Statistics and Data Science Seminar

Random Forest Prediction Intervals

Dan Nettleton (Iowa State University)

Abstract: Breiman's seminal paper on random forests has more than 30,000 citations according to Google Scholar. The impact of Breiman's random forests on machine learning, data analysis, data science, and science in general is difficult to measure but unquestionably substantial. The virtues of random forest methodology include no need to specify functional forms relating predictors to a response variable, capable performance for low-sample-size high-dimensional data, general prediction accuracy, easy parallelization, few tuning parameters, and applicability to a wide range of prediction problems with categorical or continuous responses. Like many algorithmic approaches to prediction, random forests are typically used to produce point predictions that are not accompanied by information about how far those predictions may be from true response values. From the statistical point of view, this is unacceptable; a key characteristic that distinguishes statistically rigorous approaches to prediction from others is the ability to provide quantifiably accurate assessments of prediction error from the same data used to generate point predictions. Thus, we develop a prediction interval – based on a random forest prediction – that gives a range of values that will contain an unknown continuous univariate response with any specified level of confidence. We illustrate our proposed approach to interval construction with examples and demonstrate its effectiveness relative to other approaches for interval construction using random forests.

Wednesday, September 25 at 4:00 PM in 636 SEO