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

Intellectual Carlos Vega

Tuesday 13th 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?

The term predict is used to make a more definitive statement about when an event will occur, such as the release date of the new iPhone, while forecast is giving a probabilistic statement, such as there is a 30 percent chance of rain on Tuesday. In my experience, the terms are used interchangeably in a casual setting but in a scientific setting, there is a distinction to be made.

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

John P. Ioannidis's findings were that the reliability of research, especially in the biomedical field, can be very questionable. He found that a significant portion of positive findings reported in medical journals were likely to fail when applied in the real world. This was later supported by Bayer Laboratories which found 2/3 of positive findings cannot be replicated in their own experiments.

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

The most powerful tool that human beings possess is our "wits". Despite the fact that we do not possess claws, speed, fangs, etc., our brains were the key factor to our survival since we historically depended on our cognitive abilities. As powerful as this sounds in this day and age of information overload, it can be a double-edged sword as we are also susceptible to bias in the information we process. This is crucial for our critical thinking to be aware of this cognitive bias as we would be able to minimize how much ignorance errors our models could experience before calculations.

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

While information is increasing at a rapid pace, useful information is not keeping up the same speed. As stated in the book, "the amount of useful information almost certainly

isn't. Most of it is just noise, and the noise is increasing faster than the signal." What this is trying to illustrate is that in this modern era of Big Data, the growth of information is not parallel to the growth of knowledge or understanding.

(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 in class was discussed as "reality" and not the result of models we have made in class. The class began by stating that "all models are wrong, but they are useful." With this, we made an initial model

$$f\left(\begin{bmatrix}\mathbf{b}\\ \mathbf{w}\end{bmatrix}\right) = \begin{bmatrix}\mathbf{l}\\ \mathbf{n}\\ \mathbf{s}\end{bmatrix} \tag{1}$$

where f is a function with parameters "settings" which outputs our "phenomena", in the form in which we construct our model is where we will always encounter a subjective bias. It is our job to filter out as much bias as possible which is by approximating our "settings".

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

Karl Popper defined science through the principle of falsiability, according to Karl a hypothesis is not scientific unless it is able to be proven wrong through empirical testing and observation.

(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 reason for such drastic differences in predictions are due to faulty statistical models, the models used were not based on historical data but were assumptions made using flawed statistical model which did not accurately account for the risk associated with these new and highly complex securities. As we discussed in class they were not able to see the drastic difference their estimation error would actually effect their predictions

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

According to Nate Silver the difference between risk and uncertainty is that risk is something that can be measured and quantified. For example, as an example in poker, you can quantify the odds of your opponent winning from a card on the river. but uncertainty in the other hand, refers to where risk is hard to measure as if you had no knowledge of how many aces are in the deck in a hypothetical sense, there could be 4 or 5 or 20. It is uncertain.

(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.

In the context of the book Silver defined out of sample to be situations where the predictions are tested against data that was not used in the model's training. In lab we made a model to predict the flower type using data given to us, then later made out "predicate" to predict based on an arbitrary input length.

(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.

Bias refers to the error introduced by approximating which may be complex. It is the difference between the prediction we obtain from the model and the actual value we try to predict. If your model has high bias the difference would be significantly greater. Variance in the other hand describes how much the prediction of a model varies for a given input in our data. With this being said, accuracy in the context of the book is how close the model's prediction to the actual value is, in contrast, precision is how consistent the model is at providing a reasonable prediction when given subsets of data.

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?

For the purpose of this class, data is regarded as the information collected from reality that can be quantified to help us in the creation of models, to attempt at approximating reality.

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

Predictions, for the purpose of this class, are the results we collect from a model after inputting data that is not a subset of the training data, to help us predict reality.

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

We claim that all models are wrong as for all of the settings we make in a model they are just approximations of said setting, we can never know with 100 percent certanty these settings so our best solution is to approximate them.

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

Models are simplifications of reality; they are built to capture useful information about a phenomenon being studied while omitting irrelevant details. Another reason they are useful is that they assist us in our decision-making by providing insight into the phenomenon being studied.

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

The differences between a good model and a bad model include, but are not limited to, the following: A good model possesses a high level of accuracy in predicting the outcomes of the phenomenon in question and is generalized well enough to be applicable to previously unseen data.

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.

Although this aphorism specifies quantities it is ambiguous as to the reasons why a single apple is good and doesn't provide a reasonable measurement of health.

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

The inputs in this model is the action of eating an apple.

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

The output is good health/keeping the doctor away.

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

I think this model is bad as it is ambiguous, it leaves us with no useful parameters to alter to see the changes in the output, it is binary in the sense that the only measure we have is eating apples.

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

My metric is numeric as it counts days in a row that apples were eaten in the time frame.

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

My metric is to count the amount of doctors visits in a row (good health)

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

Y mathematically represents the set of values calculated that count the amount of doctors visits in a row (good health)

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

The z's describe the proximal causes or the drivers in the function t, which takes them as inputs for y, equaling our phenomena which is how many consecutive days were there without a doctor visit.

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

We can assume that the value of p is low as we have a great δ due to the ignorance error from not including more "inputs" into our Y.

(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.

$\mathcal{X}$ mathematically represents a numerical integer representing the consecutive days in a row that an apple was eaten.

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

The functional relationship between y and x_1 can be termed as an approximation rather than an equals. This approximation is a result of the model not accounting for all possible values, (z_1, ..., z_t) that can affect health such as dietary habits, genetics etc. Therefore, we term the functional relationship as an approximation of $f(x_1)$ indicating that while there is a relationship it is an approximation.

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

Supervised learning is a type of machine learning where an algorithm is trained on a labeled dataset, which means that each training example is paired with an output label. We start with $\mathbb{D}$ and $\mathcal{H}$. $\mathbb{D}$ represents the input and output of x's and y. $\mathcal{H}$ represents the set of candidate functions because the space of all functions is too large

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

Supervised learning is an empirical solution because it is purely reliant on the data provided for the "learning" instead of being derived from a mathematical or logical analysis. Since is purely dependant on the data provided we have to assure the quality, quantity and reliability of the data provided

(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.

For us $\mathbb{D}$ would look like a set of X's with numerics for the number of consecutive days of apples eaten, with a second component for y, for all the outputs for their respective X's

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

The role of $\mathcal{H}$ is to efficiently pick a set of all candidate functions we need because the space of all functions is too large to be useful. $\mathcal{A}$ is a function that takes in $\mathbb{D}$ and $\mathcal{H}$ and outputs g which is the result of the model.

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

The domain for g is the set of all X's on $\mathbb{D}$ for our case it is only $\S_\infty$ which has a domain of non-negative integers, the range is the set of all possible functions that approximate y.

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

g must exists in $\mathcal{H}$ since $\mathcal{H}$ is a subset of all possible candidate functions so g must exist in $\mathcal{H}$

(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 $\hat{y}^*$.

We use the g function to predict the corresponding output, $\hat{y}^=$, for the never-before-seen value of $x^=$.

(s) [harder] f 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?

It is not reasonable to assume that $f \in \mathcal{H}$ as $\mathcal{H}$ is a subset of all candidate functions in g that approximate f, there is a chance that might exist in $\mathcal{H}$ but it is unreasonable to assume so.

(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.

In the modeling setup when $f \notin \mathcal{H}$ the three sources of errors are estimation error, misspecification error and ignorance error.

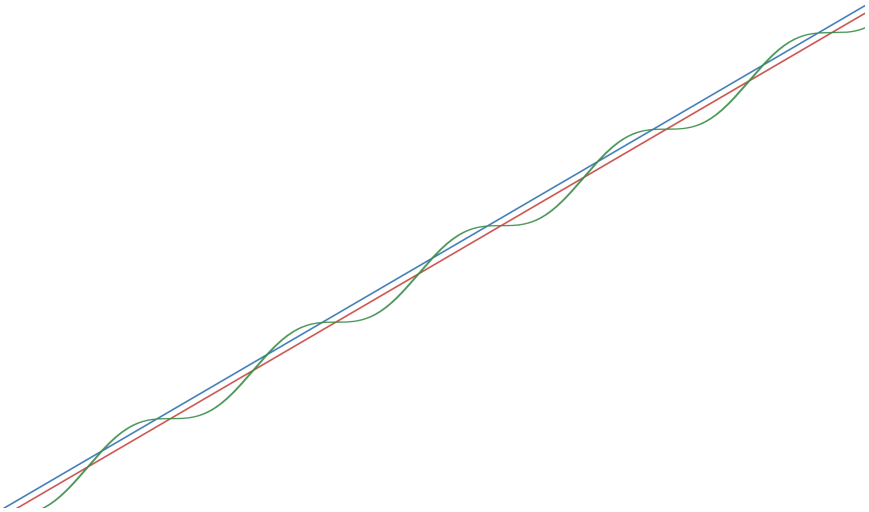
$$y = g(x_1, ..., x_p) + (h^*(x_1, ..., x_p) - g(x_1, ..., x_p)) + (f(x_1, ..., x_p) - h^*(x_1, ..., x_p)) + \delta$$

where δ = ignorance error. In this equation the estimation error is accounting for the difference between our model g and the true representation of the phenomena h^*, the misspecification is trying to equate for the difference between the true model h^* and the chosen model f, lastly the ignorance error is the difference between t which is our original model with Z's as inputs minus the chosen model f.

(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.

For this general modeling setup, I would include more relevant data to reduce the estimation error, secondly to reduce the misspecification I would refine $\mathcal{H}$ to be more expressive by increasing the size of the data set, Lastly to reduce the ignorance variables we need to carefully consider more settings for our model that can provide useful information to out model.

(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?



Green represents f, Red represents h^* and Blue represents g. The estimation error is how far off g is from h^*. The misspecification error is how far h^* is from f. We do not have ignorance error in this plot since it is derived from comparing f to reality model of $(t = \delta = t - f)$.

(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 g_0 is the mode of $\vec{y}$, therefore $g_0 = Mode[\vec{y}]$ and is unable to "see" any of our parameters, $\mathcal{X}$.

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

$\mathcal{H}$ is a set of all candidate functions. In our case, a parameter is a function, y, with parameter x_1, which was our original x, for counting the number of consecutive days apples were consumed and the number of consecutive days without a doctor's visit.

(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?

Coercing $\mathcal{Y}$ - 0,1, The null model will output whether or not the individual has visited the doctor or not, Our new g_0 would convert from a numeric to a binary as the output is in the range of true or false.

(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?

Using the threshold model, $\mathcal{H}$ would be a binary step function with activation steps where its the first 0 to 1 jump for all 0's. This new function called $\mathcal{T}$ will output a numerical integer indicating the number of days in a row that the indiviual needs to consume apples to avoid the doctor the longest consecutive days.

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

A example of g is the threshold model below:

$$g(x) = \begin{cases} 1 & \text{if } x < 300, \\ 0 & \text{if } x > 300. \end{cases}$$

here, 300 represents our threshold number of days in a row.

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)

Feynman's scientific method consists of making a guess, then computing the consequences of the guess (the errors and implications), and comparing the observations to reality. Learning from data is a different approach, as we are able to collect data first, then seek models that help us understand the structure of the phenomena.

(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)

In class we used the word predictor to refer to "computer consequences".

(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)

In class we called "compare consequences" as Model Validation which was to compare the predicted phenomenon to the measured phenomenon from reality.

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

The part of the diagram in 2a we are comparing too when "comparing consequences to experiment" is the part where the models creates predictions and we compare those predictions to the measured data from reality (this is found on the bottom of the diagram)

(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?

A data scientist would agree with this statement as it reflects the fundamental principles of the scientific method, where we emphasize the importance of empirical evidence over theoretical models or hypotheses.

(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?

He means that models are not immutable truths but are best understood as hypotheses subject to constant change and revision. This idea is deeply rooted in the principle of falsifiability previously mentioned in the homework.

(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?

What he means in the context of mathematical models and metrics is to highlight the difficulties in their development. He is trying to illustrate the importance of specificity and falsifiability in both the development and the definition of the metrics used.

(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)?

The remedy he proposes is to make psychology theories ore specific and testable. The specificity comes in the form of a clear hypothesis and having outcomes that can be tested empirically.

(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?

"Then you can't claim to know anything about it" refers to the idea that without the capability to evaluate a theory (or model) to empirical testing, it's impossible for anyone to gain knowledge about the subject at hand that they are attempting to explain.

Just to demonstrate that this modeling enterprise is all over science (not just Physics), I present to you the controversial theoretical political scientist John Mearsheimer. He's all over youtube and there's nothing special about this video that I will link here about Can China Rise Peacefully? Feel free to watch the entire clip, but for the purposes of this class, we are only interested in the following segments referenced in the questions which has nothing to do with China, only his theory of "power politics".

(j) [difficult] Is Mearsheimer's model of great power politics / international relations (i.e., modern history) 9:35-17:22 simple or complicated? Explain.

From my perspective, Mearsheimer's models of great power politics and international relations can be seen as both simple and complex. They attempt to explain the essence of politics and international relations through a power dynamic of states. However, the complexity arises from the intricacies of the modeling. The multitude of factors, including the settings and such, elevate the complexity to extreme heights.

(k) [difficult] Summarize his ideas about limitations of his theory from 39:18-40:00 using vocabulary from this class.

He explains that theories simplify a very complex reality. He thinks a good theory is one that can produce a correct result 75 percent of the time. Like we have learned in class, our models don't capture everything perfectly but can still be useful. He mentions that theories might not always work because they miss some details. In terms of modeling, this means there could have been more information or different settings used to make the model better, but they weren't included.