MEASURING DISTANCE BETWEEN TWO DATA POINTS

FOR CONTINUOUS VARS. IN $\mathbb{R}^2$: $\{(x_1, y_1), (x_1, y_1), \cdots, (x_n, y_n)\}$

① Euclidean – $\varepsilon = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

② Manhattan – $\mathrm{d} = |x_1 - x_2| + |y_2 - y_1|$

③ Minkowski – $\mathrm{d} = \left[(x_2 - x_1)^h + (y_2 - y_1)^h\right]^{\frac{1}{h}}$

④ Supremum – $\mathrm{d} = \max\left(|x_2 - x_1|, |y_2 - y_1|\right)$

⑤ COSINE SIMILARITY – see next page ...

FOR DISCRETE VARS. i.e., all data must be binary 0-1

⑥ Jaccard – $\mathrm{d} = 1\text{-}\frac{|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A} \cup \mathrm{B}|}$

⑦ Hamming – $\mathrm{d} = \sum_{\mathrm{i}=1}^{\mathrm{n}} |\mathrm{x_i} - \mathrm{y_i}|$

⑧ SIMPLE MATCHING COEFFICIENT (SMC)

distance (proximity) is commonly used to measure similarity; thats the whole idea behind KNN which happens to use euclidean distance.

Cosine similarity = $\frac{\langle \vec{a}, \vec{b} \rangle}{\|\vec{a}\| \left\|\vec{b}\right\|}$

- is bounded by -1 and 1
- is similar to the Pearson correlation coef:

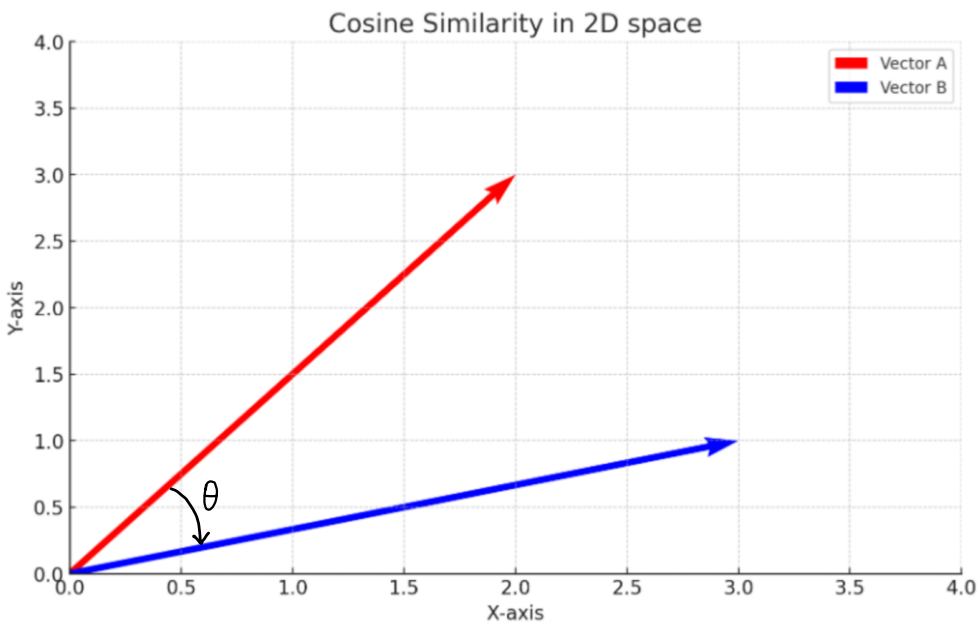
$$\rho_{\vec{a},\vec{b}} = \frac{cov(a,b)}{\sigma_a \sigma_b} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(a_i-\bar{a})(b_i-\bar{b})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(a_i-\bar{a})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(b_i-\bar{b})^2}}$$

- e.g.)

Consider two vectors, $\mathbf{A}$ and $\mathbf{B}$, each with two elements (for simplicity in a 2D space). For instance:

Vector	Element 1	Element 2
A	a1	a2
B	b1	b2

The cosine similarity is calculated by taking the dot product of A and B (which is $a1 \times b1 + a2 \times b2$) and dividing it by the product of the magnitudes of A and B. The magnitude of a vector $\mathbf{V}$ with elements $v1$ and $v2$ is calculated as $\sqrt{v1^2 + v2^2}$.



The chart above illustrates two vectors, A (in red) and B (in blue), originating from the same point (the origin) in a two-dimensional space. The angle between these two vectors represents the basis for calculating their cosine similarity. In this context, a smaller angle corresponds to a higher cosine similarity, indicating that the two vectors are more similar in direction.

Jaccard vs. Hamming Distance

see bookmarked Bing video.

- HAMMING IS USED TO COMPARE TWO SUBJECTS BASED ON HOW MANY **MUTUALLY EXCLUSIVE** ATTRIBUTES THEY SHARE; HAMMING IS TYPICALLY BETTER SUITED FOR THE "STRICTLY" BINARY CASE

FOR EXAMPLE:

	subject 1		subject 2	
sex	male	≠	woman	0
Birthplace	IA	≠	MN	0
race	black	≠	white	0
age	GenZ	=	GenZ	1
Status	middle class	=	middle class	1

subject 1 and subject 2 share two out-of five mutually exclusive, binary attributes. Hence, the Hamming distance between them is $2/5 = .4$ or in other words subject 1 and subject 2 are 40% alike.

- JACCARD, ON THE OTHER HAND, DOES NOT HAVE THE "MUTUALLY EXCLUSIVE" REQUIREMENT OF ITS BASES; JACCARD IS TYPICALLY BETTER SUITED FOR CATEGORICAL DATA ANALYSIS

FOR EXAMPLE:

	subject 1	subject 2	
HOBBIES	Golf, Soccor, Piano, Books, travel	Golf, Basketball, Guitar, Books, travel	$\Rightarrow$ JACCARD DISTANCE $= \frac{3}{5} = .6$