

Non-Euclidean machine learning methods in systems biology

Poleksic A.

University of Northern Iowa, Cedar Falls, Iowa, USA

poleksic@cs.uni.edu

Accurate modeling of the geometry of a biological system enables the discovery of novel relationships from sparse, noisy and biased data. Machine learning methods in systems biology rely on representing biological objects as points in a low dimensional space, known as the *feature* or *latent* space. Traditionally, the native (latent) biological space is assumed to be the classical Euclidean space. While this assumption has been convenient (due to the availability of efficient statistical and numerical optimization tools in the flat space), recent theoretical studies suggest that a better modeling of the proximity of biological entities in a biological system can be achieved using the hyperbolic distance metric. The authors of these studies argue that, in contrast to the Euclidean space, the negatively curved (hyperbolic) space can accommodate the exponential growth of a number of relevant biological features, since the circumference of a hyperbolic circle and the disk area are exponential functions of the radius. Nevertheless, these theoretical claims have never been validated in practice nor they have been built upon to create improved methods for biological relationship inference.

To test the abovementioned hypotheses, we have developed a variant of the classical matrix factorization method, termed *hyperbolic matrix factorization*, that uses the hyperbolic distance metric to predict interactions and associations between drugs, genes, diseases, and other objects in a complex biological network. To assess the practical benefits of our technique, we carried out a head-to-head comparison of the Euclidean and hyperbolic matrix factorization on the task of predicting drug-target interactions. The results of our benchmarking experiments demonstrate a substantial increase in the accuracy of hyperbolic representation compared to Euclidean representation. In addition, we show that the latent dimension of an optimal hyperbolic embedding is, by an order of magnitude, smaller than the latent dimension of an optimal Euclidean embedding. We see this as direct evidence that hyperbolic geometry (rather than Euclidean) underlines the biological system.