Predictive Modeling and Scientific Machine Learning

Contents

Reports that say that something hasn't happened are always interesting to me, because as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns—the ones we don't know we don't know. And if one looks throughout the history of our country and other free countries, it is the latter category that tend to be the difficult ones. Donald Rumsfeld, United States Secretary of Defense, DoD news briefing, February 12, 2002.

Machine learning

Machine learning is about making models out of data. One of these model-building tasks we will see in this class is *regression*. In regression you are given, for example, the characteristics of a house and you have to predict its price. You gather some data of house characteristics and the corresponding prices. You make a mathematical model that connects the characteristics to the prices and, finally, you fit the parameters of the model by minimizing the prediction error. Now if someone gives you a new set of building characteristics, you can predict the price.

Causality

In simple regression, there is nothing special that distinguishes the price from the house characteristics. You can also make a model that tells you what are the house characteristics from the price. But, there is a big difference between the former model and this one. The first model (house characteristics to price) is a *causal model*, i.e., a model that follows the causal mechanisms (better materials, better design, more rooms, etc. cost more). The second model (price to house characteristics) just captures *statistical correlation*. You cannot rennovate your house simply by increasing its price! You cannot even *intervene* to increase its price. It is not sufficient to ask for a higher price. There must be someone who wants to pay it!

Without causality, machine learning is useless. We must always strive to make our models *causal*. A *causal model* is a model that attempts to capture the mechanisms that govern a given phenomenon. We will use the language of *structural causal models* (SCM), developed by the computer scientist Judea Pearl, to formalize the concept. A structural causal model is a collection of three things:

- A set of variables. These are variables that our model is trying to explain (endogenous), but also other variables that may just be needed (exogenous).
- A set of functions that give values to each variable based on the values of all other variables.

Example: Asthma model (J. Pearl)

Suppose that we are trying to study the causal relationships between a treatment X and lung function Y for individuals who suffer from asthma. However, it is plausible that Y also depends on the air pollution levels Z. The final ingredient is the set of function that connects X and Z to Y:

$$Y = f(X, Z).$$

Graphical representation of causal models

Graphical representation of causal models

Every SCM is corresponds to a *graphical causal model*. These are usually *directed acyclic graphs* (DAGs). These can be read trivially from the SCM form. Let's look at an example.

Example: Asthma model - Graphical causal model

Here I am representing each variable with a node. The node at the beginning on an arrow is the direct cause of the node at the end of the arrow.

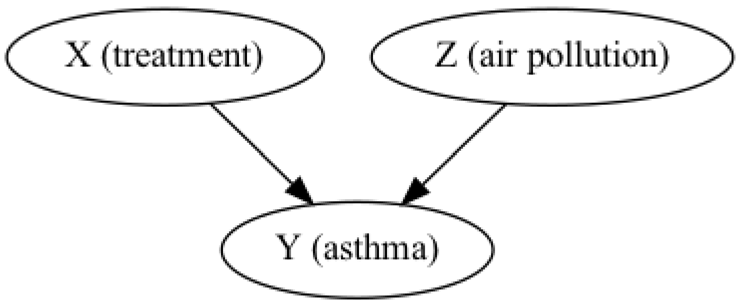

Fig. 1 An example of a causal graph.

Predictive Modeling

What if we have a good, causal model, but it requires inputs that are *unknown* or *random*? Then, it is clearly impossible to be certain about the model predictions. *Predictive modeling* is all about dealing with such *uncertainties*.

Some examples of uncertain model inputs are:

- the value of a model parameter,
- the initial conditions of an ordinary differential equations,
- the boundary conditions of a partial differential equation,
- the value of an experimental measurment we are about to perform, and even
- the mathematical form of a model.

There are two kinds of uncertainties. If something is truly random, then we say that it corresponds to *aleatory uncertainty*. If something is just unknown, then we say it corresponds to *epistemic uncertainty*. In other words:

- Aleatory uncertainty is associated with inherent system randomness.
- Epistemic uncertainty is associated with lack of knowledge.

Epistemic uncertainty can be reduced by collecting more data, improving measurement instruments, and thinking… Aleatory uncertainty cannot be reduced. There is a long philosophical debate about the distinction between aleatory and epistemic uncertainties. Sometimes, it is hard to decide if you have one kind or the other. But, we do not care. Probability theory are sufficient to describe both uncertainties. This is the stance we are going to take in this course. So, to make predictive models we will need to formulate our models in a probabilistic way.

Scientific Machine Learning

Scientific machine learning is machine learning for scientific and engineering applications. Of course, models have to be causal. Typically, scientific and engineering problems have quite a bit of uncertainty. So, ideally we want to wrap things in a probabilistic framework and make predictive models.

Finally, when dealing with scientific and engineering problems, we know quite a few things on top of the data. For example, we may know that the temperature satisfies a differential equation, or that the mass is conserved, and so on. If we use this physical information, we can make models that can extrapolate beyond the data we have seen. If we do use such information, we may say that we are doing *physics-informed* machine learning.

By Ilias Bilionis (ibilion[at]purdue.edu)