MATH 342W / 650.4 / RM742 Spring 2024 HW #1

Benjamin Minkin

Monday 12th February, 2024

Problem 1

These are questions about Silver's book, the introduction and chapter 1.

(a) [easy] What is the difference between *predict* and *forecast*? Are these two terms used interchangably today?

Silver distinguishes between predict and forecast. In his view, a prediction is a guess about how the future will unfold. This can either be true or false. For example, saying it will rain tomorrow. He describes a forecast as a probabilistic range of predictions. A forecast cannot be true or false but is instead either accurate or inaccurate. An example of a forecast is claiming there is a 30% chance of rain tomorrow.

(b) [easy] What is John P. Ioannidis's findings and what are its implications?

He published a paper claiming that the majority of conclusions from scientific papers are wrong. This has broad reaching implications as scientific studies dictate what is considered the so called "truth."

(c) [easy] What are the human being's most powerful defense (according to Silver)? Answer using the language from class.

Silver believes that one should predict using a Bayesian approach. This involves incrementally improving his prior with every new data point observed.

(d) [easy] Information is increasing at a rapid pace, but what is not increasing?

The theory required to turn that data into useful insights.

(e) [difficult] Silver admits that we will always be subjectively biased when making predictions. However, he believes there is an objective truth. In class, how did we describe the objective truth? Answer using notation from class i.e. $t, f, g, h^*, \delta, \epsilon, t, z_1, \ldots, z_t, \delta, \mathbb{D}, \mathcal{H}, \mathcal{A}, \mathcal{X}, \mathcal{Y}, X, y, n, p,\ x_{\cdot 1}, \ldots, x_{\cdot p}, x_{1\cdot}, \ldots, x_{n\cdot}$, etc.

In our lecture's notation, Y, $\mathcal{Y}$,the formula t, the parameters $z_1..z_n$, and δ were all components of the objective truth.

(f) [easy] In a nutshell, what is Karl Popper's (a famous philosopher of science) definition of *science*?

Popper defines science as any claim that can be proved either true or false. Unfalsifiable claims are inherently unscientific.

(g) [harder] Why did the ratings agencies say the probability of a CDO defaulting was 0.12% instead of the 28% that actually occured? Answer using concepts from class.

They calculated based on the mortgages being $\overset{iid}{\sim}$ R.V.'s. This was foolish as there are economic conditions that effect all homeowners, and makes the mortgages at least partly dependant.

(h) [easy] What is the difference between *risk* and *uncertainty* according to Silver's definitions?

He defines risk as a calculable amount of loss that can occur. For example, if you bet a single number in roulette you have a $\frac{37}{38}$ (or $\frac{38}{39}$) or losing your bet. That chance is your risk of losing your bet. Uncertainty is the inability to calculate risk accurately. This is akin to betting a number on a roulette wheel with an unknown amount of numbers. There may be less or more risk depending on if there are fewer or more total numbers. This is uncertainty.

(i) [difficult] How does Silver define *out of sample*? Answer using notation from class i.e. $t, f, g, h^*, \delta, \epsilon, z_1, \ldots, z_t, \delta, \mathbb{D}, \mathcal{H}, \mathcal{A}, \mathcal{X}, \mathcal{Y}, X, y, n, p, x_{\cdot 1}, \ldots, x_{\cdot p}, x_{1\cdot}, \ldots, x_{n\cdot}$, etc. WARNING: Silver defines *out of sample* completely differently than the literature, than practitioners in industry and how we will define it in class in a month or so. We will explore what he is talking about in class in the future and we will term this concept differently, using the more widely accepted terminology. So please forget the phrase *out of sample* for now as we will introduce it later in class as something else. There will be other such terms in his book and I will provide this disclaimer at these appropriate times.

He uses out of sample to describe a model trained off data in a unique time period. When the behavior regresses towards the mean it is seemingly out of sample.

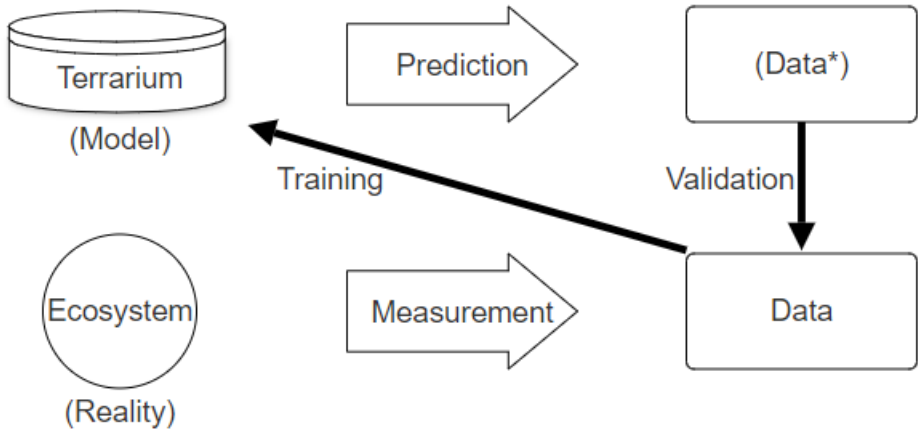
(j) [harder] Look up *bias* and *variance* online or in a statistics textbook. Connect these concepts to Silver's terms *accuracy* and *precision*. This is another example of Silver using non-standard terminology.

Silver defines precision as having low variance. A cluster of guesses can be dead wrong but if they are a close grouping, they are precise. He defines accurate as having low bias. As long as the average over time converges to the truth, the predictions are accurate. This occurs even when the guesses are very different from each other.

Problem 2

Below are some questions about the theory of modeling.

(a) [easy] Redraw the illustration of Earth and the table-top globe except do not use the Earth and a table-top globe as examples (use another example). The quadrants are connected with arrows. Label these arrows appropriately.



(b) [easy] Pursuant to the fix in the previous question, how do we define *data* for the purposes of this class?

Data is the result of both measuring reality and the results of simulations from models.

(c) [easy] Pursuant to the fix in the previous question, how do we define *predictions* for the purposes of this class?

Predictions are the output of the model that leads to gathering data.

(d) [easy] Why are "all models wrong"? We are quoting the famous statisticians George Box and Norman Draper here.

All models are by definition wrong. In order to be a model, there must be simplifications and assumptions made.

(e) [harder] Why are "[some models] useful"? We are quoting the famous statisticians George Box and Norman Draper here.

Using an accurate approximation of reality can help us better understand the world and adapt to it. You do not need to have a perfect model to gain any insight.

(f) [harder] What is the difference between a "good model" and a "bad model"?

A good model has practical uses. There are no bad models as they can be used as an example of what not to do.

Problem 3

We are now going to investigate the famous English aphorism "an apple a day keeps the doctor away" as a model. We will use this as springboard to ask more questions about the framework of modeling we introduced in this class.

(a) [easy] Is this a mathematical model? Yes / no and why.

Yes. It has the recipe with ingredients and measurements. One apple per day eliminates the need for a primary care physician.

(b) [easy] What is(are) the input(s) in this model?

$$\frac{1_{apple}}{1_{day}}$$

(c) [easy] What is(are) the output(s) in this model?

Spending no time with doctors.

(d) [harder] How good / bad do you think this model is and why?

I think this model is bad because it is overly simplistic. One apple per day is not a healthy diet.

(e) [easy] Devise a metric for gauging the main input. Call this x_1 going forward.

let:

$$x_1 = \frac{apples}{day}$$

(f) [easy] Devise a metric for gauging the main output. Call this y going forward.

let:

$$y_1 = \frac{appointments}{year}$$

(g) [easy] What is $\mathcal{Y}$ mathematically?

$$\mathcal{Y} \in \mathbb{N}_0$$

(h) [easy] Briefly describe $z_1, \ldots, z_t$ in English where $y = t(z_1, \ldots, z_t)$ in this *phenomenon* (not *model*).

Let: z_1 = making healthy choices, z_2 = genetics, z_3 = unfortunate events

(i) [easy] From this point on, you only observe x_1. What is the value of p?

p is unknown.

(j) [harder] What is $\mathcal{X}$ mathematically? If your information contained in x_1 is non-numeric, you must coerce it to be numeric at this point.

$$\mathcal{X} \in \mathbb{N}_0$$

(k) [easy] How did we term the functional relationship between y and x_1? Is it approximate or equals?

This relationship will be approximate.

(l) [easy] Briefly describe *supervised learning*.

Supervised learning is the process of collecting data and incrementally improving the model with each new observation.

(m) [easy] Why is *supervised learning* an *empirical solution* and not an *analytic solution*?

This method is empirical because it makes decisions based on data rather than intuition. Even if the underlying theory is sound, if the data is "garbage," the output will be unusable.

(n) [harder] From this point on, assume we are involved in supervised learning to achieve the goal you stated in the previous question. Briefly describe what $\mathbb{D}$ would look like here.

Let: $\mathbb{D} = \beta_0 + \beta_1 z_1$

This model would attempt to understand the correlation between eating apples and going to the doctor.

(o) [harder] Briefly describe the role of $\mathcal{H}$ and $\mathcal{A}$ here.

$\mathcal{H}$ is the subset of possible models that we select from. In this case, I have chosen a linear model.
$\mathcal{A}$ is the model in $\mathcal{H}$ that best minimizes error.

(p) [easy] If $g = \mathcal{A}(\mathbb{D}, \mathcal{H})$, what should the domain and range of g be?

$g \in \mathbb{N}$

(q) [easy] Is $g \in \mathcal{H}$? Why or why not?

Most likely not. This is due to specification error.

(r) [easy] Given a never-before-seen value of x_1 which we denote x^*, what formula would we use to predict the corresponding value of the output? Denote this prediction $\hat{y}^*$.

(s) [harder] In lecture I left out the definition of f. It is the function that is the best possible fit of the phenomenon given the covariates. We will unfortunately not be able to define "best" until later in the course. But you can think of it as a device that extracts all possible information from the covariates and whatever is left over δ is due exclusively to information you do not have. Is it reasonable to assume $f \in \mathcal{H}$? Why or why not?

(t) [easy] In the general modeling setup, if $f \notin \mathcal{H}$, what are the three sources of error? Copy the equation from the class notes. Denote the names of each error and provide a sentence explanation of each. Denote also e and $\mathcal{E}$ using underbraces / overbraces.

Ignorance error:
This is the error caused by not including every relevant variable in the model.
Misspecification error:
This is the error caused by not accurately modeling the relationship between the dependant and independant variables.
Estimation error:
This is the error caused by inferring towards theta with the data you collected. This will always cause some amount of error as you can never be perfectly certain of an inference.

(u) [easy] In the general modeling setup, for each of the three source of error, explain what you would do to reduce the source of error as best as you can.

To minimize ignorance error, you can include more variables.
To minimize misspecification error, you can increase H, expanding the possible relationships you will model.
To minimize estimation error, you can increase the amount of data you collect to better infer towards theta.

(v) [harder] In the general modeling setup, make up an f, an h^* and a g and plot them on a graph of y vs x (assume $p = 1$). Indicate the sources of error on this plot (see last question). Which source of error is missing from the picture? Why?

(w) [easy] What is a null model g_0? What data does it make use of? What data does it not make use of?

The null model assumes that all variables have no effect at all on the output. It makes use of the mean to make a baseline prediction.

(x) [easy] What is a parameter in $\mathcal{H}$?

(y) [easy] Regardless of your answer to what $\mathcal{Y}$ was above in (g), we now coerce $\mathcal{Y} = \{0, 1\}$. What would the null model g_0 be and why?

The null could output the mode of the data.

(z) [easy] Regardless of your answer to what $\mathcal{Y}$ was above in (g), we now coerce $\mathcal{Y} = \{0, 1\}$. If we use a threshold model, what would $\mathcal{H}$ be? What would the parameter(s) be?

$\mathcal{H} = (\mathbb{1}_{\boldsymbol{x}>\theta})$

(aa) [easy] Give an explicit example of g under the threshold model.

Let:
1 = no visits(doctor away)
0 = requires doctor visits

$$y = (\mathbb{1}_{\boldsymbol{x}>\frac{2apples}{day}})$$

Problem 4

As alluded to in class, modeling is synonymous with the entire enterprise of science. In 1964, Richard Feynman, a famous physicist and public intellectual with an inimitably captivating presentation style, gave a series of seven lectures in 1964 at Cornell University on the "character of physical law". Here is a 10min excerpt of one of these lectures about the scientific method. Feel free to watch the entire clip, but for the purposes of this class, we are only interested in the following segments: 0:00-1:00 and 3:48-6:45.

(a) [harder] According to Feynman, how does the scientific method differ from learning from data with regards to building models for reality? (0:08)

The scientific method provides a framework to learn from data. This involves incrementally improving the hypothesis to build better models.

(b) [harder] He uses the phrase "compute consequences". What word did we use in class for "compute consequences"? This word also appears in your diagram in 2a. (0:14)

We called this analyzing the data outputted from the model.

(c) [harder] When he says compare consequences to "experiment", what word did we use in class for "experiment"? This word also appears in your diagram in 2a. (0:29)

We called this prediction modeling.

(d) [harder] When he says "compare consequences to experiment", which part of the diagram in 2a is that comparison?

we called this model validation.

(e) [difficult] When he says "if it disagrees with experiment, it's wrong" (0:44), would a data scientist agree/disagree? What would the data scientist further comment?

(f) [difficult] [You can skip his UFO discussion as it belongs in a class on statistical inference on the topic of H_0 vs H_a which is *not* in the curriculum of this class.] He then goes on to say "We can disprove any definite theory. We never prove [a theory] right...

We can only be sure we're wrong" (3:48 - 5:08). What does this mean about models in the context of our class?

No model will ever be perfect. However, a model does not need to be perfect to be useful. In our class we are trying to learn useful insights from models even though we know the model is wrong.

(g) [difficult] Further he says, "you cannot prove a *vague* theory wrong" (5:10 - 5:48). What does this mean in the context of mathematical models and metrics?

We should strive to make clear predictions based on clear models. This would allow our process to be repeatable and our theories to be bolstered. Otherwise, our predictions will be infallible and by Popper's definition unscientific.

(h) [difficult] He then he continues with an example from psychology. Remember in the 1960's psychoanalysis was very popular. What is his remedy for being able to prove the vague psychology theory right (5:49 - 6:29)?

In order to get around epistemological vagueness in language it is important to use numbers to quantify amounts.

(i) [difficult] He then says "then you can't claim to know anything about it" (6:40). Why can't you know anything about it?

If your experiments use vague and subjective terms, your conclusions will be equally subjective.