

**The DIVISION OF BIOSTATISTICS
of the
MEDICAL COLLEGE OF WISCONSIN**

Proudly Presents
A Special Talk
By:



Yuanjia Wang, PhD

Associate Professor

Department of Biostatistics, Mailman School of Public Health
Columbia University

Statistical Learning Methods for Complex Data: Targeted Local Classification and Risk Stratification for Censored Outcomes

We discuss two examples of statistical learning methods to build effective models for predicting disease diagnosis and progression. In the first example, we develop methods to predict whether a pre-symptomatic individual is at risk of a disease based on their biomarker profiles, which offers an opportunity for early intervention well in advance of receiving a definitive clinical diagnosis. To identify effective prediction rules using nonparametric decision functions, standard statistical learning approaches treat markers with clear biological importance (e.g., genomic disease burden) and other markers without prior knowledge on disease etiology interchangeably as input variables. Therefore, these approaches may be inadequate in singling out and preserving the effects from the biologically important variables, especially in the presence of high-dimensional markers. Using a disease burden score derived from genetic structural variation as an example of a salient marker to receive special care in the analysis, we propose a local smoothing large margin classifier to construct effective varying-coefficient classification rules. In the second example, we discuss how to learn risk scores for time-to-event outcomes subject to right censoring. Our method is based on predicting the counting process associated with the time-to-event outcomes among subjects at risk via a series of support vector machines. Our method leads to a connection between support vector machines in supervised learning and hazards regression in standard survival analysis. We show that the method is optimal in discriminating covariate-specific hazard function from population average hazard function, and establish the consistency and learning rate of the predicted risk using the estimated risk scores. Performance of both developed methods are demonstrated through simulations and real data applications to biomarker studies of Huntington's disease.

Tuesday, November 1, 2016

3:30 – 4:30 PM

Medical College of Wisconsin

Room M2050 – 2nd floor of the MEB

Refreshments 3:00 – 3:30 PM in H2030