DISSERTATION PROPOSAL

Xinying (Violet) Chen

"Optimization Methods for Attaining and Understanding Fairness in AI"

Monday, December 7, 2020
1:00pm EST
Zoom: https://cmu.zoom.us/s/99777999554

With increasing deployment of Artificial Intelligence (AI) methods to assist high-stake real life decisions, fairness has become an essential factor of consideration for both the designers and users of AI tools. Unified by the central theme of fairness in AI, this dissertation is inspired by two fundamental questions in the design of AI methods: what is fair and how to be fair. We utilize optimization as the primary theoretical foundation and aim to contribute insights to these high level questions from three directions.

In chapter one, we study fairness in mathematical modeling. While optimization models generally focus on efficiency goals such as cost minimization or benefit maximization, there is a wide range of applications where fairness is also critical. Moreover, many decisions require a reasonable trade-off between fairness and efficiency. We propose a principled and practical method for balancing these two criteria in an optimization model. We define a set of social welfare functions (SWFs) that combine leximax fairness and utilitarianism and overcome some of the weaknesses of existing trade-off schemes. In particular, our SWFs regulate the equity/efficiency trade-off with a single parameter that has a meaningful interpretation in practical contexts. We formulate the SWFs using mixed integer constraints and provide a sequential procedure to maximize these SWFs subject to problem specific constraints. We demonstrate the practical potentials of our method on problems of realistic size involving healthcare resource allocation and disaster preparation.

In chapter two and three, we explore optimization techniques for fair classification in machine learning. Chapter two studies a new algorithm for existing fairness constrained support vector machine (SVM) models: we design a sequential minimal optimization (SMO) type algorithm to solve SVM formulations that contain a set of convex and non-repetitive fairness constraints. Our fair SMO algorithm can be considered as the standard SMO algorithm, a popular algorithm for training SVM, complemented by fairness-seeking steps. We prove the asymptotic convergence and finite termination of fair SMO, and demonstrate with experiments on synthetic and real datasets that fair SMO incurs comparable computational costs relative to standard SMO.

In Chapter three, we propose to explore new formulation and computation techniques for fair classification. The majority of fairness definitions proposed for ML equate fairness with elimination of certain disparity among groups or individuals, and the chosen fairness conditions are often encoded with their convex, continuous variations to generate tractable fair classification models. We note two insufficient aspects of these methods: one is the lack of theoretical guarantees in terms of the exact fairness measures based on discrete classification outcomes; the other one is that parity based fairness notions may be misaligned with utility based social justice standards. We aim to study strategies to address both issues. Specifically, we wish to design computational techniques to handle integer variables in fair classification models, and utilize these techniques to develop efficient, scalable and justice-driven fair classification methods.

In Chapter four and five, we shift our focus from making fair decisions to learning preference information relevant for fairness. Chapter four studies online learning (OL) from revealed preferences: a learner wishes to learn an agent's private utility function through interacting with the agent in a changing environment. Through designing a new loss function that is convex under relatively mild assumptions, we design a flexible OL framework that enables a unified treatment of usual loss functions from literature and supports a variety of online convex optimization algorithms. We demonstrate with theoretical and empirical evidence that our framework based on the new loss function (in particular online Mirror Descent) has significant advantages in terms of regret performance and solution time over other OL algorithms from the literature. This general framework is broadly applicable to decision making tasks that are affected by inferred preferences; for example, one potential fair AI related application is that a central agency seeks to fairly allocate limited resources to agents with diverse preferences.

Lastly, in chapter five, we propose a new project to investigate the learning of dynamic fairness preferences. A recent line of work considered inferring people's fairness preferences from querying their fair decisions under different contexts; one example of a query is to ask a person to evaluate whether two individuals or groups should receive similar or disparate treatment. The inferred fairness or higher level moral preference knowledge is useful for designing AI systems that align with stakeholders' moral principles. This emerging thread has worked with a static set of preferences, but as commonly observed in reality, fairness preferences can be influenced and changed by internal and external factors. We wish to formulate possible preference dynamics and preference elicitation queries as mathematical definitions, then study how different query designs affect the learning of different preferences.