



STATISTICS DEPARTMENT

SEMINAR

TITLE: **NONPARAMETRIC VARIABLE SELECTION: THE EARTH ALGORITHM - WITH APPLICATIONS TO REGRESSION AND CLASSIFICATION**

SPEAKER: Shijie Tang

TIME: 4:00 P.M.

DATE: Wednesday, February 21, 2007

ROOM: 140 BARDEEN

ABSTRACT:

We consider regression experiments involving a response variable and a large number of predictor variables, many of which may be irrelevant for the prediction of the response and thus need to be removed before predicting the response from the predictors. Similarly, the variables that are related to the response need to be selected and their relationship to the response analyzed. This paper uses local polynomial methods with bandwidths chosen to provide a high probability of selecting the relevant variables. Our approach avoids the curse of dimensionality by basing bandwidth selection on a local signal to noise ratio, called efficacy, which automatically and adaptively selects relatively large local neighborhoods. We develop an algorithm called EARTH (Efficacy Adaptive Regression Tube Hunting) based on the conditional expectation of the response given all but one of the predictor variables, and we derive some of its properties. Computer simulations show that EARTH successfully and efficiently selects the relevant variables in situations with a large number of irrelevant predictor variables for a variety of models. When it is combined with the model selection and prediction procedure MARS or the tree-based prediction procedure GUIDE, the combinations lead to improved prediction accuracy. We also discuss the case where the response variable is categorical and the case where the input variables may be categorical. We show how our procedure can improve classification algorithms such as the widely used Support Vector Machine and the tree-based algorithm QUEST.

Coffee and Cookies at 3:30 p.m. in Room 1210 MSC