Limits to prediction: pre-read

SOC 555 / COS 598J, Princeton University, Spring 2024

Matthew J. Salganik and Arvind Narayanan

Is everything predictable given enough data and powerful algorithms? Researchers and companies have made many optimistic claims about the ability to predict phenomena ranging from crimes to earthquakes using data-driven, statistical methods. These claims are widely believed by the public and policy makers. However, even a cursory review of the literature reveals that state-of-the-art predictive accuracies fall well short of expectations.

This course aims to understand today's limited predictive abilities by synthesizing the existing literature and augmenting it with hands-on activities. This will help us understand the present and predict the future of prediction. More specifically, will limits to prediction melt away as datasets get bigger and computational abilities improve, or are we seeing fundamental practical limits that will remain for the foreseeable future?

These questions are interesting and important to social scientists, machine learning researchers, and policy makers, for many overlapping reasons. They pique our scientific interest and intellectual curiosity. They help us identify the types of problems or situations where machine learning techniques might be improved to provide better predictions. They guide policy makers on investing in AI research as a solution to thorny social problems. If we are entering a world where the future is predictable, we need to start preparing for the consequences, both good and bad. If, on the other hand, commercial claims are overhyped, we need the knowledge to push back effectively.

1. Preliminaries

The word prediction is often used loosely to refer to all applications of supervised machine learning. In contrast, our primary interest is in applications that involve predicting future events. The distinction is crucial: although deep neural networks have achieved breakthroughs in the last decade at many tasks such as "predicting" words in text, none of these tasks are true prediction problems, because they do not involve future events.

If we model a natural phenomenon as a process by which some input state is transformed into some output state, we can hope to learn the transformation function from past examples using machine learning. This simplified description immediately suggests at least three limits to prediction:

1. the possible nondeterminism of the universe (and, hence, phenomena of interest);
2. limits to measuring input/output states accurately and collecting sufficiently many training examples; these are highly dependent on the nature of the system
3. computational limits, whether hardware or algorithms.

The metaphysical question of the determinism of the universe is out of scope for this course. We will also assume that hardware and algorithms don't pose a serious limitation. We don't offer a fully principled

justification of this assumption but rather adopt it axiomatically. That enables us to focus our attention on what we subjectively consider to be more interesting research questions. In any event, when seeking to identify relatively hard limits, betting against Moore's law or against the ingenuity of the ML community seems unwise.

2. Why make predictions?

Predictive modeling has three main uses. It is essential to be clear about which goal we are pursuing in any given application. Failure to do so can lead to modeling approaches that don't let us effectively accomplish the goal and can lead to serious ethical lapses (see the ethics section below).

Adaptation. We might be interested in predicting an outcome in order to adapt to it or make decisions based on it. Predictive decision making has seen explosive use by companies and governments. Firms predict which candidate will do well if hired, banks predict who will repay a loan, and courts predict which defendants will be rearrested for a crime if released on bail. Predicting market trends can help investors mitigate risk. Predicting the path of a hurricane can help evacuate people and save lives.

Intervention. In public policy, the goal of predicting an outcome is often to change that outcome: making good outcomes more likely and bad ones less so. While we (usually) don't try to change the weather, we can try to forestall climate change, and that is one reason we try to predict it. If we can predict pandemics or disease outbreaks, we can try to prevent them. If we can predict population growth in a city, we can invest in traffic infrastructure. Unlike the previous category, we don't take the (probability of the) outcome as a given; rather, we seek to change it.

Scientific understanding. Ptolemy's model of the universe with the Earth at its center — once its parameters had been adequately tweaked based on observation — generated remarkably accurate predictions of the apparent movements of celestial objects, much like how machine learning makes predictions. It was used successfully for millennia for many purposes including navigation. Of course, it inhibited scientific progress.

These limitations of prediction have long been understood, so prediction per se has never been the primary goal of science, although prediction is an important way of validating theories or falsifying hypotheses. In the last decade, there has been interest in building models to maximize predictive accuracy and then interrogating those models as a way of generating scientific understanding, which has led to work on improving how causal explanation and prediction can best be integrated in social science.

3. Overview of evidence

In the first half of the course we will study a few prominent areas where researchers have built predictive models. We will start with the domain of geopolitical events (wars, regime change, …). Philip Tetlock, in a notable 20-year study of forecasting experts, took a critical look at how good they actually are at doing the thing they claim to be experts at, and compared them to statistical models. The findings were eye opening. This module will also serve as a foundation for learning how to think about making and evaluating predictions. The first assignment is related to predicting geopolitical events.

Another domain of importance is life trajectories. Some of the best evidence for limits to prediction comes from the Fragile Families Challenge, a mass collaboration organized by a Princeton team led by Matt, using a dataset resulting from a 20-year study of thousands of children and their families. The headline is that the best machine learning models were little better than a 4-variable logistic regression model, but the project has yielded and continues to yield many rich insights about predictability.

We will also study cultural products (i.e. predicting which books will be bestsellers or which movies will be blockbusters or even which social media posts will go viral), a domain where effective prediction would be extremely commercially valuable. It is scientifically interesting because it shows temporal dynamics unlike most of the other domains that we will study.

In these and other domains, attempts to predict social phenomena have almost always ended in disappointment. One notable exception we saw in the previous edition of this course was a line of papers claiming to predict civil war highly accurately, but we discovered that this breakthrough was illusory. Predictive decision making algorithms are employed in many areas, including criminal justice and hiring, which may lead you to assume that they must be highly accurate. The evidence says the opposite. For example, a state-of-the-art criminal risk prediction algorithm has an AUC of about 65% (random guessing is 50%).

Could accuracy be much better if we could access orders of magnitude more data? It is possible, but we will also identify many mechanisms that could generate limits to prediction that won't be overcome by improving data size or quality. In the previous version of the course, we had hoped to find ways to quantify such "irreducible error", but didn't succeed. This time we will look at this question qualitatively.

We will contrast this with weather, which is a physical system. As such, weather prediction has seen steady increases in accuracy over the decades. But there are also fundamental limits to prediction which are quantifiable and well understood. Weather is also interesting because the primary approach to prediction has historically been simulation rather than machine learning, although machine learning has recently been making inroads. We will try to understand the relative merits and limitations of the two approaches.

The final domain is existential risk. We are primarily concerned with existential risk from advanced AI, but we will briefly discuss other types of existential risk as well. This is utterly unlike the other domains because extinction is by definition a one-time event. It is a topic of keen interest to the AI community, to policy makers, and to the general public. There is little consensus among those who study the topic about methods or conclusions; in fact, it is notable for the acrimonious nature of the debate. Our discussion will necessarily be heavily shaped by the instructors' personal perspectives.

4. Mechanisms that generate limits to prediction

We now outline several concrete mechanisms that give rise to limits to prediction and which we expect will not be fully overcome with more and better data, algorithms, or hardware. Understanding these mechanisms can help anticipate how the achievable predictive accuracy will change over time in a given domain.

Sensitive dependence on inputs: A butterfly's wings, according to an aphorism, can trigger a tornado. Weather is notoriously a system in which arbitrarily small divergences in initial conditions tend to amplify over time. Thus, any fixed limit to the resolution of measurements implies a limit to predictive accuracy that gets more severe the farther out one wants to forecast.

Accumulation and amplification of advantage: In markets for some cultural products such as books, movies, or music, success can lead to increased attention, which can lead to more success. This process means that small differences in initial success, even ones that were essentially random can be magnified over time, which makes prediction difficult. A similar process can also happen in reverse, whereby failure can lead to more failure. For example, a person can be evicted from their home, which could cause them to lose their job, which could lead to substance abuse and other problems. This accumulation of disadvantage can magnify small differences or random fluctuations.

Strategic behavior: In the stock market, intelligent agents aim to incorporate all available information to act to maximize their profits, which has the effect of making it difficult to predict the future movement of stock prices. Many other systems may have a similar quality. For example, it has been argued that if an armed conflict can be seen coming, one of the sides will have the incentive to take steps to avoid it. Strategic behavior by political candidates, such as changing one's platform to appeal to a broader swath of voters, could make elections difficult to predict.

The previous three mechanisms can be seen as properties of the system we are trying to predict. The next three can be seen as properties of our ability to observe the system.

Unobserved or unobservable inputs: One reason it's difficult to predict who will get evicted is that landlords vary considerably in how aggressively they will attempt to evict tenants. Thus, a dataset that tracked tenants thoroughly but not landlords will be limited in effectiveness. Perhaps surveillance of people's activities will one day become so comprehensive that companies or governments will not be limited by unobserved attributes. The present reality, however, is that relevant attributes are often unavailable for prediction.

The 8 billion problem: Unobserved inputs are missing columns in a dataset. There is also the possibility that our data has too few rows, i.e., training samples. The more complex the phenomenon we are trying to model and predict, the more samples we need. Even as computing power and storage plummet in cost, we are fundamentally limited by the number of training instances that the real world can furnish us. Also, we are limited by the fact that in social settings, the mapping between the inputs and the outcomes might vary across societies.

Distribution shift: No real-world system or phenomenon is perfectly static over time. Yet the use of machine learning for prediction involves learning a relationship from past observations and applying it to future observations. This is not a problem if the task is to predict, say, which stars will go supernova, because the phenomenon does not change at human timescales. But for most problems of interest, the joint distribution of the predictors and the target changes fast enough to pose challenges. For example, influenza models built using pre-2020 data may need to be adjusted because people's response to epidemics has been profoundly altered due to the experience of covid-19. This type of temporal distribution shift is unavoidable. In addition, in practice, there are usually additional types of distribution shift to contend with, such as geographic: a model may be built using data from in one area but then deployed in others.

Contrast all this with a domain where there isn't a significant limit to predictability: image classification (say, cats vs. dogs). The task doesn't involve predicting the future or strategic agents, so the first three items in the list above don't apply. There are no unobserved inputs in the sense that the image contains all the information we need for correct classification. And we can find an effectively endless supply of training data online.

5. Measuring accuracy is surprisingly subjective

What makes a prediction "good"? This turns out to be a fundamentally hard question to answer; there is no one single best way to evaluate predictions. But we can identify three important criteria:

- *Consistency* measures the correspondence between the forecaster's beliefs and the predictions. This might seem like a low bar, but sometimes people make forecasts that differ from their true beliefs if they are worried about being harshly penalized for certain kinds of errors. Interestingly, something like this can happen even with algorithmic predictions. Designing the scoring function or loss function turns out to be non-trivial and essential for achieving consistency.
- *Quality* measures the correspondence between predictions and outcomes. Quality (model performance in machine learning) is especially hard to measure. That's because we usually try to collapse the correspondence between the predictions and outcomes into a single number. Unsurprisingly, this number rarely tells us everything we want to know, and the best-performing model may depend on the choice of scoring function (such as R^2, AUC, RMSE, cross entropy, etc.) Although quality is often the only thing people measure, it is not the only criterion, nor even the most important in many cases.
- *Value* measures the incremental benefit of forecasts to users. For example, if we are using forecasts to adapt to the world or intervene, then we should try to evaluate not just how close the forecast was to events, but also how much some critical outcome was improved based on this prediction.

The subjectivity of evaluation means that straightforward numerical comparisons between domains are meaningless due intrinsic differences between domains and degrees of freedom in problem and task formulation. To make this more concrete, suppose we find that the success of books is more predictable than that of movies. Can we conclude that book publishing is more meritocratic than the movie industry? No, because there are many other possible explanations, such as:

- current methods fall well short of the actual limits to predictability.
- our finding is reversed if we change some seemingly trivial details of the formulations of the two prediction tasks
- there is different information routinely collected about books and movies before launch and so what appears to be a difference between these two products is actually a difference in the data routinely available about these products
- there are many inherent differences between books and movies, such as the fact that movies earn much of their revenue in a single season and thus face a greater variance in the competitive landscape.

In the previous edition of this course, we'd hoped to make statements about relative predictability between domains. Arvind has since given up on this goal, while Matt continues to hope that it will become worthwhile to pursue. At any rate, it is not a major theme of this edition.

6. Measuring accuracy: pitfalls

There is a long list of pitfalls in machine learning that may lead us to biased estimates — usually overestimates — of predictive accuracy. Some but not all of these pitfalls are well known. Published research often falls into these traps and industrial applications even more so. Here we will briefly review some of them.

The pervasiveness of these errors may explain some of the unfounded optimism about the capabilities of machine learning.

We briefly review a few major pitfalls. Many of these are elaborated in David J. Hand's paper "Classifier Technology and the Illusion of Progress" that we will read in week 1.

- Problem uncertainty: there may be inherent arbitrariness in the class definition (in the hiring context, who is a good employee?), or the way we define the task may not faithfully capture what we have in mind (we may use performance reviews to measure employee productivity).
- Errors in class labels: even if classes can be clearly defined and labeled in theory, real-world data usually has errors.
- Researcher degrees of freedom in task formulation: in a typical machine learning problem, there is a wide array of choices necessary to concretely formulate the task. (Example: image dimensions for an object recognition problem.) These choices greatly affect the accuracy that can be achieved.
- Overfitting: textbook machine learning includes techniques to avoid overfitting to small training sets, but there are more subtle types of overfitting that are harder to avoid, including "human-in-the-loop overfitting".
- Leakage: this refers to a spurious relationship between the features (independent variables) and the target (dependent variable) that is an artifact of the data collection, sampling, or pre-processing strategy. In an apocryphal story from the early days of computer vision, a classifier was trained to discriminate between images of Russian and American tanks with seemingly high accuracy, but it turned out that this was only because the Russian tanks had been photographed on a cloudy day and the American ones on a sunny day.
- Drift: the statistical relationship between the features and the target may change over time; this is particularly salient to us given our focus on predicting the future.
- Demographic biases: human societies consist of subpopulations (e.g. ethnic groups) that differ in the distributions of predictor and target variables, often a continuing effect of historical prejudice. Machine learning tends to perform better for the majority group than minority groups for many reasons including the availability of a greater number of training instances. Aggregated performance metrics often hide disparities in performance that lead to unfair decision-making systems.
- Selective labels: this is an insidious type of sample bias in which the ability to observe an instance is correlated with the outcome we're trying to predict. For example, in a college admissions context, if we want to use the performance of past students to learn to predict whether an applicant will succeed (e.g. earn a high GPA) if admitted, we may be limited to a training set that is already filtered based on attributes thought to correlate with college success.
- Other problem-specific sample biases: in addition to the biases above that tend to recur across domains, in most problems there are other idiosyncratic sample biases.
- Acting on predictions changes the outcome: the goal of prediction is often to make a decision. But that decision may in turn impact the outcome. For example, a bank may set a loan interest rate based on the predicted risk that the borrower will default, but a higher interest rate makes a default more likely. This creates a self-fulfilling prophecy. Prophecies may also be self-defeating: see "strategic behavior" above.

Awareness of these pitfalls and difficulties will inform our approach to the readings and provide a natural opportunity for original student research.

7. Ethics

There are two categories of ethical concerns that recur in predictive decision making systems. The first relates to how they are built, and the second is about whether they should be deployed at all. Privacy and bias concerns belong in the first category. Machine learning requires surveillance in order to collect training data. The surveillance and centralization of data concentrates power the hands of often non-transparent and unaccountable entities. And when training data reflects societal biases, stereotypes and prejudices, those biases are propagated to the decisions made by the system, perpetuating cycles of inequality.

A failure to recognize limits to prediction and limits *of* prediction means that prediction is often deployed where it is not ethically appropriate. One example is denying criminal defendants their freedom based on a prediction of their future behavior (recidivism or failing to appear at trial) that is only slightly more accurate than a coin flip. The effect is to punish people for crimes they have not committed. More importantly, it turns out that many cases of failure to appear have benign explanations such as needing to care for a child. A more humane approach would be to find opportunities to reduce failures to appear, e.g. by having the state provide childcare services to defendants and thus avoid the need for prediction altogether.

Understanding limits to prediction allows us to gain a more nuanced understanding of different ways to design predictive systems — if a predictive system is indeed what we want — and the tradeoffs involved. One long-running debate concerns the pros and cons of human judgment versus machine predictions. But we will see that there is a third option: simple hand-computable statistical formulas with just a few predictor variables. Empirically, these approaches are almost as accurate as black-box machine learning systems in many domains and avoid many (but not all) of their drawbacks.

8. A few other things to know about the course

An unusual feature of our course is the diversity of domains and disciplines that we draw our readings from. Most of these papers are not presented in terms of limits to prediction and their authors may be surprised to be included in this list. This admixture is both an opportunity and a challenge. In the second half of the course, there will be fewer readings and more focus on original activities.

In the first week we will learn about the "two cultures" of statistical modeling (broadly, statistics and machine learning). This distinction is not just of intellectual interest: it is directly applicable to your interactions in the class. Most of you have been socialized into one or the other culture. These are two profoundly different ways of thinking about the same problems. They differ in terms of which questions are valued, what constitutes rigorous evidence, and other cultural assumptions, even when the technical tools being used are identical (say, logistic regression). We hope that students from the two cultures will partner with and learn from each other, despite its frustrations. The two of us come from different cultures and have benefited greatly from learning the other culture.

A grad seminar is an opportunity for research, especially in a course that is pushing the boundaries of knowledge. The earlier iteration of the course led to published research and we hope that this one will, too. We hope you will pick your final project with an eye toward producing original research. After the course is over, we (Arvind and/or Matt) may, at our discretion, offer to help you develop your project into a paper to submit for peer review. You are welcome to work with us to produce a publishable paper, or do so on your own, or choose to leave it as a class project. It is completely up to you.