AI for Equitable, Data-Driven Decisions in Public Health

Bryan Wilder[1]

[1]Machine Learning Department, Carnegie Mellon University
bwilder@andrew.cmu.edu

Abstract

As exemplified by the COVID-19 pandemic, our health and wellbeing depend on a difficult-to-measure web of societal factors and individual behaviors. This effort requires new algorithmic and data-driven paradigms which span the full process of gathering costly data, learning models to understand and predict such interactions, and optimizing the use of limited resources in interventions. In response to these needs, I present methodological developments at the intersection of machine learning, optimization, and social networks which are motivated by on-the-ground collaborations on HIV prevention, tuberculosis treatment, and the COVID-19 response. Here, I give an overview of two lines of work.

Social network interventions for HIV prevention

Youth experiencing homelessness (YEH) have approximately 10 times higher HIV prevalence than the general population. One mechanism for behavior change is to recruit peer leaders from amongst the YEH to serve as advocates for HIV prevention. This poses the question: which youth would be most effective at disseminating prevention messages? The influence maximization problem, subject of extensive work in theoretical computer science and AI, asks how a limited number of seed nodes can be selected from a network to maximize information diffusion. However, no prior work had connected influence maximization to network interventions in health. The key computational challenge is severely limited data – previous work assumed that the network structure and model of information diffusion are known but neither are available in health domains. We developed algorithms for robust optimization and graph sampling to address this challenge (Wilder et al. 2018; Wilder 2018; Staib, Wilder, and Jegelka 2019). For example, we developed an algorithm to *subsample* the network by querying only specific nodes for their edges . On graphs drawn from the stochastic block model, this algorithm obtains a constant-factor approximation to the optimal influence spread while querying only $O(\log n)$ nodes.

With social work colleagues, we conducted a clinical trial which enrolled 713 youth (Wilder et al. 2021). The trial compared three arms: one where the intervention was planned using the system I developed, one where the highest-degree nodes were selected (the most common method in public health), and an observation-only control group. The results from this trial showed clear evidence in favor of the algorithmic approach: youth in that arm had a statistically significant reduction of 33% in their odds of engaging in condomless anal sex, a key risk behavior for HIV. No statistically significant effect was observed for the high-degree arm. This study provides, to our knowledge, the first empirical evaluation of an algorithmic approach to social network interventions in health.

Integrating machine learning and optimization

Consider a decision-maker allocating limited resources, formalized as an optimization problem. In a health context, this could be selecting a limited number of patients for additional followup. However, input parameters to the optimization problem, for example the true risk level of each patient, are not known exactly. Rather, they are predicted by a machine learning model. How should we train the model? The standard approach would maximize predictive accuracy as measured by a loss function. However, the model does not need to be equally accurate everywhere; all that matters is whether it can identify a high-quality decision. We present a new approach which embeds the optimization problem into training for the machine learning model, focusing the model directly on inducing good decisions (Wilder, Dilkina, and Tambe 2019; Ferber et al. 2020; Wilder et al. 2019; Wang et al. 2019). Previous work dealt with continuous (convex) optimization problems where the solution can be differentiated with respect to the input predictions. By contrast, many application domains require discrete decisions (e.g., we either intervene with a given patient or not), which inherently breaks differentiability. Our approach designs continuous relaxations of the discrete problems for use as surrogates in the training loop.

These techniques have shown promise in addressing a pressing challenge in global health: tuberculosis (TB) treatment. Using real data supplied by the city of Mumbai, we developed a system which predicts the probability of a patient failing to take their medication and optimizes the scheduling of house visits by health workers to keep patients in treatment (Killian et al. 2019). Results on the historical data showed a 15% improvement in intervention effectiveness compared to a model trained with standard methods.