



Spring 2010 CIS Colloquium Series

Feature Selection in Large Data Sets

Lyle Ungar
(University of Pennsylvania)

11am-12pm, Wednesday, January 20
4th Floor Conference Room (Wachman Hall, CC 447)

Abstract: When building predictive models using large data sets such as arise in biology and in text mining, one often needs to select a small number of the millions of candidate features. This talk describes our Streamwise Feature Selection (SFS) method, in which new features are sequentially considered for addition to a predictive model. SFS dynamically adjusts an information theoretic-based criterion for adding new features, giving it much more power than fixed criteria such as AIC, BIC and RIC, while still giving strong guarantees against overfitting. When the space of potential features is large, SFS offers many advantages over methods in which all features are assumed to be known in advance; Features can be generated dynamically, focusing the search for new features on promising subspaces. Empirical results show that SFS is competitive with much more compute-intensive feature selection methods. It is also easily modified to exploit weak domain knowledge about the features.

Joint work with Paramveer Dhillon, Dean Foster, and Jing Zhou.

Bio: Dr. Lyle H. Ungar is an Associate Professor of Computer and Information Science (CIS) at the University of Pennsylvania. He also holds appointments in several other departments at Penn in the Engineering, Wharton, and Medicine Schools. Dr. Ungar received a B.S. from Stanford University and a Ph.D. from M.I.T. He directed Penn's Executive Masters of Technology Management (EMTM) Program for a decade, and is currently Associate Director of the Penn Center for BioInformatics (PCBI). He has published over 100 articles and holds eight patents. His current research focuses on developing scalable machine learning methods for data mining and text mining.

Refreshments will be served!