2014

Can We Learn to Be Fair?

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, Rich Zemel

"Fairness Through Awareness" is an approach to fairness in classification, where the goal is to prevent discrimination against protected population subgroups in classification systems while simultaneously preserving utility for the party carrying out the classification (eg, an advertiser, bank, or admissions committee). We argue that a classification is fair only when individuals who are similar with respect to the classification task at hand are treated similarly, and this in turn requires understanding of sub-cultures of the population. In consequence, hiding information from a classifying algorithm can result in less fairness (and less utility): "privacy" does not yield fairness.

We obtain a computational solution that, given a similarity metric defining, for each pair of individuals, their similarity with respect to the given classification task, achieves our fairness goals. The metric should represent ground truth, but how can it be obtained? Can learning help?

We also discuss the crescendo of calls for "comprehensible" or "interpretable" classifiers that "explain" individual classifications, and suggest a new desideratum, which we call "negotiability," as a direction for future research.

Learning Rich But Fair Representations

Richard Zemel

We propose a learning algorithm for fair classification that achieves both group fairness (the proportion of members in a protected group receiving positive classification is identical to the proportion in the population as a whole), and individual fairness (similar individuals should be treated similarly). We formulate fairness as an optimization problem of finding a good representation of the data with two competing goals: to encode the data as well as possible, while simultaneously obfuscating any information about membership in the protected group. I will present some alternative formulations for the two main components of this framework, the measure of fairness, and the form of the learned representation. I will show empirical comparisons of these with earlier formulations on two datasets.

Slides

Moving Beyond Prediction: Big Data, Transparency, and Accountability

Hanna Wallach

This talk will be structured around four talking points – intended to prompt discussion – that lie at the heart of fairness and transparency in machine learning: data, questions, models, and findings. In discussing these talking points, I will touch upon limitations of the most prevalent definitions of big data; the need for true collaborations with social scientists when analyzing social data; data-driven research vs. question-driven research; convenience data vs. carefully selected data; transparency and algorithmic accountability reporting; models for exploration/explanation vs. models for prediction; representing and maintaining uncertainty; error analysis; intuition and bias in interpreting findings; and, finally, the importance of scientific communication.

Slides

Privacy through Accountability: Information Flow Experiments

Anupam Datta, Michael Carl Tschantz, and Amit Datta

Privacy through accountability refers to the principle that entities that hold personal information about individuals are accountable for adopting measures that protect the privacy of the data subjects. Computational approaches to privacy through accountability involve developing algorithms and tools that can be used to provide internal and external oversight about the practices of such entities. After providing an overview of this emerging research area, I will focus on one of our recent results in Web privacy.

I will describe the problem of detecting personal data usage by websites when the analyst does not have access to the code of the system nor full control over the inputs or observability of all outputs of the system. A concrete example of this setting is one in which a privacy advocacy group, a government regulator, or a Web user may be interested in checking whether a particular web site uses certain types of personal information for advertising. I will present a methodology for information flow experiments based on experimental science and statistical analysis that addresses this problem, our tool AdFisher that incorporates this methodology, and findings of opacity, choice and discrimination from our experiments with Google. These results also raise interesting challenges for the design of new classes of machine learning algorithms that provide transparency, respect choice, and are non-discriminatory.

Slides

Certifying and Removing Disparate Impact

Sorelle Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian

What does it mean for an algorithm to be biased?

In U.S. law, the notion of bias is typically encoded through the idea of disparate impact: namely, that a process (hiring, selection, etc) that on the surface seems completely neutral might still have widely different impacts on different groups. This legal determination expects an explicit understanding of the selection process.

If the process is an algorithm though (as is common these days), the process of determining disparate impact (and hence bias) becomes trickier. Firstly, it might not be possible to disclose the process. Secondly, even if the process is open, it might be too complex to ascertain how the algorithm is making its decisions. In effect, since we don't have access to the algorithm, we must make inferences based on the data it uses.

We make three contributions to this problem. Firstly, we link the legal notion of disparate impact to a measure of classification accuracy that while known, has not received as much attention as more traditional notions of accuracy. Secondly, we propose a test for the possibility of bias based on analyzing the information leakage of protected information from the data. Finally, we describe methods by which data might be made "unbiased" in order to test an algorithm. Interestingly, our approach bears some resemblance to actual practices that have recently received legal scrutiny.

Slides

Accountable Algorithms

Joshua Kroll and Ed Felten

Important decisions about people are increasingly made by algorithms: Votes are counted; voter rolls are purged; financial aid decisions are made; taxpayers are chosen for audits; air travelers are selected for search; credit eligibility decisions are made. Citizens, and society as a whole, have an interest in making these processes more transparent. Yet the full basis for these decisions is rarely available to affected people: the algorithm or some inputs may be secret; or the implementation may be secret; or the process may not be precisely described. A person who suspects the process went wrong has little recourse. And an oversight authority who wants to ensure that decisions are made according to an acceptable policy has little assurance that proffered decision rules match decisions for actual users.

To address this problem, we propose to use accountable algorithms, which provide both an result and a proof that can convince a skeptical party that a con-

sistent policy was applied correctly to accurate data to produce the announced result. Critically, the proof can convince an observer while maintaining the secrecy of parts of the policy used to determine the output, and the privacy of individuals' personal data.

Our methods use the tools of computer science to cryptographycially ensure the technical properties that can be proven, while providing the necessary information so that a political, legal, or social oversight process can operate effectively. Combining the technology of verified computation with the operation of non-technical governance structures offers the best hope of governing the operation of algorithmic decision processes in practice.

Slides

Civil Rights and Machine Learning: Emerging Policy Questions

David Robinson and Harlan Yu

The key decisions that shape people's lives – decisions about jobs, healthcare, housing, education, criminal justice and other key areas – are, more and more often, being made by machine learning systems. As a result, a growing number of important conversations about civil rights, which focus on how these decisions are made, are also becoming discussions about machine learning. Policymakers and the public increasingly want to understand how machine learning systems work in general, how they reach particular decisions, and (in some cases) how their operation might be altered to advance social goals.

Some political requirements for automated decisions – such as a requirement that decisions be reached using a consistent procedure, or be intelligible to human observers – may lend themselves to technical solutions. But in the area of non-discrimination, U.S. law often avoids bright line rules, proceeding instead on a holistic, case-by-case basis. The kind of rules that engineers may need, in other words, may not exist today. Those rules may need to be developed anew, in a policy discussion that is informed by a deeper understanding of technical methods. This is both a challenge and an opportunity for the field of machine learning to help shape the future of civil rights.

2015

Discrimination- and Privacy-Aware Data Mining

Sara Hajian

In the information society, massive and automated data collection occurs as a consequence of the ubiquitous digital traces we all generate in our daily life. The availability of such wealth of data makes its publication and analysis highly desirable for a variety of purposes, including policy making, planning, marketing, research, etc. Yet, the real and obvious benefits of data analysis and publishing have a dual, darker side. There are at least two potential threats for individuals whose information is published: privacy invasion and potential discrimination. Privacy invasion occurs when the values of published sensitive attributes can be linked to specific individuals (or companies). Discrimination is unfair or unequal treatment of people based on membership to a category, group or minority, without regard to individual characteristics.

On the legal side, parallel to the development of privacy legislation, anti-discrimination legislation has undergone a remarkable expansion, and it now prohibits discrimination against protected groups on the grounds of race, color, religion, nationality, sex, marital status, age and pregnancy, and in a number of settings, like credit and insurance, personnel selection and wages, and access to public services. On the technology side, efforts at guaranteeing privacy have led to developing privacy preserving data mining (PPDM) and efforts at fighting discrimination have led to developing anti-discrimination techniques in data mining. Some proposals are oriented to the discovery and measurement of discrimination, while others deal with preventing data mining (DPDM) from becoming itself a source of discrimination, due to automated decision making based on discriminatory models extracted from inherently biased datasets. I will describe some of these techniques for discrimination prevention, simultaneous discrimination and privacy protection, and discrimination discovery and show some recent results.

Slides

Privacy Attacks and Anonymization Methods as Tools for Discrimination Discovery and Fairness

Salvatore Ruggieri

Social discrimination discovery from data aims to identify illegal and unethical discriminatory patterns towards protected-by-law groups. Fairness in machine learning aims at preventing the usage of such patterns in classifiers trained from data possibly containing them. In this talk, we introduce an intriguing parallel between the role of the anti-discrimination authority and the role of an attacker in private data publishing. The parallel leads to two approaches in re-using tools from the privacy research. On the one side, we deploy privacy attack strategies, such as Frëchet bounds attacks, as tools for indirect discrimination discovery. On the other side, we investigate the relation between attribute inference control methods and social discrimination models, showing that t closeness implies bd(t)-protection for a bound function bd(). This al-

lows us to adapt data anonymization algorithms, such as Mondrian multidimensional generalization and Sabre bucketization and redistribution, to the purpose of non-discrimination data protection—a form of pre-processing that removes discriminatory patterns from training data.

Future Directions of Fairness-Aware Data Mining: Recommendation, Causality, and Theoretical Aspects

Toshihiro Kamishima, Kazuto Fukuchi, Jun Sakuma, Shotaro Akaho, and Hideki Asoh

The goal of fairness-aware data mining (FADM) is to analyze data while taking into account potential issues of fairness. In this talk, we will cover three topics in FADM:

Fairness in a Recommendation Context: In classification tasks, the term "fairness" is regarded as anti-discrimination. We will present other types of problems related to the fairness in a recommendation context.

What is Fairness: Most formal definitions of fairness have a connection with the notion of statistical independence. We will explore other types of formal fairness based on causality, agreement, and unfairness.

Theoretical Problems of FADM: After reviewing technical and theoretical open problems in the FADM literature, we will introduce the theory of the generalization bound in terms of accuracy as well as fairness.

Fairness Constraints: A Mechanism for Fair Classification

Muhammad Bilal Zafar, Isabel Valera Martinez, Manuel Gomez Rodriguez, and Krishna Gummadi

Automated data-driven decision systems are ubiquitous across a wide variety of online services, from online social networking and ecommerce to e-government. These systems rely on complex learning methods and vast amounts of data to optimize the service functionality, satisfaction of the end user and profitability. However, there is a growing concern that these automated decisions can lead to user discrimination, even in the absence of intent.

In this paper, we introduce fairness constraints, a mechanism to ensure fairness in a wide variety of classifiers in a principled manner. Fairness prevents a classifier from outputting predictions correlated with certain sensitive attributes in the data. We then instantiate fairness constraints on three well-known classifiers—logistic regression, hinge loss and support vector machines (SVM)—and evaluate their performance in a real-world dataset with meaningful sensitive human attributes. Experiments show that fairness constraints allow for an optimal trade-off between accuracy and fairness.

Fair Boosting: A Case Study

Benjamin Fish, Jeremy Kun, and Ádám D. Lelkes

We study the classical AdaBoost algorithm in the context of fairness. We use the Census Income Dataset as a case study. We empirically evaluate the bias and error of four variants of AdaBoost relative to an unmodified AdaBoost baseline, and study the trade-offs between reducing bias and maintaining low error. We further define a new notion of fairness and measure it for all of our methods. Our proposed method, modifying the hypothesis output by AdaBoost by shifting the decision boundary for the protected group, outperforms the state of the art for the census dataset.

Towards Diagnosing Accuracy Loss in Discrimination-Aware Classification: An Application to Predictive Policing

Zubin Jelveh and Michael Luca

Prediction algorithms are increasingly used to forecast outcomes of processes that are societally sensitive. In response, algorithms have been developed to produce fair classifications but at the potential cost of accuracy. In this work, we present a framework for modeling the pathways by which sensitive variables influence—and are influenced by—nonsensitive variables. These pathways allow us to discern between two types of accuracy loss: justified reduction due to underlying discrimination in the data, and overadjustment due to the removal of nonsensitive predictive information. We also present a framework for adjusting input data to remove the association between sensitive and nonsensitive predictors and assess its ability to produce fair classifications. Finally, we apply our methodology to a new dataset in the criminal justice domain.

On the Relation between Accuracy and Fairness in Binary Classification

Indrė Žliobaitė

Our study revisits the problem of accuracy/fairness tradeoff in binary classification. We argue that comparison of non-discriminatory classifiers needs to account for different rates of positive predictions, otherwise conclusions about performance may be misleading, because accuracy and discrimination of naive baselines on the same dataset vary with different rates of positive predictions. We provide methodological recommendations for sound comparison of non-discriminatory classifiers, and present a brief theoretical and empirical analysis of tradeoffs between accuracy and non-discrimination.

2016

Equality of Opportunity in Supervised Learning

Eric Price, Nati Srebro, Moritz Hardt

We propose a criterion for discrimination against a specified sensitive attribute in supervised learning, where the goal is to predict some target based on available features. Assuming data about the predictor, target, and membership in the protected group are available, we show how to optimally adjust any learned predictor so as to remove discrimination according to our definition. Our framework also improves incentives by shifting the cost of poor classification from disadvantaged groups to the decision maker, who can respond by improving the classification accuracy. In line with other studies, our notion is oblivious: it depends only on the joint statistics of the predictor, the target and the protected attribute, but not on interpretation of individual features. We study the inherent limits of defining and identifying biases based on such oblivious measures, outlining what can and cannot be inferred from different oblivious tests. We illustrate our notion using a case study of FICO credit scores.

Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of Opportunity in Supervised Learning." In Advances In Neural Information Processing Systems, pp. 3315-3323. 2016.

Fairness in Learning: Classic and Contextual Bandits

Matthew Joseph, Michael Kearns, Jamie Morgenstern, Aaron Roth

We introduce the study of fairness in multi-armed bandit problems. Our fairness definition can be interpreted as demanding that given a pool of applicants (say, for college admission or mortgages), a worse applicant is never favored over a better one, despite a learning algorithm's uncertainty over the true payoffs. We prove results of two types. First, in the important special case of the classic stochastic bandits problem (i.e., in which there are no contexts), we provide a provably fair algorithm based on "chained" confidence intervals, and provide a cumulative regret bound with a cubic dependence on the number of arms. We further show that any fair algorithm must have such a dependence. When combined with regret bounds for standard non-fair algorithms such as UCB, this proves a strong separation between fair and unfair learning, which extends to the general contextual case. In the general contextual case, we prove a tight connection between fairness and the KWIK (Knows What It Knows) learning model: a KWIK algorithm for a class of functions can be transformed into a provably fair contextual bandit algorithm, and conversely any fair contextual bandit algorithm can be transformed into a KWIK learning algorithm. This tight connection allows us to provide a provably fair algorithm for the linear contextual bandit problem with a polynomial dependence on the dimension,

and to show (for a different class of functions) a worst-case exponential gap in regret between fair and non-fair learning algorithms.

Joseph, Matthew, Michael Kearns, Jamie Morgenstern, and Aaron Roth. "Fairness in Learning: Classic and Contextual Bandits." In Advances In Neural Information Processing Systems, pp. 325-333. 2016.

To Predict and Serve?

Kristian Lum and William Isaac

Predictive policing systems are used increasingly by law enforcement to try to prevent crime before it occurs. But what happens when these systems are trained using biased data?

Lum, Kristian, and William Isaac. "To Predict and Serve?." Significance 13, no. 5 (2016): 14-19.

Combatting Police Discrimination in The Age of Big Data

Sharad Goel, Maya Perelman, Ravi Shroff, David Alan Sklansky

The growth of available information about routine police activities offers new opportunities to improve the fairness and effectiveness of police practices. We illustrate this point by showing how a particular kind of calculation made possible by modern, large-scale datasets—determining the likelihood that stopping and frisking a particular pedestrian will result in the discovery of contraband or other evidence of criminal activity—could be used to reduce the racially disparate impact of pedestrian searches and to increase their effectiveness. For tools of this kind to achieve their full potential in improving policing, though, the legal system will need to adapt. One important change would be to understand police tactics such as investigatory stops of pedestrians or motorists as programs, not as isolated occurrences. Beyond that, the judiciary will need to grow more comfortable with statistical proof of discriminatory policing, and the police will need to be more receptive to the assistance that algorithms can provide in reducing bias.

Goel, Sharad, Maya Perelman, Ravi Shroff, and David Alan Sklansky. "Combatting Police Discrimination in the Age of Big Data." New Criminal Law Review (2016), Forthcoming.

How the Machine 'Thinks:' Understanding Opacity in Machine Learning Algorithms

Jenna Burrell

This article considers the issue of opacity as a problem for socially consequential mechanisms of classification and ranking, such as spam filters, credit card fraud detection, search engines, news trends, market segmentation and advertising, insurance or loan qualification, and credit scoring. These mechanisms of classification all frequently rely on computational algorithms, and in many cases on machine learning algorithms to do this work. In this article, I draw a distinction between three forms of opacity: (1) opacity as intentional corporate or state secrecy (2) opacity as technical illiteracy, and (3) an opacity that arises from the characteristics of machine learning algorithms and the scale required to apply them usefully. The analysis in this article gets inside the algorithms themselves. I cite existing literatures in computer science, known industry practices (as they are publicly presented), and do some testing and manipulation of code as a form of lightweight code audit. I argue that recognizing the distinct forms of opacity that may be coming into play in a given application is key to determining which of a variety of technical and non-technical solutions could help to prevent harm.

Burrell, Jenna. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." Big Data & Society 3, no. 1 (2016): 10.1177/2053951715622512.

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai

The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with word embedding, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises concerns because their widespread use, as we describe, often tends to amplify these biases. Geometrically, gender bias is first shown to be captured by a direction in the word embedding. Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding. Using these properties, we provide a methodology for modifying an embedding to remove gender stereotypes, such as the association between between the words receptionist and female, while maintaining desired associations such as between the words queen and female. We define metrics to quantify both direct and indirect gender biases in embeddings, and develop algorithms to "debias" the embedding. Using crowd-worker evaluation as well as standard benchmarks, we empirically demonstrate that our algorithms significantly reduce gender bias in embeddings while preserving the its useful properties such as the ability to cluster related concepts and to solve analogy tasks. The resulting embeddings can be used in applications without amplifying gender bias.

Semantics Derived Automatically from Language Corpora Necessarily Contain Human Biases

Aylin Caliskan-Islam, Joanna J. Bryson, Arvind Narayanan

Artificial intelligence and machine learning are in a period of astounding growth. However, there are concerns that these technologies may be used, either with or without intention, to perpetuate the prejudice and unfairness that unfortunately characterizes many human institutions. Here we show for the first time that human-like semantic biases result from the application of standard machine learning to ordinary language—the same sort of language humans are exposed to every day. We replicate a spectrum of standard human biases as exposed by the Implicit Association Test and other well-known psychological studies. We replicate these using a widely used, purely statistical machine-learning model—namely, the GloVe word embedding—trained on a corpus of text from the Web. Our results indicate that language itself contains recoverable and accurate imprints of our historic biases, whether these are morally neutral as towards insects or flowers, problematic as towards race or gender, or even simply veridical, reflecting the status quo for the distribution of gender with respect to careers or first names. These regularities are captured by machine learning along with the rest of semantics. In addition to our empirical findings concerning language, we also contribute new methods for evaluating bias in text, the Word Embedding Association Test (WEAT) and the Word Embedding Factual Association Test (WEFAT). Our results have implications not only for AI and machine learning, but also for the fields of psychology, sociology, and human ethics, since they raise the possibility that mere exposure to everyday language can account for the biases we replicate here.

How to be Fair and Diverse?

L. Elisa Celis, Amit Deshpande, Tarun Kathuria, Nisheeth K. Vishnoi

Due to the recent cases of algorithmic bias in data-driven decision-making, machine learning methods are being put under the microscope in order to understand the root cause of these biases and how to correct them. Here, we consider a basic algorithmic task that is central in machine learning: subsampling from a large data set. Subsamples are used both as an end-goal in data summarization (where fairness could either be a legal, political or moral requirement) and to train algorithms (where biases in the samples are often a source of bias in the resulting model). Consequently, there is a growing effort to modify either the subsampling methods or the algorithms themselves in order to ensure fairness. However, in doing so, a question that seems to be overlooked is whether it is possible to produce fair subsamples that are also adequately representative of the feature space of the data set - an important and classic requirement in machine learning. Can diversity and fairness be simultaneously ensured? We

start by noting that, in some applications, guaranteeing one does not necessarily guarantee the other, and a new approach is required. Subsequently, we present an algorithmic framework which allows us to produce both fair and diverse samples. Our experimental results on an image summarization task show marked improvements in fairness without compromising feature diversity by much, giving us the best of both the worlds.

Exploring or Exploiting? Social and Ethical Implications of Autonomous Experimentation in AI

Sarah Bird, Solon Barocas, Fernando Diaz, Hanna Wallach and Kate Crawford

In the field of computer science, large-scale experimentation on users is not new. However, driven by advances in artificial intelligence, novel autonomous systems for experimentation are emerging that raise complex, unanswered questions for the field. Some of these questions are computational, while others relate to the social and ethical implications of these systems. We see these normative questions as urgent because they pertain to critical infrastructure upon which large populations depend, such as transportation and healthcare. Although experimentation on widely used online platforms like Facebook has stoked controversy in recent years, the unique risks posed by autonomous experimentation have not received sufficient attention, even though such techniques are being trialled on a massive scale. In this paper, we identify several questions about the social and ethical implications of autonomous experimentation systems. These questions concern the design of such systems, their effects on users, and their resistance to some common mitigations.

Rawlsian Fairness for Machine Learning

Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel and Aaron Roth

Motivated by concerns that automated decision-making procedures can unintentionally lead to discriminatory behavior, we study a technical definition of fairness modeled after John Rawls' notion of "fair equality of opportunity". In the context of a simple model of online decision making, we give an algorithm that satisfies this fairness constraint, while still being able to learn at a rate that is comparable to (but necessarily worse than) that of the best algorithms absent a fairness constraint. We prove a regret bound for fair algorithms in the linear contextual bandit framework that is a significant improvement over our technical companion paper [16], which gives black-box reductions in a more general setting. We analyze our algorithms both theoretically and experimentally. Finally, we introduce the notion of a "discrimination index", and show that standard algorithms for our problem exhibit structured discriminatory behavior, whereas the "fair" algorithms we develop do not.

Iterative Orthogonal Feature Projection for Diagnosing Bias in Black-Box Models

Julius Adebayo and Lalana Kagal

Price of Transparency in Strategic Machine Learning

Emrah Akyol, Cedric Langbort and Tamer Basar

Based on the observation that the transparency of an algorithm comes with a cost for the algorithm designer when the users (data providers) are strategic, this paper studies the impact of strategic intent of the users on the design and performance of transparent ML algorithms. We quantitatively study the {**price of transparency**} **in the context of strategic classification algorithms, by modeling the problem as a nonzero-sum game between the users and the algorithm designer. The cost of having a transparent algorithm is measured by a quantity, named here as price of transparency which is the ratio of the designer cost at the Stackelberg equilibrium, when the algorithm is transparent (which allows users to be strategic) to that of the setting where the algorithm is not transparent.**

Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments

Alexandra Chouldechova

Recidivism prediction instruments provide decision makers with an assessment of the likelihood that a criminal defendant will reoffend at a future point in time. While such instruments are gaining increasing popularity across the country, their use is attracting tremendous controversy. Much of the controversy concerns potential discriminatory bias in the risk assessments that are produced. This paper discusses a fairness criterion originating in the field of educational and psychological testing that has recently been applied to assess the fairness of recidivism prediction instruments. We demonstrate how adherence to the criterion may lead to considerable disparate impact when recidivism prevalence differs across groups.

The Case for Temporal Transparency: Detecting Policy Change Events in Black-Box Decision Making Systems

Miguel Ferreira, Muhammad Bilal Zafar and Krishna Gummadi

Bringing transparency to black-box decision making systems (DMS) has been a topic of increasing research interest in recent years. Traditional active and passive approaches to make these systems transparent are often limited by scalability and/or feasibility issues. In this paper, we propose a new notion of black-box DMS transparency, named, temporal transparency, whose goal is to detect if/when the DMS policy changes over time, and is mostly invariant to the drawbacks of traditional approaches. We map our notion of temporal transparency to time series changepoint detection methods, and develop a framework to detect policy changes in real-world DMS's. Experiments on New York Stop-question-and-frisk dataset reveal a number of publicly announced and unannounced policy changes, highlighting the utility of our framework.

Fair Learning in Markovian Environments

Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern and Aaron Roth

We introduce the study of fairness in the setting of reinforcement learning. We adapt our notion of fairness from Joseph et al. (2016), which requires that over the course of its learning process a learning algorithm never favors a worse action over a better one, with high probability. Working in the setting of Markov Decision Processes, we begin by proving an exponential separation in the time required for fair and unfair algorithms to achieve near- optimal performance. This separation motivates a relaxation to a new notion of approximate fairness, for which we prove more favorable lower and upper bounds.

Inherent Trade-Offs in the Fair Determination of Risk Scores

Jon Kleinberg, Sendhil Mullainathan and Manish Raghavan

Recent discussion in the public sphere about algorithmic classification has involved tension between competing notions of what it means for a probabilistic classification to be fair to different groups. We formalize three fairness conditions that lie at the heart of these debates, and we prove that except in highly constrained special cases, there is no method that can satisfy these three conditions simultaneously. Moreover, even satisfying all three conditions approximately requires that the data lie in an approximate version of one of the constrained special cases identified by our theorem. These results suggest some of the ways in which key notions of fairness are incompatible with

each other, and hence provide a framework for thinking about the trade-offs between them.

A Statistical Framework for Fair Predictive Algorithms

Kristian Lum and James Johndrow

Predictive modeling is increasingly being employed to assist human decision-makers. One purported advantage of replacing human judgment with computer models in high stakes settings– such as sentencing, hiring, policing, college admissions, and parole decisions– is the perceived "neutrality" of computers. It is argued that because computer models do not hold personal prejudice, the predictions they produce will be equally free from prejudice. There is growing recognition that employing algorithms does not remove the potential for bias, and can even amplify it, since training data were inevitably generated by a process that is itself biased. In this paper, we provide a probabilistic definition of algorithmic bias. We propose a method to remove bias from predictive models by removing all information regarding protected variables from the permitted training data. Unlike previous work in this area, our framework is general enough to accommodate arbitrary data types, e.g. binary, continuous, etc. Motivated by models currently in use in the criminal justice system that inform decisions on pre-trial release and paroling, we apply our proposed method to a dataset on the criminal histories of individuals at the time of sentencing to produce "race-neutral" predictions of re-arrest. In the process, we demonstrate that the most common approach to creating "race-neutral" models– omitting race as a covariate– still results in racially disparate predictions. We then demonstrate that the application of our proposed method to these data removes racial disparities from predictions with minimal impact on predictive accuracy.

Measuring Fairness in Ranked Outputs

Ke Yang and Julia Stoyanovich

Ranking and scoring are ubiquitous. We consider the setting in which an institution, called a ranker, evaluates a set of individuals based on demographic, behavioral or other characteristics. The final output is a ranking that represents the relative quality of the individuals. While automatic and therefore seemingly objective, rankers can, and often do, discriminate against individuals and systematically disadvantage members of protected groups. This warrants a careful study

of the fairness of a ranking scheme. In this paper we propose fairness measures for ranked outputs. We develop a data generation procedure that allows us to systematically control the degree of unfairness in the output, and study the behavior of our measures on these datasets. We then apply our proposed measures to several real datasets, and demonstrate cases of unfairness. Finally, we show preliminary results of incorporating our ranked fairness measures into an optimization framework, and show potential for improving fairness of ranked outputs while maintaining accuracy.

Fairness Beyond Disparate Treatment and Disparate Impact: Learning Classification without Disparate Mistreatment

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez and Krishna Gummadi

As the use of automated decision making systems becomes widespread, there is a growing concern about their potential unfairness towards people with certain traits. Anti-discrimination laws in various countries prohibit unfair treatment of individuals based on specific traits, also called sensitive attributes (e.g., gender, race). In many learning scenarios, the trained algorithms (classifiers) make decisions with certain inaccuracy (misclassification rate). As learning mechanisms target minimizing the error rate for all decisions, it is quite possible that the optimally trained algorithm makes decisions for users belonging to different sensitive attribute groups with different error rates (e.g., decision errors for females are higher than for males). To account for and avoid such unfairness when learning, in this paper, we introduce a new notion of unfairness, disparate mistreatment, which is defined in terms of misclassification rates. We then propose an intuitive measure of disparate mistreatment for decision boundary-based classifiers, which can be easily incorporated into their formulation as a convex-concave constraint. Experiments on synthetic as well as real world datasets show that our methodology is effective at avoiding disparate mistreatment, often at a small cost in terms of accuracy.

Interpretable Classification Models for Recidivism Prediction

Jiaming Zeng, Berk Ustun, Cynthia Rudin

We investigate a long-debated question, which is how to create predictive models of recidivism that are sufficiently accurate, transparent, and interpretable to use for decision-making. This question is complicated as these models are used to support different decisions, from sentencing, to determining release on probation, to allocating preventative social services. Each use case might have an objective other than classification accuracy, such as a desired true positive rate (TPR) or false positive rate (FPR). Each (TPR, FPR) pair is a point on the receiver operator characteristic (ROC) curve. We use popular machine learning methods to create models along the full ROC curve on a wide range of recidivism prediction problems. We show that many methods (SVM, Ridge Regression) produce equally accurate models along the full ROC curve. However, methods that designed for interpretability (CART, C5.0) cannot be tuned to produce models that are accurate and/or interpretable. To handle this shortcoming, we use a new method known as SLIM (Supersparse Linear Integer Models) to produce accurate, transparent, and interpretable models along the full ROC curve. These models can be used for decision-making for many different use cases, since they are just as accurate as the most powerful black-box machine learning models, but completely transparent, and highly interpretable.

Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems

Anupam Datta, Shayak Sen, Yair Zick

Algorithmic systems that employ machine learning play an increasing role in making substantive decisions in modern society, ranging from online personalization to insurance and credit decisions to predictive policing. But their decision-making processes are often opaque—it is difficult to explain why a certain decision was made. We develop a formal foundation to improve the transparency of such decision-making systems. Specifically, we introduce a family of Quantitative Input Influence (QII) measures that capture the degree of influence of inputs on outputs of systems. These measures provide a foundation for the design of transparency reports that accompany system decisions (e.g., explaining a specific credit decision) and for testing tools useful for internal and external oversight (e.g., to detect algorithmic discrimination).

Distinctively, our causal QII measures carefully account for correlated inputs while measuring influence. They support a general class of transparency queries and can, in particular, explain decisions about individuals (e.g., a loan decision) and groups (e.g., disparate impact

based on gender). Finally, since single inputs may not always have high influence, the QII measures also quantify the joint influence of a set of inputs (e.g., age and income) on outcomes (e.g. loan decisions) and the marginal influence of individual inputs within such a set (e.g., income). Since a single input may be part of multiple influential sets, the average marginal influence of the input is computed using principled aggregation measures, such as the Shapley value, previously applied to measure influence in voting. Further, since transparency reports could compromise privacy, we explore the transparency-privacy tradeoff and prove that a number of useful transparency reports can be made differentially private with very little addition of noise.

Our empirical validation with standard machine learning algorithms demonstrates that QII measures are a useful transparency mechanism when black box access to the learning system is available. In particular, they provide better explanations than standard associative measures for a host of scenarios that we consider. Further, we show that in the situations we consider, QII is efficiently approximable and can be made differentially private while preserving accuracy.

'Why Should I Trust You?' Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

Despite widespread adoption, machine learning models remain mostly black boxes. Understanding the reasons behind predictions is, however, quite important in assessing trust, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model. Such understanding also provides insights into the model, which can be used to transform an untrustworthy model or prediction into a trustworthy one. In this work, we propose LIME, a novel explanation technique that explains the predictions of any classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction. We also propose a method to explain models by presenting representative individual predictions and their explanations in a non-redundant way, framing the task as a submodular optimization problem. We demonstrate the flexibility of these methods by explaining different models for text (e.g. random forests) and image classification (e.g. neural networks). We show the utility of explanations via novel experiments, both simulated and with human subjects, on various scenarios that require trust: deciding if one should trust a prediction, choosing between models, improving an untrustworthy classifier, and identifying why a classifier should not be trusted.

Fairness as a Program Property

Aws Albarghouthi, Loris D'Antoni, Samuel Drews and Aditya Nori

We explore the following question: Is a decision-making program fair, for some useful definition of fairness? First, we describe how several algorithmic fairness questions can be phrased as program verification problems. Second, we discuss an automated verification technique for proving or disproving fairness of decision-making programs with respect to a probabilistic model of the population.

2017

Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English

Su Lin Blodgett and Brendan O'Connor

We highlight an important frontier in algorithmic fairness: disparity in the quality of natural language processing algorithms when applied to language from authors of different social groups. For example, current systems sometimes analyze the language of females and minorities more poorly than they do of whites and males. We conduct an empirical analysis of racial disparity in language identification for tweets written in African-American English, and discuss implications of disparity in NLP.

Links: Video

Fair Clustering Through Fairlets

Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi and Sergei Vassilvitskii

We study the question of fair clustering under the disparate impact doctrine, where each protected class must have approximately equal representation in every cluster. We formulate the fair clustering problem under both the k-center and the k-median objectives, and show that even with two protected classes the problem is challenging, as the optimum solution can violate common conventions—for instance a point may no longer be assigned to its nearest cluster center!

En route we introduce the concept of fairlets, which are minimal sets that satisfy fair representation while approximately preserving the

clustering objective. We show that any fair clustering problem can be decomposed into first finding good fairlets, and then using existing machinery for traditional clustering algorithms. While finding good fairlets can be NP-hard, we proceed to obtain efficient approximation algorithms based on minimum cost flow.

We empirically demonstrate the price of fairness by comparing the value of fair clustering on real-world datasets with sensitive attributes.

Links: Video

Calibrated Fairness in Bandits

Yang Liu, Goran Radanovic, Christos Dimitrakakis, David Parkes and Debmalya Mandal

We study fairness within the stochastic, multi-armed bandit (MAB) decision making framework. We adapt the fairness framework of "treating similar individuals similarly" to this setttting. Here, an 'individual' corresponds to an arm and two arms are 'similar' if they have a similar quality distribution. First, we adopt a smoothness constraint that if two arms have a similar quality distribution then the probability of selecting each arm should be similar. In addition, we define the fairness regret, which corresponds to the degree to which an algorithm is not calibrated, where perfect calibration requires that the probability of selecting an arm is equal to the probability with which the arm has the best quality realization. We show that a variation on Thompson sampling satisfies smooth fairness for total variation distance, and give an O((kT)^2/3) bound on fairness regret. This complements prior work, which protects an on-average better arm from being less favored. We also explain how to extend our algorithm to the dueling bandit setting.

Links: Video

The Authority of "Fair" in Machine Learning

Michael Skirpan and Micha Gorelick

We argue for the adoption of a normative definition of fairness within the machine learning community. After characterizing this definition, we review the current literature of Fair ML in light of its implications. We end by suggesting ways to incorporate a broader community and generate further debate around how to decide what is fair in ML.

Links: Video

Runaway Feedback Loops in Predictive Policing

Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger and Suresh Venkatasubramanian

Predictive policing systems are increasingly used to determine how to allocate police across a city in order to best prevent crime. Observed crime data (arrest counts) are used to update the model, and the process is repeated. Such systems have been shown susceptible to runaway feedback loops, where police are repeatedly sent back to the same neighborhoods regardless of the true crime rate. In response, we develop a model of predictive policing that shows why this feedback loop occurs, show empirically that this model exhibits such problems, and demonstrate how to change the inputs to a predictive policing system (in a black-box manner) so the runaway feedback loop does not occur, allowing the true crime rate to be learned.

Links: Video

Fairness at Equilibrium in the Labor Market

Lily Hu and Yiling Chen

Recent literature on computational notions of fairness has been broadly divided into two distinct camps, supporting interventions that address either individual-based or group-based fairness. Rather than privilege a single definition, we seek to resolve both within the particular domain of employment discrimination. To this end, we construct a dual labor market model composed of a Temporary Labor Market, in which form strategies are constrained to ensure group-level fairness, and a Permanent Labor Market, in which individual worker fairness is guaranteed. We show that such restrictions on hiring practices induces an equilibrium that Pareto dominates those arising from strategies that employ statistical discrimination or a "group-blind" criterion. Individual worker reputations produce externalities for collective reputation, generating a feedback loop termed a "self-fulfilling prophecy." Our model produces its own feedback loop, raising the collective reputation of an initially disadvantaged group via a fairness intervention that need not be permanent. Moreover, we show that, contrary to popular assumption, the asymmetric equilibria resulting from hiring practices that disregard group-fairness may be immovable without targeted intervention. the enduring nature of such equilibria that are both inequitable and Pareto inefficient suggest that fairness interventions are of critical importance in moving the labor market to be more socially just and efficient.

Links: Video

Preference vs. Parity-based Notions of Fairness

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi and Adrian Weller

Many notions of fairness in data-driven decision making are inspired by the concept of discrimination in social sciences and law, and focus on ensuring parity (equality) in treatment or outcomes for different social groups. In this paper, we propose preference-based notions of fairness with the goals of avoiding potential 'reverse-discrimination' and enabling high decision accuracy. We introduce tractable proxies to design convex boundary-based classifiers that satisfy these new notions of fairness and show on the ProPublica COMPAS dataset that these notions allow for greater decision accuracy than parity-based fairness.

Links: Video

Logics and Practices of Transparency and Opacity in Real-world Applications of Public Sector Machine Learning

Michael Veale

Machine learning systems are increasingly used to support public sector decision-making across a variety of sectors. Given concerns around accountability in these domains, and amidst accusations of intentional or unintentional bias, there have been increased calls for transparency of these technologies. Few, however, have considered how logics and practices concerning transparency have been understood by those involved in the machine learning systems already being piloted and deployed in public bodies today. This short paper distills insights about transparency on the ground from interviews with 27 such actors, largely public servants and relevant contractors, across 5 OECD countries. Considering transparency and opacity in relation to trust and buy-in, better decision-making, and the avoidance of gaming, it seeks to provide useful insights for those hoping to develop socio-technical approaches to transparency that might be useful to practitioners on-the-ground.

Links: Video

Papers

Fair Algorithms for Infinite Contextual Bandits ∗

Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, Aaron Roth

University of Pennsylvania

We study fairness in infinite linear bandit problems. Starting from the notion of meritocratic fairness intro- duced in Joseph et al. [9], we expand their notion of fair- ness for infinite action spaces and provide an algorithm that obtains a sublinear but instance-dependent regret guarantee. We then show that this instance depen- dence is a necessary cost of our fairness definition with a matching lower bound. This provides a strong contrast with the traditional non-fair setting, where instance- independent regret bounds are achievable. Finally, we exhibit an action space in which fair algorithms cannot even obtain nontrivial instance-dependent bounds.

Better Fair Algorithms for Contextual Bandits ∗

Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, Aaron Roth

University of Pennsylvania

We study fairness in the linear bandit setting. Start- ing from the notion of meritocratic fairness introduced in Joseph et al. [11], we introduce a sufficiently more gen- eral model in which meritocratic fairness can be imposed and satisfied. We then perform a more fine-grained anal- ysis which achieves better performance guarantees in this more general model. Our work therefore studies fairness for a more general problem and provides tighter performance guarantees than previous work in the sim- pler setting.

Multisided Fairness for Recommendation

Robin Burke

Recent work on machine learning has begun to con- sider issues of fairness. In this paper, we extend the concept of fairness to recommendation. In particular, we show that in some recommendation contexts, fair- ness may be a multisided concept, in which fair outcomes for multiple individuals need to be considered. Based on these considerations, we present a taxonomy of classes of fairness-aware

recommender systems and suggest possible fairness-aware recommendation archi- tectures.

Interpretability via Model Extraction*

Osbert Bastani, Carolyn Kim, Hamsa Bastani

The ability to interpret machine learning models has become increasingly important now that machine learn- ing is used to inform consequential decisions. We pro- pose an approach called model extraction for interpret- ing complex, blackbox models. Our approach approximates the complex model using a much more inter- pretable model; as long as the approximation quality is good, then statistical properties of the complex model are reflected in the interpretable model. We show how model extraction can be used to understand and debug random forests and neural nets trained on several datasets from the UCI Machine Learning Repository, as well as control policies learned for several classical reinforcement learning problems.

A Reductions Approach to Fair Classification

Alekh Agarwal, Miroslav Dudík, Alina Beygelzimer, John Langford

We present a systematic method for training-time en- forcement of definitions of fairness in binary classifi- cation with protected attributes. Taking two popular definitions, demographic parity and equalized odds, we show how to reduce fairness-constrained classification to cost-sensitive classification. The reduction is agnos- tic to the representation and form of the cost-sensitive classifier, allowing a range of existing algorithms. Em- pirical evaluation shows that the systematic incorpo- ration of fairness constraints allows us to outperform prior baselines for both definitions.

A Convex Framework for Fair Regression*

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph Michael Kearns, Jamie Morgenstern, Seth Neel, Aaron Roth University of Pennsylvania

We introduce a flexible family of fairness regularizers for (linear and logistic) regression problems. These regular- izers all enjoy convexity, permitting fast optimization, and span the range from group fairness to strong indi- vidual fairness. We study the accuracy-fairness trade-off on any given dataset, and we measure the severity of this

trade-off via a numerical quantity we call the Price of Fairness (PoF). The centerpiece of our results is an ex- tensive comparative study of the PoF across six different datasets in which fairness is a primary consideration.

Is it ethical to avoid error analysis?

Eva García-Martín, Niklas Lavesson

Machine learning algorithms tend to create more accu- rate models with the availability of large datasets. In some cases, highly accurate models can hide the pres- ence of bias in the data. There are several studies pub- lished that tackle the development of discriminatory-aware machine learning algorithms. We center on the further evaluation of machine learning models by do- ing error analysis, to understand under what condi- tions the model is not working as expected. We focus on the ethical implications of avoiding error analysis, from a falsification of results and discrimination per- spective. Finally, we show different ways to approach error analysis in non-interpretable machine learning algorithms such as deep learning.

Fair Pipelines∗

Amanda Bower, Sarah N. Kitchen, Laura Niss, Martin J. Strauss, Alexander Vargo, Suresh Venkatasubramanian

This work facilitates ensuring fairness of machine learning in the real world by decoupling fairness considerations in compound de- cisions. In particular, this work studies how fairness propagates through a compound decision-making processes, which we call a pipeline. Prior work in algorithmic fairness only focuses on fair- ness with respect to one decision. However, many decision-making processes require more than one decision. For instance, hiring is at least a two stage model: deciding who to interview from the applicant pool and then deciding who to hire from the interview pool. Perhaps surprisingly, we show that the composition of fair components may not guarantee a fair pipeline under a (1 +)-equal opportunity definition of fair. However, we identify circumstances that do provide that guarantee. We also propose numerous direc- tions for future work on more general compound machine learning decisions.

Decoupled classifiers for fair and efficient machine learning

Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson

When it is ethical and legal to use a sensitive attribute (such as gender or race) in machine learning systems, the question remains how to do so. We show that the naive application of machine learning algorithms using sensitive attributes leads to an inherent tradeoff in accuracy between groups. We provide a simple and efficient decoupling technique, that can be added on top of any black-box machine learning algorithm, to learn different classifiers for different groups. The method can apply to a range of fairness criteria. In particular, we require the application designer to specify as joint loss function that makes explicit the trade-off between fairness and accuracy. Our reduction is shown to efficiently find the global optimum loss as long as the objective has a certain natural monotonicity property. Monotonicity may be of independent interest in the study of fairness in algorithms.

Penalizing Unfairness in Binary Classification

Yahav Bechavod, Katrina Ligett

We present a new approach for mitigating un- fairness in learned classifiers. In particular, we focus on binary classification tasks over individu- als from two populations, where, as our criterion for fairness, we wish to achieve similar false pos- itive rates in both populations, and similar false negative rates in both populations. As a proof of concept, we implement our approach and empiri- cally evaluate its ability to achieve both fairness and accuracy, using datasets from the fields of criminal risk assessment, credit, lending, and col- lege admissions.

Fairer and more accurate, but for whom?

Alexandra Chouldechova, Max G'Sell

Complex statistical machine learning models are in- creasingly being used or considered for use in high- stakes decision-making pipelines in domains such as financial services, health care, criminal justice and hu- man services. These models are often investigated as possible improvements over more classical tools such as regression models or human judgement. While the modeling approach may be new, the practice of using some form of risk assessment to inform decisions is

not. When determining whether a new model should be adopted, it is therefore essential to be able to compare the proposed model to the existing approach across a range of task-relevant accuracy and fairness metrics. Looking at overall performance metrics, however, may be misleading. Even when two models have comparable overall performance, they may nevertheless disagree in their classifications on a considerable fraction of cases. In this paper we introduce a model comparison frame- work for automatically identifying subgroups in which the differences between models are most pronounced. Our primary focus is on identifying subgroups where the models differ in terms of fairness-related quantities such as racial or gender disparities. We present experi- mental results from a recidivism prediction task and a hypothetical lending example.

Fair Personalization

∗ L. Elisa Celis, Nisheeth K. Vishnoi

Personalization is pervasive in the online space as, when combined with learning, it leads to higher efficiency and revenue by allowing the most relevant content to be served to each user. However, recent studies sug- gest that such personalization can propagate societal or systemic biases, which has led to calls for regula- tory mechanisms and algorithms to combat inequality. Here we propose a rigorous algorithmic framework that allows for the possibility to control biased or dis- criminatory personalization with respect to sensitive attributes of users without losing all of the benefits of personalization.

Discriminatory Transfer

Chao Lan, Jun Huan

We observe standard transfer learning can improve pre- diction accuracies of target tasks at the cost of lowering their prediction fairness – a phenomenon we named discriminatory transfer. We examine prediction fairness of a standard hypothesis transfer algorithm and a stan- dard multi-task learning algorithm, and show they both suffer discriminatory transfer on the real-world Com- munities and Crime data set. The presented case study introduces an interaction between fairness and trans- fer learning, as an extension of existing fairness studies that focus on single task learning.

On Fairness, Diversity and Randomness in Algorithmic Decision Making

Nina Grgić-Hlača1, Muhammad Bilal Zafar1, Krishna P. Gummadi1, and Adrian Weller

Consider a binary decision making process where a single machine learning classifier replaces a mul- titude of humans. We raise questions about the resulting loss of diversity in the decision making process. We study the potential benefits of using random classifier ensembles instead of a single classifier in the context of fairness-aware learning and demonstrate various attractive properties: (i) an ensemble of fair classifiers is guaranteed to be fair, for several different measures of fairness, (ii) an ensemble of unfair classifiers can still achieve fair outcomes, and (iii) an ensemble of classifiers can achieve better accuracy-fairness trade-offs than a single classifier. Finally, we introduce notions of distributional fairness to characterize further potential benefits of random classifier en- sembles.

Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations

Alex Beutel, Jilin Chen, Zhe Zhao, Ed H. Chi

How can we learn a classifier that is "fair" for a protected or sen- sitive group, when we do not know if the input to the classifier belongs to the protected group? How can we train such a classifier when data on the protected group is difficult to attain? In many set- tings, finding out the sensitive input attribute can be prohibitively expensive even during model training, and sometimes impossible during model serving. For example, in recommender systems, if we want to predict if a user will click on a given recommendation, we often do not know many attributes of the user, e.g., race or age, and many attributes of the content are hard to determine, e.g., the language or topic. Thus, it is not feasible to use a different classifier calibrated based on knowledge of the sensitive attribute. Here, we use an adversarial training procedure to remove infor- mation about the sensitive attribute from the latent representation learned by a neural network. In particular, we study how the choice of data for the adversarial training effects the resulting fairness properties. We find two interesting results: a small amount of data is needed to train these adversarial models, and the data distribution empirically drives the adversary's notion of fairness.

The causal impact of bail on case outcomes for indigent defendants

Kristian Lum, Mike Baiocchi

New Fairness Metrics for Recommendation that Embrace Differences

Sirui Yao, Bert Huang

We study fairness in collaborative-filtering recommender systems, which are sensitive to discrimination that ex- ists in historical data. Biased data can lead collabora- tive filtering methods to make unfair predictions against minority groups of users. We identify the insufficiency of existing fairness metrics and propose four new met- rics that address different forms of unfairness. These fairness metrics can be optimized by adding fairness terms to the learning objective. Experiments on syn- thetic and real data show that our new metrics can bet- ter measure fairness than the baseline, and that the fair- ness objectives effectively help reduce unfairness.

Identifying Significant Predictive Bias in Classifiers

Zhe Zhang, Daniel Neill

We present a novel subset scan method to detect if a prob- abilistic binary classifier has statistically significant bias — over or under predicting the risk — for some subgroup, and identify the characteristics of this subgroup. This form of model checking and goodness-of-fit test provides a way to interpretably detect the presence of classifier bias or regions of poor classifier fit. This allows consideration of not just subgroups of a priori interest or small dimensions, but the space of all possible subgroups of features. To address the difficulty of considering these exponentially many possible subgroups, we use subset scan and parametric bootstrap-based methods. Extending this method, we can penalize the complexity of the detected subgroup and also identify subgroups with high classification errors. We demonstrate these methods and find interesting results on the COMPAS crime recidivism and credit delinquency data.

Interpretable & Explorable Approximations of Black Box Models

Himabindu Lakkaraju, Rich Caruana, Ece Kamar, Jure Leskovec

We propose Black Box Explanations through Transparent Approximations (BETA), a novel model agnostic framework for explaining the behavior of any black-box classifier by simultaneously opti- mizing for fidelity to the original model and interpretability of the explanation. To this end, we develop a novel objective function which allows us to learn (with optimality guarantees), a small num- ber of compact decision sets each of which explains the behavior of the black box model in unambiguous, well-defined regions of feature space. Furthermore, our framework also is capable of ac- cepting user input when generating these approximations, thus allowing users to interactively explore how the black-box model behaves in different subspaces that are of interest to the user. To the best of our knowledge, this is the first approach which can produce global explanations of the behavior of any given black box model through joint optimization of unambiguity, fidelity, and in- terpretability, while also allowing users to explore model behavior based on their preferences. Experimental evaluation with real- world datasets and user studies demonstrates that our approach can generate highly compact, easy-to-understand, yet accurate ap- proximations of various kinds of predictive models compared to state-of-the-art baselines.

Causal Falling Rule Lists

Fulton Wang, Cynthia Rudin

A causal falling rule list (CFRL) is a sequence of if-then rules that specifies heterogeneous treatment effects, where (i) the order of rules determines the treatment ef- fect subgroup a subject belongs to, and (ii) the treatment effect decreases monotonically down the list. A given CFRL parameterizes a hierarchical bayesian regression model in which the treatment effects are incorporated as parameters, and assumed constant within model- specific subgroups. We formulate the search for the CFRL best supported by the data as a Bayesian model selection problem, where we perform a search over the space of CFRL models, and approximate the evidence for a given CFRL model using standard variational tech- niques. We apply CFRL to a census wage dataset to identify subgroups of differing wage inequalities between men and women.