Arizona State University Machine Learning Day 2021

Some Unique Problems with Social Media Data for Machine Learning and AI, Huan Liu

Abstract: Social media data is distinctive from its traditional counterpart and opens the door for interdisciplinary research and allows researchers to collectively study large-scale human behavior otherwise impossible. The study of social media data brings about new challenges for machine learning and data mining. In this talk, we will introduce some unique issues of big social media data, e.g., the big data paradox, the privacy-utility 'trade-off', and the evaluation dilemma. We will also mention some efforts of using AI for good if time allows. With data abundance and algorithmic development, we are better equipped than ever to answer challenging and novel research questions and advance AI and CS.

Designing surveillance: The critical role of ethical frameworks in AI Games research, Florence Chee

Abstract: This talk discusses the application of ethical frameworks to research on game communities, data-driven game design, and their role in defining the present and future of everyday sociotechnical practices. The issues of consent, privacy, race, and gender that arise out of this research present numerous implications for policy and are an increasing part of much needed interdisciplinary discussions and collaborations.

Excavating the "unintended consequences" of algorithms: The case of Facebook's "interest" classification system for ad targeting, Kelley Cotter

Abstract: This talk presents a case study that exemplifies a critical analysis of an algorithmic classification system. I offer some examples of the subtle ways that such systems can reproduce social biases and inequities and demonstrate one approach to surfacing these.

Anna Jobin (Title/Abstract TBD)

Efficiently Evaluate Social Network Interventions, Johan Ugander

Abstract: When trying to maximize the adoption of a behavior in a community such as a school or rural village, how should one target individuals? This talk will explore the frontiers of this research question, discussing strategies that leverage the structure of social networks to target individuals expected to have an outsized impact on adoption, as well as how to evaluate the efficacy of such ideas using statistical ideas from importance sampling.

A super scalable algorithm for short segment detection, Selena Niu

Abstract: In many applications such as copy number variant (CNV) detection, the goal is to identify short segments on which the observations have different means or medians from the background. Those segments are usually short and hidden in a long sequence, and hence are very challenging to find. In this talk, we introduce a super scalable short segment (4S) detection algorithm. This nonparametric method clusters the locations where the observations exceed a threshold for segment detection. It is computationally efficient and does not rely on Gaussian noise assumption. Moreover, we develop a framework to assign significance levels for detected segments. We demonstrate the advantages of our proposed method by theoretical, simulation, and real data studies.

Designing disease-tracking metrics with causal machine learning, P. Richard Hahn

Abstract: By combining ideas from causal inference with machine learning we are able to design diagnostic metrics that track neurological disease status better than popular (purely associative) supervised machine learning approaches which have been unsuccessful in this application. Our new approach defines the neurological state as the unobserved physical changes that mediate the causal effect of primary biomarkers (as measured by PET scans) on various clinical measures of cognition; this conceptualization allows us to "causally denoise" the clinical metrics, resulting in a more reliable gauge of disease progression. We relate our new approach to a traditional approach of norming measured patient symptoms with respect to the distribution of that symptom in a healthy population. A small simulated example illustrates the idea and its potential.

Resilience and adaptation in social networks, Will Hobbs

Abstract: This talk will describe how online social networks change in response to deaths, sudden censorship, and inflammatory political rhetoric. Technical aspects of the talk will discuss causal inference, applications of automated text analysis in social science, and new behavioral measurements made possible through large-scale digital trace data.

Measuring physical and mental health using social media, Johannes Eichstaedt

Abstract: The content shared on social media the largest data set on human thoughts, emotions, and behaviors in history. We use Natural Language Processing and machine learning to leverage this data for the social good and psychological science. I will demonstrate how Facebook data can be used to predict depression of patients before it appears in their medical records, and how Twitter can predict the heart disease of communities better than standard risk factors. Across these studies, I argue that AI-based approaches to social media can augment clinical practice, guide prevention, and inform public policy.

Erika Salomon, University of Chicago's Center for Data Science and Public Policy (Title/Abstract TBD)

Efficient active learning of halfspaces: noise tolerance and exploiting sparsity, Chicheng Zhang

Abstract: Motivated by the abundance of unlabeled data and the expensiveness of obtaining label annotations, the paradigm of active learning has been proposed and studied in the literature. By adaptively selecting examples for label queries, an active learner learns accurate classification models with a substantially reduced label complexity than conventional supervised learning. In this talk, we focus on active learning of d-dimensional homogeneous halfspaces (linear classifiers), a fundamental concept class in machine learning. Specifically: 1) We give an efficient algorithm for active learning halfspaces with tolerance to Massart and Tsybakov noise. In the special case of Massart noise, our algorithm is the first to achieve both computational efficiency and information-theoretic optimal label complexity, under a broad family of unlabeled distributions. 2) We propose an efficient algorithm that exploits the sparsity of the target halfspace. Specifically, if the target halfspace is s-sparse (s << d), our algorithm achieves a label complexity guarantee of order O(s polylog(d,1/eps)), which significantly improves over the O(d polylog(1/eps)) label complexities by full-dimensional active learning algorithms. This is based on joint works with Jie Shen (Stevens Institute of Technology), Pranjal Awasthi (Rutgers), and Yinan Li (University of Arizona).

Why does functional pruning yield such fast algorithms for optimal changepoint detection? Toby D Hocking

Abstract: In this talk I will present a review of recently proposed algorithms for optimal changepoint detection, which are empirically very fast, but we don't have any good theoretical justification as to why this is the case in realistic data settings. Detecting abrupt changes is an important problem in N data gathered over time or space. In this setting, maximum likelihood inference amounts to minimizing a loss function (which encourages data fitting) plus a penalty on the number of changes in the model parameters (which discourages overfitting). Computing the optimal solution to this non-convex problem is possible using classical dynamic programming algorithms, but their O(N^2) complexity is too slow for large data sequences. The functional pruning technique of Rigaill involves storing the optimal cost using functions rather than scalar values. Empirical results from several recent papers show that the functional pruning technique consistently yields optimal algorithms of O(N log N) complexity, which is computationally tractable for very large N. The theoretical results of Rigaill prove that functional pruning is O(N^2) in the worst case and O(N log N) on average (for a special loss function). For future work it would be interesting to further study the average complexity of these algorithms, in order to provide more theoretical justification for these very fast empirical results.

Optimizing for whom? The role of robustness in equitable algorithms, Kenneth Nieser

Abstract: Many of the objective functions that we use to fit models to data rely on some form of an average across the data sample. This might result in nice statistical properties, but for whom? Models might perform well for one subgroup of a population but poorly for another. This can have meaningful ramifications when the model is deployed in some consequential setting, like informing a medical diagnosis or treatment decision. Algorithmic fairness challenges us to develop models and algorithms that are both accurate and “fair”. In this talk, I will discuss what role I think robustness can play in developing

fair models and algorithms. I will present a robust expectation-maximization algorithm that can be used in latent variable settings to identify subgroups for whom the model is performing poorly. This method could be incorporated into model building pipelines to assess model fairness.