STATISTICS SEMINAR

Daniel S. Nettleton, Ph.D.
Iowa State University

Thursday, November 12, 2015
Dickens Hall, Room 207, 4:00-5:00 pm
Refreshments: Dickens 108, 3:30-4:00 pm

Case-Specific Random Forests for Big Data Prediction

A Random Forest (RF) is a collection of classification or regression trees, where each tree is constructed from a bootstrap sample of a training dataset. A Case-Specific Random Forest (CSRF) is a variation of a RF in which uniform bootstrap resampling probabilities are replaced with proximity-weighted bootstrap resampling probabilities that are largest for training cases in closest proximity to a target case for which a response prediction is desired. When a training set is large, the construction of each tree in a forest may be computationally expensive. In such cases, it may be advantageous to build each tree from a bootstrap sample of size substantially less than the size of the training dataset. This makes particularly good sense when only a small fraction of the training dataset is relevant for predicting the response of a specific target case. We discuss the challenge of developing proximity-weighted bootstrap resampling probabilities that are concentrated on cases in a large training dataset that are relevant for predicting the response of a specific target case. Such probabilities can be used to generate computationally efficient CSRF predictions in big data problems.