BIOSTATISTICS SEMINAR

Cynthia D. Rudin

PhD, Professor of Computer Science, Associate Professor of Electrical and Computer Engineering, of Mathematics and of Statistical Science, Duke University

Interpretable Machine Learning: Optimal Decision Trees and Optimal Scoring Systems

How do patients and doctors know that they can trust predictions from a model that they cannot understand? Transparency in machine learning models is critical in high stakes decisions, like those made every day in healthcare. My lab creates machine learning algorithms for predictive models that are interpretable to human experts. I will focus on two historical hard optimization problems whose solutions are important in practice: **(1)** Optimal sparse decision trees and optimal sparse rule list models. Our algorithms are highly customized branch and bound procedures. These are an alternative to CART and other greedy decision tree methods. The solutions are globally optimal according to accuracy, regularized by the number of leaves (sparsity). This problem is NP-hard with no polynomial time approximation. I will present the first practical algorithms for this problem. **(2)** Optimal scoring systems. Scoring systems are sparse linear models with integer coefficients. Traditionally, scoring systems have been designed using manual feature elimination on logistic regression models, with a post-processing step where coefficients have been rounded. However, this process does not produce optimal solutions. I will present a novel cutting plane method for producing scoring systems from data. The solutions are globally optimal according to the logistic loss, regularized by the number of terms (sparsity), with coefficients constrained to be integers.

These algorithms have been used for many medical applications and criminal justice applications.

Thursday August 29, 2019   **3:30 pm - 4:30 pm**
Blue Cross and Blue Shield of North Carolina Foundation Auditorium

UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH