z) = e^{-\|x-z\|^2}$$