Advances in Plug-in and Non-Plug-in Nearest Neighbor Methods for Utilizing Density Functionals

Yung-Kyun Noh

Hanyang University / Korea Institute for Advanced Study

The nearest neighbor distances in data space can be used for estimating probability densities. P'lugging in a well-estimated probabilities can, in principle, generate various useful machine learning methods, such as Bayes classification and feature selection using information-theoretic measures. However, regardless of how well the method performs in practical, real-world applications, an underlying issue with this straightforward approach is the misconception that the plug-in methodology is universally applicable with asymptotic guarantees. In this tutorial, I will provide a foundational understanding for constructing non-plug-in methods, beginning with bias-variance behavior in high-dimensional space, clarifying which aspect should be prioritized for reduction, and designing algorithms that use nearest neighbor information when plug-in methods are invalid for certain problems, such as f-divergence estimation. Historically, our estimators are connected to the analysis of nearest neighbors established by T. Cover in the 1960s, who have demonstrated potential in both theoretical and practical realms of machine learning, particularly in addressing the asymptotic behaviors of classification errors and density estimations.