MATH 390.4 / 650.2 Spring 2020 Homework #1

Kurt Werber

Friday 14th February, 2020

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 interchangeably today?

Silver likens prediction to looking into a crystal ball, as opposed to forecasting which requires using prior knowledge, similar to foresight. These two terms today are used mostly interchangeably.

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

Ioannidis claimed most positive research findings are false. This means that despite what happens in research, most clinical trials were not replicable in real scenarios.

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

Our most powerful defense is our ability to recognize patterns.

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

Useful information.

(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.

The objective truth is emergent from all the minutia that influence phenomena. We represented this in class with the $t(z_1, \ldots, z_t)$, the beyond-our-knowledge function that incorporates all true parameters that affect an outcome.

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

For Popper, science is the practice of testing a prediction.

(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.

The results were so skewed because the ratings agencies lacked sufficient historical data. This makes it nearly impossible for an algorithm to spit out a good function.

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

Risk is quantifiable while uncertainty is not.

(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.

Silver defines out of sample to be when a change of context causes $\mathbb{D}$ to be irrelevant. It essentially means there was an additional z in $t(z_1, \ldots, z_t)$ not accounted for by an x in $f(x_1, \ldots, x_n)$

(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.

According to Silver's definitions, bias reduces accuracy because it causes estimators to miss their target, while higher variance reduces precision because it increases the chance that estimates will not be close to each other.

Problem 2

Below are some questions about the theory of modeling.

(a) [easy] Redraw the illustration from lecture one except do not use the Earth and a tabletop globe. 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 natural result of measuring a phenomenon.

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

Predictions are guesses to the what a phenomenon will be based on measured proxies assumed to have some causal relation with said phenomenon.

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

All models are wrong due to the inherent ignorance error caused by what we don't know. If we knew everything that influenced a phenomenon, there would be no need to model that phenomenon.

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

Some models are useful because they are more accurate then simply guessing, often incredibly so.

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

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.

No. One could argue that the "apple a day part" is, but there is no quantification in "keeps the doctor away".

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

The input is whether or not one has eaten an apple on a given day.

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

The output is the frequency at which one gets sick.

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

This is *very* bad model. Unfortunately, apples are not magical foods capable of keeping illness at bay. There will be a lot of error due to ignorance thanks to omitting so many important variables.

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

x_1 could be the number of days one has eaten an apple in a given period of time, let's say a year.

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

y could be the number of illnesses contracted in the same time period measured by x_1.

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

$\mathcal{Y} = \{0, \mathbb{N}\}$

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

While not fully understood, immune system performance would be a function of the number of nutrients someone eats a sufficient amount of each day, the average temperature and humidity, and whether or not they have been immunocompromised by drugs or conditions, such as immune suppressants or AIDS.

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

$p = 1$

(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.

$x_1 = \{0, 1, \ldots, 365\}$

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

It is very much approximate.

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

Supervised learning is a machine learning paradigm in which a functional relationship between an outcome variable and one or more explanatory variables is established by a an algorithm refining a set of candidate functions using a data set.

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

Its result is derived directly from data as opposed to attempting to achieve some closed form expression in which the data could be plugged in.

(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.

$\mathbb{D}$ would be a matrix with n rows and 2 columns.

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

$\mathcal{A}$ is the means by which we can achieve the best possible function $\in \mathcal{H}$

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

The domain and range should be respectively the possible values of x_1 and y listed above.

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

Yes

(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 $\hat{y}^*$.

$\hat{y}^* = f(x^*)$

(s) [harder] Is it reasonable to assume $f \in \mathcal{H}$? Why or why not?

No, $\mathcal{H}$ must be chosen ad hoc so there is no way of knowing.

(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.

$$y = g(\vec{x}) + \overbrace{\underbrace{[h^*(\vec{x}) - g(\vec{x})] + \overbrace{[f(\vec{x}) - h^*(\vec{x})]}^{\mathcal{E}} + \overbrace{[t(\vec{z}) - f(\vec{x})]}^{\delta}}_{e}}$$

δ : error due to ignorance
$\mathcal{E}$: misspecification error
e : the residual

(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 reduce error due to ignorance, get more or better xs. To reduce misspecification error, $\mathcal{H}$ must be expanded to include more complicated functions. To reduce estimation error, collect more data to increase n.

(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?