Trustworthy Machine Learning and Artificial Intelligence

How can we add the most important ingredient to our relationship with machine learning?

Kush R. Varshney, IBM Research AI

krvarshn@us.ibm.com

Decision making in high-stakes applications such as educational assessment, credit, employment, healthcare, and criminal justice is increasingly data-driven, and supported by machine learning models. Machine learning models are also enabling cyber-physical systems such as self-driving automobiles. These models either work to augment human abilities, or act fully autonomously. In all cases, they have a profound effect on our lives.

Advancements in the field of machine learning over the last few years have been nothing short of amazing. Nonetheless, even as these technologies become increasingly integrated into our lives, journalists, activists, and academics uncover characteristics that erode the trustworthiness of these systems.

For example, a machine learning model that supports judges in pretrial detention decisions was reported to be biased against black defendants. Similarly, a model supporting resume screening for employment at a large technology company was reported to be biased against women. Machine learning models for computer-aided diagnosis of disease from chest x-rays were shown to give importance to text contained in the image, rather than details of the patients' anatomy. Fatalities by self-driving cars have occurred in unusual confluences of conditions that the underlying machine learning algorithms had not been trained on.

In short, while each day brings a new story of a machine learning algorithm achieving superhuman performance on some task, these marvelous results are only in the average case. The reliability, safety, security, and transparency required for us to trust these algorithms in *all* cases remains elusive. As a result, there is growing popular will to have more fairness, interpretability, robustness, and provenance in these systems.

They say "history doesn't repeat itself, but it often rhymes." We have seen the current state of affairs many times before with technologies that were new to their age. The 2016 book, *Weapons of Math Destruction* by Cathy O'Neil, catalogues numerous examples of machine learning algorithms gone amok. In the conclusion, she places her work in the tradition of Progressive Era muckrakers Upton Sinclair and Ida Tarbell. Sinclair's classic 1906 book, *The Jungle,* tackled the processed food industry. It helped spur the passage of the Federal Meat Inspection Act and the Pure Food and Drug Act, which together regulated that all foods must be cleanly prepared and free from adulteration.

For computer scientists and engineers, the history and evolution of processed food from inside the industry may be more instructive. Henry J. Heinz was the progenitor of one of the largest food companies in the world today. In the 1870s, at a time when food companies were adulterating their products with wood fibers and other fillers, Heinz started selling horseradish, relishes and sauces made

of natural and organic ingredients. Heinz offered these products in transparent glass containers when others were using dark containers. His company innovated processes for sanitary food preparation, and was the first to offer factory tours that were open to the public. The Heinz company lobbied for the passage of the aforementioned Pure Food and Drug Act, which became the precursor to regulations for food labels and tamper-resistant packaging. These practices increased trust in, and adoption of the products. They provided Heinz a competitive advantage, but also advanced industry standards and benefitted society.

And now to the rhyme. What is the current state of machine learning, and how do we make it more trustworthy? What are the analogues to natural ingredients, sanitary preparation, and tamper-resistant packages? What are machine learning's transparent containers, factory tours, and food labels?

Let's begin by considering *supervised* machine learning – the most common form of machine learning, and the one behind many of the algorithms obtaining superhuman performance. The goal of supervised machine learning is to find a mathematical function that takes data points composed of numerical, ordinal, or categorical features as input, and predicts correct labels for those data points. For example, in credit, the features may be the income, education level and occupation of an applicant, and the label may be whether or not the applicant defaults on a given loan three years later. The algorithm finds the desired function by *training* on a large set of already labeled examples. And, although, the function is *fit* to the training data, it is applied to new unseen data points. It must, therefore, generalize to predict well on these new points.

In its basic form, the fitting is done to optimize a criterion such as average accuracy. To encourage generalization, the set of mathematical functions is restricted. Different machine learning algorithms, such as neural networks, decision trees, and support vector machines, have different restricted sets of functions. These are called hypothesis classes. A key assumption underlying these algorithms is that the training data points are – and all new data points will be – sampled independently from the same probability distribution. This is known as the independent and identically distributed (i.i.d.) assumption.

Returning to our food analogy, let us view the training data as the ingredients and ask if they are natural, untampered, and organic. The biggest problem in deploying machine learning systems is that the training data distribution does not always match the desired distribution.

Robustness to Data Set Shift

The probability distribution governing the data may change over time, resulting in the training data distribution drifting away from the data distribution. Or, it may be difficult to obtain sufficient labeled training data to correctly model the current data distribution. These situations are known as data set shift: it is as if we have ingredients that are not a reflection of the natural state of the world as it exists during deployment. Models trained on mismatched data usually have severely degraded prediction accuracy and lead to mistrust because promised performance numbers are not achieved when the systems are deployed.

Covariate shift	Distribution of features is different
Prior probability shift	Distribution of labels is different
Concept drift	Distribution of labels given features is different
Confounding shift	Distribution of labels given a variable that influences both features and labels is different.

Table 1. The different kinds of data set shift encountered in practice.

Protection from Data Poisoning

Data set shift is an inadvertent phenomenon. A deliberate data poisoning attack, however, can yield even worse performance degradation. This involves adversaries imperceptibly injecting just a few carefully-designed data points into the training data. A more sophisticated data poisoning attack involves creating a so-called backdoor. Data points are added to the training set so that the fitted predictor function will output the attacker's desired label for a given set of features. This label is not what would be expected from an unadulterated data set, and can be exploited maliciously. Fortunately, we have not witnessed such an event yet. But if it were to happen, an attack on a machine learning system via tampered data would cause distrust – not to mention damage – at a massive scale. In 1982, Tylenol poisoned with cyanide caused seven deaths in Chicago. As a result, the company's share in the over-the-counter painkiller market dropped from 35% to 8%. The potential for something similar happening to machine learning today is very real.

Fairness

The third factor that may dissipate people's trust in machine learning algorithms is bias. The desired training data distribution is not always the one that reflects reality if the present reality puts certain individuals at a systematic disadvantage. The sorts of biases and prejudices described at the beginning of the article—against blacks in criminal justice and women in employment—that are already present in decisions made by judges and hiring managers get reflected in training data and are subsequently baked into machine learning models. If practicing computer scientists have *the audacity of equality* – to borrow a phrase from the commentator Hasan Minhaj – then training data should not capture what is, but what could be. There are several mathematical definitions of algorithmic fairness for different notions of unequal decisions and outcomes among individuals and groups. Like other behaviors that are not aligned with the values of society, unfairness diminishes trust. Since algorithms are endowed with the values of their creators, striving for equality requires diverse deployment teams with broad sets of values to choose among attributes to be considered as features and labels, hypothesis classes, and fitness criteria.

Overcoming Mismatch Issues

The best way to overcome data set shift, data poisoning, and bias is to acquire and train on data that is matched to the desired distribution. If this is not possible – good, clean, labeled, i.i.d. data is a hard commodity to get hold of – then we can turn to the machine learning equivalent of sanitary (and sanitizing) food processing. Domain adaptation and transfer learning are a general category of

techniques that transform a training data set, or a model learned from it, to match a desired distribution. One way is to give different weights to different training data points. An extreme version of the weighting approach is discarding some data points altogether by setting their weights to zero. Many defenses against data poisoning attacks adopt this approach by discarding anomalous data points. Another way is by generating new i.i.d. data points from the desired distribution; we can use the given training set to help learn the generation process.

To perform domain adaptation and transfer learning, we need to know the desired distribution or its properties. When the desired distribution is uncertain, we can change the criteria by which we define the best fit of the predictor function. Robust formulations have criteria that yield predictors whose performance does not degrade severely in the face of distribution mismatch. As an example, we may choose to select a predictor function that maximizes the minimum accuracy across different distributions, rather than one which maximizes the average accuracy only for the given training data set. Adapting to data set shift has been a topic of machine learning research for many years, and advances continue to be made. Adversarial robustness and algorithmic fairness are currently hot research topics; our team at IBM Research has recently released open-source toolkits for both [1, 2].

Interpretability

A machine learning model may very well be accurate, reliable, fair, and robust but people won't trust it unless they have a way of knowing that this is true. Certain hypothesis classes admit predictor functions with millions of coefficients composed in complicated ways that do not allow people to understand how the labels are being predicted. Such models are known as black-boxes and are the equivalent of opaque ketchup bottles we cannot peer into. Deep neural networks are one example of black-box models. The counterpoint to black-box machine learning is *interpretable* machine learning. Interpretable hypothesis classes only contain simpler functions like decision trees and Boolean rule sets that people can more easily understand. Through this understanding, it can easily be determined whether the model has picked up on some quirks of the training data set that will not generalize robustly, whether the model contains any backdoors, and whether the model is explicitly and unwantedly discriminating against certain people.

If we have features that describe the prediction task well, interpretable functions tend to have equivalent, if not superior, accuracy to black-box functions. While doing so, their transparency imbues much more trust than black-boxes. Post-hoc explanation of black-box models, although currently a popular research topic and useful for other purposes, does not create trust in nearly the same capacity. The main idea behind post-hoc explanations is to first fit a black-box model to data, and then either fit simpler and more interpretable models around each individual data point, or perform sensitivity analysis of the black-box. The reason post-hoc explanations do not provide the same level of trust is because the predictions continue to be made by the black-box. Thus, explanations are sometimes inconsistent with the prediction, most often precisely where there is a lack of robustness.

System-Level Transparency

Interpretable machine learning provides transparency at the function level. But just as clear bottles are not enough for overall trust, system-level transparency, akin to factory tours and standardized food labels, is also needed for machine learning systems and services. Our team at IBM Research has recently proposed and demonstrated the use of *factsheets* for machine learning and artificial intelligence (AI) services that report the intended usage, ethical concerns, and development team in a standardized way. A factsheet also reports the lineage of the data sets and models used, the provenance of model development and training, results of accuracy and reliability tests, results of tests for robustness to data set shift and adversarial attacks, and results of fairness tests, all standardized. When voluntarily released by machine learning service providers, a factsheet may be called a supplier's declaration of conformity (SDoC), the technical term used in many industries and sectors for such documentation.

We envision the evolution of the SDoC for AI services to proceed as follows. A convening of corporations, standards bodies, and professional and civil society organizations will create standardized tests and testing protocols for AI services. Providers will then use these standards to populate and release SDoCs to be competitive in the market. Much of the SDoC computation will become automated as part of training, testing, and adaptation pipelines. The resulting living documentation will be automatically posted to distributed immutable ledgers. These can be supported by blockchain technologies, which are already being used for trust and transparency across the food supply chain, from growers to processors, and retailers. Then will come regulation, and an ecosystem of third-party testing and verification laboratories.

Social Good

The final element of trust is intended use, which is one of the items in the factsheet. Characteristics of trustworthy people include kindness, selflessness, and benevolence. Heinz was committed to the good of others and his ketchup's mission was to be a "blessed relief for mothers, and other women in the household." When machine learning is used to uplift humanity, a trust for the technology is further developed. A burgeoning AI-for-social-good movement has produced a portfolio of machine learning projects that help reduce poverty, hunger, inequality, injustice, ill health, and other causes of human suffering. Moreover, in partnering with non-profits, social enterprises, and international agencies to address specific problems, computer scientists and engineers encounter unique problem settings that naturally lead to machine learning innovations along the other elements of trust.

Conclusion

Machine learning is increasingly affecting our daily lives. In order to make it trustworthy, computer scientists need to consider measures beyond average predictive accuracy. Robustness to data set shift, robustness to poisoning, fairness, interpretability, end-to-end service-level provenance and transparency, and application for social good are not just condiments, but are essential to ushering in a world in which machine learning is a nonmaleficent and beneficent partner that humanity can count on.

References

[1] Adversarial Robustness Toolbox. Accessed January 18, 2019. https://developer.ibm.com/code/open/projects/adversarial-robustness-toolbox/.

[2] AI Fairness 360 Toolkit. Accessed January 18, 2019. http://aif360.mybluemix.net.

Author Biography

Kush R. Varshney is a principal research staff member and manager with IBM Research AI, working from the T. J. Watson Research Center, Yorktown Heights, NY. He is the founding co-director of the IBM Science for Social Good initiative. He conducts academic research on the theory and methods of statistical signal processing and machine learning, and applies data science to human capital management, healthcare, olfaction, computational creativity, public affairs, international development, and algorithmic fairness.