Challenges in Learning From Social Data

William Brown
May 9, 2018

Introduction

The role of data in society today is larger than ever before. With advances in cloud computing and machine learning, the barrier to entry for integrating data science into a business or organization is getting lower every year. Mobile phones and social media have spawned massive and rich data sources which allow us to study people with more depth than ever before. Tech companies are racing to learn as much as they can about their users in order to craft personalized experiences. Institutions such as prisons and universities are leveraging all the information they can find when deciding to grant parole or admission. Governments are using data to paint more vivid pictures of the populations they serve to determine the optimal allocations of resources. Nearly every academic discipline, from political science to linguistics to biology, has found new ways to incorporate data drawn from populations of people into its problem-solving paradigm. Increasingly, decisions which affect the lives of individuals are being made using data collected from the individuals themselves.

The pervasiveness of data-driven decision-making in our society has and will continue to make efficient problem-solving much easier. Yet this also presents a new array of challenges. Recent data breaches from Facebook and Equifax have shown that this level of mass data collection can put sensitive information about people at risk. Concerns about privacy of one’s data are not only justified in cases where social security and credit card numbers are at risk. Individuals may also worry that by providing seemingly innocuous information to surveys or online applications, complex profiles about their habits and preferences can be constructed and used in undesirable ways. As an example, the retail giant Target was able to predict that a young girl was pregnant based on shopping patterns and sent related product recommendations, which subsequently tipped off her parents [7]. One frequently proposed solution is to simply remove information which directly identifies a person from their record in a database. Yet as we will later see, this level of anonymization is often insufficient to prevent sensitive information about individuals from being discovered. Questions of fairness also arise when data-driven decisions are made about resource allocation. Algorithms used to filter job applicants or determine bail or parole eligibility may result in allocations which are discriminatory in some sense. These algorithms may be very complex to understand and difficult to evaluate for discrimination, allowing their designers to feign ignorance. Deciding exactly what it means for an algorithm to be “fair” is a challenge itself. Another increasingly pressing issue is the misuse of adaptive analyses on a dataset. Getting enough of the right data is not always easy or cheap. The pressure to show results in both academia and industry have caused some to forgo the guidance of statistical best practices. The same dataset may be repeatedly scrutinized in order to get results which seem significant, but these exploratory findings are often not reproducible in later experiments.

The aim of this paper is to consider these dilemmas as they relate to analyzing data drawn from people, or “social data”, and review the techniques which have been presented as solutions. We select these topics not only because they address a connected set of issues, but because there is a common thread of beautiful mathematics which connects them as well. Surprisingly, by injecting a small amount of noise into the answers of queries asked
about a dataset, we can ensure that analysis procedures satisfy certain definitions of privacy, fairness, and robustness to adaptivity. These techniques have drawn much recent attention in theoretical research, and we discuss their practical benefits and shortcomings for these modern crises in data science.

Characterizing Social Data

“Social data” is not an incredibly technically precise term, but for our purposes we will broadly consider it to mean any information about people in a society. This will first call to mind information about people’s friends, habits, occupations, or preferences, as drawn from social media services. We will also consider it to include data about one’s financial, educational, or medical history, as well as anything else useful for studying trends and behaviors in society, including answers and actions in behavioral surveys and experiments. This data is often privately held by an individual, or at least is not entirely publicly accessible. Frequently it is difficult to make guarantees about the accuracy of this data in terms of what underlying truth it claims to represent, as people can easily lie in surveys, and Facebook likes are not always indicative of real world preferences. Crucially, this means we need to be particularly careful about how we elicit data from individuals and cannot always assume that our data is drawn i.i.d from some perfect distribution, as our method of elicitation may itself directly bias the data we receive.

Analyzing patterns of social media behavior has become an effective tool for researchers to explore the way in which people communicate. Recent studies involving Twitter and Facebook have begun to answer questions about how lies and false news stories spread, and how posting patterns correlate with personality traits [1,2]. However, these approaches can draw criticism from users who are concerned about their privacy and feel that they are being taken advantage of. A study concerning the emotional reactions of users to various types of content sparked outrage when it was discovered that researchers were manipulating the content users saw without their permission. In some cases, this resulted in users being shown a large number of disturbing or depressing articles in order to prompt and study their negative reactions [3,4].

Despite these possible shortcomings and misuses, the ability to access and analyze large swathes of social data has opened up incredible doors in business and social science. Researchers have seen such remarkable recent success applying these data sources for important prediction and classification tasks that we ought not to discourage these analyses altogether, but rather explore precautions which will better allow us to forge ahead. Our discussion of privacy-preserving techniques for data analysis will showcase techniques which are currently being practiced to mitigate these issues, and which have potential to be adopted much further.

Typical problems in machine learning take the form of a prediction task. Researchers begin with a data set featuring many explanatory variables and some labeling for each point, and they aim to “train” a model which can accurately predict the labeling for future data points. These tasks range from predicting election results based on historical and demographic traits to identifying whether or not an image contains a cat based on the array of pixels. Increasingly, they inform decisions at every level of society, such as whether to grant a loan to an applicant or where to deploy a police force. In these cases, we have reason to worry that algorithms may be discriminatory in some sense; for example, an algorithm which aims to minimize loan default rates may end up disproportionately rejecting applicants from underprivileged backgrounds even if they would be completely capable of repaying their loans. Algorithms with fairness concerns are becoming increasingly widespread. Justice departments have begun using algorithmic techniques which estimate the likelihood of recidivism for an inmate to determine parole eligibility, yet these tools can exhibit systemic bias along demographic lines [5]. Our discussion of fairness-preserving techniques will explore partial solutions to this as well as impossible barriers.
Interpreting a successful machine learning model from a scientific perspective can be a difficult task. Many common machine learning algorithms derive their effectiveness from identifying some sort of latent structure in the dataset, in the form of correlations between combinations of variables and the target label. For model classes from a simple linear regression to a multi-layer neural network, the training process is effectively searching a space of parameter values for the coefficients of the hypothesis to discover the model which most accurately predicts labels. It is rarely the case that we believe a dataset is actually generated by the exact hypothesis class that we are using to learn. We might be able to frequently predict weather, or stock prices, or adoption of language patterns with some sort of regression model, but the dynamics of physical and social systems which cause these behaviors are far too complex to be entirely described with a small handful of numerical coefficients. In training a successful model, we have discovered something; not a law about how things must work, but a reasonably strong heuristic for some pattern in reality. Too often, the results from an experiment or model are treated as fact. The dilemma of “p-hacking” has led to a number of published results which turned out to be unreproducible, prompting concerns throughout the scientific community about the need for revisions in methodology [6]. Our discussion of techniques for data reuse will explore solutions to this problem as well.

Privacy

It is quite common for people to value their privacy, and most developed countries have legal codes codifying privacy as a right within certain domains. The idea that privacy is important is generally non-controversial, but answers to questions about when privacy matters the most, why it matters, and how it can be truly protected are more elusive. When a person wishes to keep some piece of information withheld from public knowledge, there typically is a concern that the release of the information could result in some sort of adverse harm to the person. We may be inclined to apply caution about our information in a wide variety of situations, ranging from not telling an embarrassing story to our friends, to lying on a survey about medical history out of a concern that our insurance premiums may go up if the data is shared with our provider. If someone desires to conduct research requiring data from individuals which may be considered sensitive, the possibility that someone may misreport or intentionally withhold some of their data damages the ability to trust their findings.

For years, the reigning mentality about conducting analyses involving sensitive data was that anonymizing the dataset by stripping away information directly linked to an individual, such as their name, email, or address, was enough to protect participants from adverse harm. Yet it has become increasingly apparent that this approach is often insufficient. In 2006 (?), the movie service Netflix announced a competition for data scientists, who would compete to improve upon Netflix’s algorithm for content recommendations in hopes of winning a million dollar prize. Along with the competition, they released a dataset containing the star rankings assigned to movies by hundreds of thousands of users whose personal identifiers were removed. One might assume that the ranking out of five stars assigned to a movie by someone would reveal very little about them, as there are likely many other users who assigned the same rating to the movie. However, when the full vector of rankings (or lack of rankings) assigned by a user to all movies on Netflix is available, as was the case for the Netflix Challenge, much more is possible. In 2008, Arvind Narayanan and Vitaly Shmatikov from the University of Texas published a paper demonstrating how this “anonymized” information could be used to recover the exact identities of many of the dataset participants [8]. Their approach was to gather another large dataset of movie rankings from the popular movie ranking site IMDb where personal information was directly available, and they observed that many of these profiles matched up very closely with a specific record in the Netflix database. Apparently, if someone is a frequent user of both Netflix and IMDb, they tend to submit highly similar movie rankings to both sites. By matching up nearly identical
profiles they were able to attach names to much of the Netflix database, but this poses a bigger risk to the participants of that dataset than just revealing information which they already shared online. In many cases, someone may have an opinion about a movie which they may not want revealed publicly, such as in the case of a film which is highly sexual or political in nature. These ratings may have been present in the Netflix database and not IMDb, but if the remainder of their ratings are close enough between the two datasets, they can still be uniquely identified, and their perceived privacy will have disintegrated. Their paper discusses other similar linkage attacks, such as a deanonymization of medical records using public voter registration data, and argue that privacy-preservation tactics must extend beyond removing names from a dataset.

Differential Privacy

The inability to rely on anonymization for protecting people’s privacy has motivated a great deal of exploration into new techniques, resulting in a mathematical guarantee known as “differential privacy”, invented by Cynthia Dwork in 2006 [9]. Differential privacy gives a promise to an individual which aligns with the initial reason why we care about privacy in the first place: a promise to an individual that they will be not be adversely affected by participating, regardless of any post-processing or joining with additional datasets.

Before providing the mathematical definition of differential privacy, we illustrate its application via a paradigm of accurately eliciting sensitive information from survey participants which is commonly used in the social sciences. Imagine you wish to understand the prevalence of some stigmatized behavior in a population, such as drug use or cheating on an exam, but you worry that survey participants are untrustworthy about the way the data will be used, and may lie in order to avoid the possibility of punishment. One workaround is to ask every participant to flip a coin before reporting their answer a question, and to not reveal the result of the flip. If heads, they are instructed to tell the truth. If tails, they flip again, and report one predetermined answer (such as "I cheated") for heads and another for tails ("I didn’t cheat"). Once all reports are collected, the true proportion of an answer can be estimated with high accuracy by adjusting for the expected results of the coin flips. By introducing randomization, anyone worried about their answer being revealed has the plausible deniability that they were simply following the coin flips, and thus they shouldn’t be dissuaded from telling the truth.

Formally, differential privacy is a mathematical guarantee about randomized algorithms which take a dataset as an input and produces some output as a result [9]. Two datasets are considered to be “neighboring” if they differ only in a single entry. A randomized algorithm $M$ is $(\epsilon)$-differentially private if, for all possible sets $S$ of outputs of $M$ and for all neighboring datasets $x$ and $y$:

$$\Pr[M(x) \in S] \leq \exp(\epsilon) \Pr[M(y) \in S]$$

Intuitively, this means that regardless of whether an individual person decides to participate in a dataset, the probability that a differentially private algorithm behaves differently as a result of their inclusion is quite low. And because this guarantee holds over the entire output of the algorithm, any sort of post-processing which attempts to identify an individual person’s record is unable to tell whether that person is even in the dataset, so a sense of robustness greatly exceeding that of simple anonymization procedures is obtained. The $\epsilon$ term acts as a measurement of the degree to which privacy is preserved. If it is quite high, the promise that increased likelihood of some negative outcome is bounded by a large factor is not very comforting. Thankfully, there exist many differentially private mechanisms for practical data analysis scenarios where quite useful results can be obtained even with a miniscule value of $\epsilon$. A common way for such a mechanism to work is to only allow analysts to ask certain types of questions, such as the proportion of the dataset exhibiting a specific attribute. The answers to these queries are then perturbed with a small amount of noise.
These algorithms typically require explicit tradeoffs between accuracy and privacy; however, any desired combination of privacy and accuracy (other than perfection) is achievable given enough data.

If a problem involving social data can be answered by asking questions of the form permitted by a differentially private algorithm, the issues presented by privacy concerns may be surmountable. However, this is not a one-size-fits-all solution. Differentially private algorithms tend to perform poorly if the records in a dataset are relatively sparse. This is the case for the Netflix dataset, as any user has typically only reviewed a tiny fraction of the movies available. The amount of noise necessary here to provide meaningful privacy would nearly eliminate the ability to learn anything useful about the data. Another issue which can arise when using differentially private methods is that the promise of privacy will never be perfect. There is always some probability, even if it is quite small, that a person’s inclusion in a dataset may result in an output which reveals some information about them. The $\epsilon$ term gives us a way to quantify privacy, and some agents may be more sensitive to privacy than others and wish to be appropriately compensated for their data. It is common in research involving human experiments to offer compensation for participation, and we might think that we can simply adjust compensation accordingly with sensitivity. There’s a catch, however. In many cases, asking someone how much money they would need to be willing to answer a question is effectively equivalent to asking them the question itself. If responding “no” is benign and “yes” is embarrassing or incriminating, we can likely infer that anyone asking for a large sum of money before reporting their answer will say yes. This is known as the Sensitive Surveyor’s problem, and differential privacy offers a solution under certain conditions. As long as the maximum value any agent has for their privacy is bounded by some reasonable constant, a mechanism can be implemented which is differentially private with respect to both answers and valuations while maintaining reasonable accuracy with high probability; essentially, a survey conductor would not know how much each individual was paid, only the total cost of conducting the entire survey [9].

When the goal of a study is to accurately estimate attributes and behaviors in a population, it is imperative that the traits expressed in the data are aligned with their true underlying values. Privacy concerns of individuals need to be considered in any scenario where people need to opt in to a dataset, as well as any case where the data in question may cause harm if made public. Rational incentives for agents to lie poses a risk to the usefulness of social data, and differential privacy provides a possible path forward in an increasingly expanding number of scenarios.

Fairness

In 1973, UC Berkeley’s graduate school admissions came under a great deal of scrutiny when graduate admissions statistics revealed that overall, male applicants were offered admission at a significantly higher rate than female applicants. However, upon looking at individual departments, the vast majority showed unbiased statistics, and slightly more departments appeared to be biased in favor of women rather than men. It turned out that the gender balance of applicant pools varied drastically by department, and the apparent university-level bias could be primarily explained by the fact that men applied at higher rates to departments who admit a greater proportion of applicants [10]. This is a prominent instance of the statistical phenomenon of Simpson’s Paradox, which illustrates the difficulty of identifying the specific causes of observed discrepancies, and is highly relevant to questions about algorithmic fairness.

As algorithmic decision-making creeps its way into more areas of life with every passing year, there have been rising concerns that algorithms may be systematically biased to discriminate on factors based on factors such as race and gender. In theory, this can occur even if explicit information about race and gender is not even in the dataset; there may be a strong enough correlation with some other attribute such as location, education, arrest
record, or income to enable the same form of discrimination. One of the foremost challenges about evaluating the fairness of some statistical procedure is knowing what you’re looking for in the first place. Interestingly, for all of the fairness definitions we discuss, isolated solutions can be found using techniques from the differential privacy literature [11]. The mathematical formulations behind these mechanisms in fact show an intimate connection between the statistical problems of relying too much on one person’s data and assigning too much weight to one attribute in a dataset.

Several various definitions have been proposed as targets, each with their own drawbacks. Initially, a prevailing mentality is that an algorithm is fair if it “treats similar people similarly”. While this may seem reasonable, it lacks the ability to encapsulate group-level discrimination. Taken to the extreme, it promises almost nothing about discrimination along protected attributes; as evidenced by the Netflix data, when someone’s data contains a large number of attributes, very few data points are actually close together [12]. Other definitions include normalizing rates of false positives and false negatives in classification across protected groups, as well as the criterion that an algorithm should be equally successful at predicting some target value regardless of what protected category a person belongs to.

If we could simply gather enough criteria for fairness such that any reasonable person would find the combination of them reasonably fair, without giving up much predictive power, we could avoid a great deal of concern about algorithmic discrimination. Yet, an analysis of an algorithm called COMPAS used in criminal sentencing across the U.S., coupled with theoretical research, showed that this problem may not be easy to solve. Upon hearing criticism claiming that COMPAS was discriminating along racial lines when determining eligibility for early release of inmates, its makers sought to defend it. On a large sample of former inmates’ COMPAS scores and recidivism rates, they demonstrated that for both black and white defendants, the assigned scores resulted in identical probabilities that recidivism would occur. But while COMPAS may be able to satisfy the desire for predictive parity, it failed miserably with regard to false negative rates. Records for released black inmates who eventually reoffended made up a large portion portion of the dataset, and the algorithm had inadvertently trained itself to lock on to this trend. When examining the scores of defendants who ultimately did not reoffend, the typical scores for black defendants were far higher than those for white defendants [13]. As in the case with Simpson’s paradox, looking at data on different scales can result in a great deal of disagreement about whether a procedure is fair. Accompanying research provided a mathematical proof that simultaneously normalizing predictive parity and false negative rates while training an accurate algorithm is computationally intractable (unless P = NP).

While future research may uncover alternate models of fairness which achieve all relevant desiderata, or methods for sufficiently approximating the existing approaches for the two criteria in a feasible manner, for now we are at a loss. Whereas we had a uniform definition and feasible solution mechanism with privacy, the waters are less clear for fairness. If a computer can’t produce a perfectly fair classifier, we shouldn’t expect humans to be much better when relying on heuristics. Whether by algorithm or human, decisions about how to assign desirable resources such as college admission, a bank loan, or parole are made all the time. We should be skeptical of anyone believing that their approach is bias-free because they claim to “choose” to ignore a person’s protected attributes. But at least by solidifying which guarantees we can and cannot obtain, we can hope to proceed with further caution and minimize any undue harm as a result of algorithmic discrimination.

Adaptivity

False discovery is a pervasive issue in nearly every scientific discipline. There is a substantial gap between the traditional statistical guarantees for generalization and the way in which data analysis is performed in practice. Whereas traditional guarantees often require defining the entire scope of experimentation before beginning analysis, practical data science often
requires numerous rounds of adaptivity in order to improve a hypothesis or model based on prior results. When the statistical theory is abandoned, this can give rise to the dilemma of \( p \)-hacking, where the goal simply becomes finding a hypothesis with a “statistically significant” \( p \)-value on the dataset, rather than one which will generalize out of sample. This is exacerbated by the pressure in academia to publish significant results, and it can be difficult to distinguish between true discoveries and overfitting to the data when the methodology is cloudy. False discovery involving social data is particularly pernicious, as it can result in the adoption of policies or decision-making heuristics which are based on an untruthful heuristic and resulting in either ineffective or actively harmful. As we have seen from the Sensitive Surveyor issue, the process of collecting data may be expensive enough to prevent the possibility of simply going out and getting more data whenever we want.

Modern data science rarely consists of only asking a single query to a dataset and then discarding the data. Data is expensive, and it can be difficult or impossible to extract all relevant information from a dataset in a single round of analysis. In the context of a prediction problem using a machine learning-based approach where overfitting is common, data is separated into training and holdout sets, so that a model’s accuracy can be measured out of sample. Repeated analyses are necessary in this setting in order to compare and contrast different model techniques and hyperparameter values, and so an analyst may make many passes over their holdout set to inform their model choices, even if the model is not being trained on the holdout set directly.

In both the cases of hypothesis testing and model training, information from the holdout set can leak into the analysis process. Even if a data analyst is correctly isolating a holdout set and using proper cross-validation to select the parameters for their model, adaptively choosing queries based on the answers to prior queries can easily result in overfitting to the holdout set [6]. Randomly chosen holdout sets will inevitably have statistical properties which differ from the underlying data universe, and adaptively chosen sequences of queries can be led astray by these differences. There are existing tactics such as Bonferroni’s correction for avoiding issues of multiple hypothesis testing, but they often demand that the size of the dataset scales proportionally with the amount of testing iterations needed. In an analysis-heavy experiment where data is hard to come by, this may be infeasible.

Perhaps surprisingly, differential privacy can provide a solution. Recent work in adaptive data analysis has produced mechanisms which allow a much greater degree of “legal” holdout reuse – a number of tests which scales exponentially with the sample size – than any previous approach using the same mechanics as the previously discussed tools for privacy protection. By injecting appropriate noise into the answers to queries, guarantees can be derived proving that the values observed in testing are highly unlikely to deviate from the values in the entire universe of data by more than a small margin. Upon realizing that this can be done, we might actually see why it makes sense. The problem of overfitting occurs when a hypothesis or model focuses too much on specific points in a dataset and fails to capture the overall trends. Because differential privacy prevents us from learning too much about individual points, we are forced to focus on the big picture.

**Conclusion**

The solutions presented to problems concerning privacy, fairness, and adaptivity are still in their early stages. Much of the research literature on these topics has been. Apple has begun using differential privacy in analyzing user data. Loan providers and judicial institutions are consulting with researchers to evaluate the fairness of their predictive algorithms. Data science competition hosts are abundantly aware of the overfitting which can arise from repeated reuse, and are eager for solutions. As the scale at which data guides decisions in society inevitably expands, these issues will become increasingly pressing. At some point these techniques may be improved to a point where their downsides are negligible enough, their applicability broad enough, and their promises essential enough to warrant a paradigm
shift in the way universities, companies, and governments look at data about people. Until then, awareness that these issues are occurring is a first and necessary step toward finding better solutions.

References


