Individuals and Organizations Completing Research in the IMEDS Lab
Massachusetts Institute of Technology

Scope of Research - Proposal to request access to IMEDS Laboratory from Cynthia Rudin, Massachusetts Institute of Technology.

Research Objectives and Aims:
We are developing new statistical tools to learn simple scoring systems from data. In many areas of medicine scoring systems are the predictive model of choice for assessing the risk of an adverse event. For instance, the $CHADS_2$ score uses five features and integer coefficients to estimate the risk of stroke in atrial fibrillation patients. These scoring systems are preferred by physicians because of their interpretability – it is clear which features are being used to make a prediction. However, scoring systems are generally constructed by hand, and so their predictive accuracy suffers in comparison to standard black-box machine learning tools like support vector machines. Our aim is to develop statistical learning methods to learn scoring systems from data, without any manual construction. Such a model would give the interpretability of a scoring system with the predictive power that comes with statistical learning. Applying the method to the massive observational data at IMEDS will be useful for determining a simple, sparse set of correlative factors that underlie an adverse event, and will provide a suite of accurate scoring systems that will be used by physicians to make better treatment decisions.

Scope/Proposed Approach:
We are developing two novel approaches to train scoring systems from data. Both approaches aim to train scoring systems that can balance accuracy and interpretability in a scalable manner. In addition, these approaches also aim allow practitioners to use their own measures of interpretability.

The first is a Bayesian approach that uses Markov Chain Monte Carlo (MCMC) techniques. Here, the likelihood function is the logistic regression likelihood, and will ensure solutions have high classification accuracy. The prior will be used to favor solutions that are interpretable (e.g., sparse, integer coefficients). Posterior sampling via MCMC will then yield a scoring system that can balance accuracy and interpretability. The core methodological contribution of this approach consists of developing the prior and sampling procedure are the core of the methodological work.

The second is an optimization-based approach that is based on a class of algorithms known as decomposition or localization techniques. This approach decouples a difficult and non-convex optimization problem into two smaller parts and solves these problems in an iterative manner. The resulting approach can yield the exact same scoring system as the Bayesian approach, but can also provide a guarantee of optimality, thereby allowing end-users to know when the technique has produced a suitable scoring system.

Impact:
Twenty years after the invention of support vector machines, many prediction models in medicine are still developed through the use of simple heuristics. The opacity of high-performance machine learning tools is hurting their adoption. The methods that we are developing will provide physicians with predictive models that not only mimic models that are currently being used in practice, but are likely to be far more accurate and easier to produce. We expect this work to have a high impact in spreading statistical learning-based predictive models

to daily use in medicine. Both the scoring systems that we develop from the IMEDS data, and the method itself will be useful to the public.

Experience:

Ben Letham, Berk Ustun, and Ramin Moghaddass are the individuals involved in this study, and are requesting access to the IMEDS lab. Ben is a fourth-year PhD student in Operations Research at MIT. He has published research papers in several of the top journals in machine learning, including *Machine Learning*, *Journal of Machine Learning Research*, *ICML*, and *Data Mining and Knowledge Discovery*. Berk Ustun is a second-year PhD student in Electrical Engineering and Computer Science at MIT. Ramin Moghaddass is a post doc at MIT Sloan School of Management. Cynthia Rudin will be involved in the project but does not require direct access to the IMEDS lab. She is an Associate Professor in Statistics at the MIT Sloan School of Management.

Timeline:

We expect the whole research project (methodological developments, simulations, benchmark dataset studies, IMED lab experiments, and finally publication) to take less than 12 months, and are thus requesting access until 31 December 2014, with access beginning as soon as possible.