What is Cosine Similarity?

Barum Park

Department of Sociology
Cornell University
b.park@cornell.edu

January 21, 2023

In the paper by Aral and Van Alstyne (2011), a lot of measures depend on the cosine similarity between two vectors. This measure will turn up in future papers we read for this course as well, so let me summarize it here.

Let $v = [v_1, v_2]^\top$ be a two-dimensional vector of real numbers. We often write this as $v \in \mathbb{R}^2$. As mentioned in the previous note on eigenvectors and Markov chains, we can represent v as an arrow pointing from the origin to the point (v_1, v_2). Now, consider the two vectors $x, y \in \mathbb{R}^2$ and the angle θ_{xy} between them, as shown in the figure below.

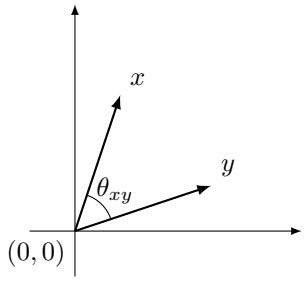

Figure 1: $x, y \in \mathbb{R}^2$ and their angle

The angle between x and y represents how similar x and y are in terms of the *directions* they point to. For example, the coordinates of the two vectors, x and y^*, in Figure 2 will be more distant from each other than those in Figure 1. However, the angle between the two vectors remain the same. So, we can think of the angle between the two vectors as a *standardized* measure of similarity—standardized in the sense that it is not influenced by the length of the vectors.

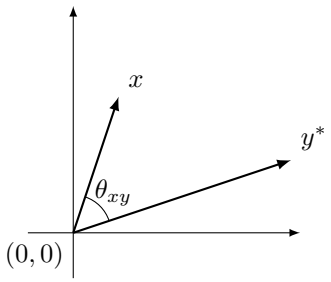

Figure 2: $x, y^* \in \mathbb{R}^2$ and their angle

The *cosine similarity* between two vectors x and y is just the cosine of the angle between x and y for $0 \leq \theta \leq \pi$. That is,

$$\text{Cosine similarity}(x, y) = \cos \theta_{xy},$$

where θ_{xy} is the angle between x and y. This definition is not restricted to $\mathbb{R}^2$ but applies to higher dimensional spaces, as far as we can define angles between vectors in these spaces.[1]

The cosine similarity can be easily calculated using the inner product between two vectors. By definition, the inner product between x and y in $\mathbb{R}^n$ is

$$x^\top y = \|x\|\|y\| \cos\theta_{xy}$$

where $\|x\| = \sqrt{\sum_{i=1}^n x_i^2}$ is the length of the vector x—i.e., the Euclidean distance from the vector x to the origin—and $^\top$ denotes the transpose. Hence, the cosine similarity measure is simply

$$\text{Cosine similarity}(x, y) = \cos\theta_{xy} = \frac{x^\top y}{\|x\|\|y\|} = \frac{\sum_{i=1}^n x_i y_i}{\left(\sqrt{\sum_{i=1}^n x_i^2}\right)\left(\sqrt{\sum_{i=1}^n y_i^2}\right)}.$$

By now, you must have thought "I think I saw this equation before" and, indeed, if you think of x and y as two "variables" in a dataset that are mean-centered—i.e., both x and y have a mean of zero—then

$$\text{Cosine similarity}(x, y) = \frac{\text{Cov}(x, y)}{\text{SD}(x)\text{SD}(y)} = \text{Corr}(x, y).$$

So, whenever you encounter cosine similarity measures, you can think of correlations. On the other hand, whenever you see "variables" represented as vectors—such as so-called "loading plots" in factor analysis—you can look at the angles between the vectors and guess their correlation. For example, if the angle between them is about $90° = \pi/2$, it would mean that they have approximately zero correlation, since $\cos\frac{\pi}{2} = 0$.

[1] Indeed, the measure makes sense for any vector space equipped with an inner product, as the inner product let's us define angles between vectors.