Limits to prediction: pre-read

COS 597E / SOC 555, Princeton University, Fall 2020

Dear students,

This document aims to give you a rough summary of our thoughts on this topic before we teach the class. There are points where we are not as precise as we would like to be, and there are places where the two of us don't agree. This is to be expected for a topic that's at the cutting edge of research, and we will be learning through this class along with you. Still, we hope this document gives you a sense of the goals and motivations of the class.

Arvind Narayanan and Matt Salganik
September 1, 2020

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 object recognition, 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. Hypotheses

We now outline several concrete hypotheses for why limits to prediction arise. These are purely illustrative and not exhaustive, nor mutually exclusive. Our course goals will include testing some of these hypotheses, generating new hypotheses, and understanding how the nature of the system gives rise to these limits.

H1: 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.

H2: Shocks

Life trajectories are also sometimes upended by the kinds of inputs that seem likely to remain unmeasurable for the foreseeable future: a lottery jackpot; an accident; a crime of passion committed in the heat of the moment; a college admission for which one just made the cut. What is unclear is how common these are in the typical life course and to what extent they limit predictability.

H3: Accumulation and amplification of advantage

There is a particular kind of sensitive dependence that is ubiquitous in society and is worth discussing separately: when success breeds further success. For example, in the 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.

H4: 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.

Going further, imagine a premeditated crime, but one where the plans existed entirely in the minds of the perpetrator(s). As long as people's thoughts remain inaccessible to predictive algorithms, that will impose limits to the predictability of some types of events.

H5: 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.

H6: Drift

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 -- drifts -- fast enough that predictive algorithms must explicitly account for it. When we build algorithms or models that explicitly account for drift, we call it a forecasting. In other words, forecasting is one approach to prediction problems. Note that while we adopt this terminology for our course, it is not universal.

Drift is a notorious problem in applications such as epidemic forecasting. 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.

H7: Ill-conceived target variable

Some target variables will be hard to predict because they are poorly measured or unstable. For example, imagine trying to predict a student's performance on a math test. As the number of questions on the test increases, the measurement of math performance should happen with less error (assuming that it is a well-designed test). However, we also hypothesize that there will be diminishing returns to improved measurement and that predictive performance will plateau well below perfection.

H8: Self-equilibration and strategic behavior

In some systems, limits to prediction arise because of strategic behavior of participants. 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.

H9: Forward and inverse "prediction"

Finally, here's a heuristic to recognize tasks for which there may *not* be a strong limit to predictive accuracy. As we noted above, many so-called prediction tasks aren't actually about predicting the future. Instead of thinking in terms of time, let's imagine data-generating processes in nature. For a physical system such as weather, the data generating process is simply the evolving state of the system: the vector of observed variables at time t+1 is a function of the vector at time t plus noise. Social systems can also be thought of roughly similarly. Since noise accumulates over time, it's harder and harder to predict states further out from observed states.

For a task such as distinguishing between images of cats and dogs, what's the "data generating process"? There's no correct answer, but here's an arguably useful one. Each species corresponds to a probability distribution over genotypes that is relatively stable over time, and nature "samples" individuals from this distribution. Then, there is a well-studied stochastic process by which genotypes are transformed into phenotypes. Finally, photography transforms 3-dimensional beings into 2-dimensional images. The "data generating" process is thus the composition of these three functions.

The point of this seemingly convoluted exercise is simply: recognizing species from images is an "inverse" prediction problem. Given the output of the data generating process, the task is to predict the input. Since noise tends to accumulate in the forward direction, inverse prediction problems tend to be easier than forward problems. For this particular task, there are many other ways to arrive at the same conclusion, such as the fact that humans can easily tell cats from dogs. But this heuristic can be applied to many other tasks for it is not intuitively obvious whether there are strong limits to prediction.

3. Quantifying predictability: pitfalls and opportunities

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" and nonindependence of training samples.
- Drift: the statistical relationship between the input variables 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.

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

We have a particular interest in scoring functions — that is, how we can measure predictability. A predictive model maps each point in the input space to a probability distribution over outputs. It is a multidimensional beast. Yet we measure predictive performance by collapsing the comparison between the model and the test data (or distribution) into a single number. Unsurprisingly, this number rarely tells us everything we want to know about performance, and the best-performing model may depend on the choice of scoring function (such as R^2, AUC, RMSE, cross entropy, etc.)

We hope that the breadth of readings will help us understand the pros and cons of different scoring functions. We will start by identifying important properties of scoring functions such as absolute vs. relative scoring and whether it is a proper scoring rule. Then we will determine which functions satisfy which properties, and discuss which properties are important in which domains.

4. Understanding the domains

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.

To be sure, domain understanding is important for understanding the papers and their findings. Let us share an anecdote based on personal experience. When initially discussing the possibility of this course, Matt expressed his surprise to Arvind about the low R^2 values of the best-performing models in the Fragile Families Challenge. But Arvind was confused by this since he lacked any a-priori expectation of what the R^2 values "should" be. If we have better predictive models than we did before, shouldn't the results be considered a success? Is improved predictive performance even part of scientific success? A normative, domain-specific understanding of the purpose and context of prediction is important for expressing judgment about whether a given level of accuracy is good or bad; interesting or dull; important or meaningless. Many, many nuances about prediction are domain specific.

Neither of us is an expert in all the domains covered in the course. However, we opted for this cross-domain approach because we think it is the way forward to new insight about the limits of prediction; no field has yet developed a complete approach to thinking about these problems. One way that we have tried to make the cross-domain approach more likely to be successful is diversity of thought and experience. The course is taught by faculty members from two different fields (computer science and sociology), and we hope that we will have students from different fields. Learning together in an interdisciplinary class can itself be difficult, but we think this approach increases our chances from successful cross-domain comparisons. Students who have more domain expertise are expected to share it, and students that have less domain expertise are expected to listen generously.

Cross-domain comparisons can be difficult for other reasons as well. For example, 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? Not directly, 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.

Although it seems hopeless in light of these limitations, we will cautiously undertake to develop heuristics that may enable us to make statements about relative predictability between domains. We also seek to compare and contrast theories of unpredictability across domains.

When selecting domains for this course we had a preference for domains where there was research both measuring the limits to prediction empirically and developing theories that help explain those limits. We also had a preference for domains about which we ourselves had some expertise, and where we thought there was fertile ground for new research.

5. Implications

Now we turn to several high-level questions which we hope to better understand through the empirical knowledge gained in this course.

Implications of high and low accuracy

There are many tasks where there has been great progress in predictive accuracy, either gradually over decades, as in the case of weather prediction, or relatively rapidly, as in the case of object recognition in the early 2010s (the latter is, of course, not a true prediction task). When accuracy increases sufficiently, then for some (but not all) problems the predictions become practically useful and the implications tend to be profound. This is the case even when predictive accuracy remains low in an absolute sense, with a good example being the pervasiveness of targeted advertising online. Thus, "predicting the future of prediction" is an important skill for anticipating and preparing for upcoming disruptions to specific industries, scientific fields, or society as a whole. A pitfall that's just as common as failing to anticipate advances is to overreact by assuming that a breakthrough is just around the corner.

The implications of improvements in accuracy may not always be positive. In particular, machine learning is often used to predict or infer things about people that they may not want others to know. This is bad news for privacy. Computer vision furnishes many examples, such as Clearview AI, a commercial product that dredges up people's long-forgotten photos online based on facial recognition. Even when predictive accuracy is low, such technology may be deeply problematic, such as "predicting" sexual orientation from facial images.

When is prediction the right question?

Predictive models are overused in computer science and industry because they are convenient to apply. Often, what is framed as a prediction problem can be better understood as a problem of explanation, intervention, or decision making.

Explanation is about generating scientific insight into how a process works rather than simply predicting its input-output behavior. For example, 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. Many modern applications of prediction are arguably similarly shortsighted.

Intervention is about figuring out how to change a process for the better rather than treating it as a given and confining oneself to making predictions. A poignant example: in response to the problem that some defendants released on bail won't show up to court on their trial date, risk prediction systems are in widespread use today to determine who should be released and who should be held until their trial. These are statistical algorithms that use factors such as age and income to predict the probability of failing to appear at trial. The effect is to punish people for crimes they have not committed. Further, these systems have a disproportionate impact on racial minorities.

It turns out that many cases of failure to appear have benign explanations such as needing to care for a child. The prediction approach is thus often criticized as needlessly cruel; an intervention approach would

seek 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.

Finally, the science of **decision making** recognizes that many considerations go into making good decisions beyond maximizing predictive accuracy, especially because the decisions themselves have causal effects. For example, recommender systems are often built as predictive systems that maximize the probability of a click (or some other metric). Unfortunately, this formulation ignores the fact that recommending some types of items to users changes their interests and preferences over time. Failing to model these consequences can exacerbate filter bubbles, political polarization, and other harmful effects on social media.

How to design predictive systems

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, in domains including criminal justice, social services, and employment. 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.

A second debate concerns the importance of domain expertise in designing predictive systems. Through the course, we hope to develop a better understanding of how domain experts should contribute to building prediction systems, interpreting their outputs, and evaluating their performance.

Many of the debates referenced in this section speak to the Fairness, Accountability, Transparency, and Ethics (FATE) of predictive systems. While there is a robust and ongoing FATE critique of machine learning, that critique has rarely contested the predictive analytics industry's claim that machine learning methods are delivering great improvements in accuracy compared to human experts and traditional statistics. Questioning that assumption changes the debate completely.