Machine Learning on Geospatial Big Data

Terence van Zyl

Abstract

When trying to understand the difference between machine learning and statistics, it is important to note that it is not so much the set of techniques and theory that are used but more importantly the intended use of the results. In fact, many of the underpinnings of machine learning are statistical in nature. When considering statistics, the main intent of statistics is in gaining an understanding of the underlying system, in this case geospatial system, through an analysis of observations or data about the system. Here, the geostatistician or environmental modeller is interested in cause and effect in the underlying system and gaining a deeper understanding of system itself. As a result of the need for environmental modellers and geostatisticians to gain an understanding of the underlying system, it is important that the eventual statistical model be interpretable, that is, not a black box. In fact, one reason for the limited use of machine learning algorithms has historically been exactly the lack of interpretability. Machine learning, on the other hand, is more focused on learning from observations of a system so as to be able to automate functionality. Here, the intention is not one of understanding but more one of engineering. For instance, in machine learning, a model may be trained so as to do automated classification of new unlabelled observations, to forecast future observations of some system or automatically spot anomalous events (Vatsavai et al. 2012). Geospatial big data present two opportunities for the increased use of machine learning in the geospatial analytics domain. First, geospatial big data have created a shift towards considering large amounts of data as a resource that can be used to add value to an organization. Second, by virtue of the three V's, volume, velocity, and variety, of big data, there is a shift away from complex models that require extensive computational and memory resources to techniques that instead can produce results in a more computationally efficient manner. Both of these opportunities provide a space in which black box solutions that produce *usable* results are more valuable than a strict need for interpretability and transparency.