MATH 342W / 650.4 / RM742 Spring 2022 HW #1

Antonio D'Alessandro

Thursday 10th February, 2022

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 two terms are used interchangably today however in the past the word prediction was something a fortune teller would provide while a forecast was a carefully designed plan taking into account uncertain or incomplete information.

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

Ioannidis found that most findings published in peer-reviewed journals, predictions coming from laboratory experiments, would fail if applied to real world situations. Bayer Labs confirmed Ioannidis's paper - they could not reproduce 2/3 of the "positive findings" published in scientific journals when they attempted the experiments themselves.

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

Our ability to "learn from the data" or detect patterns in information and react to threats in our environment.

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

Our understanding of what to do with all the information i.e. how to process it into useful knowledge.

(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.

We said the objective truth was represented by the function $y = t(z_1, \ldots z_t)$ where y is metric of the phenomenon, t is the true unknowable function and $z_1, \ldots, z_t$ are the true unknowable causal inputs.

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

It is an iterative process where falsifiable hypothesis are put forth, tested and then revised based on empirical results.

(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 approximations / assumptions used to model the probability of a CDO defaulting were wrong. They assumed that the risk of default was uncorrelated (someone defaults in Orlando Florida has no effect on a mortgage holder in Queens New York), when in fact it was. So a feature in their model x_i = default rate, was not being combined properly in their function (model) g being used to approximate t (the true unknowable default probability function).

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

Risk is uncertainty that can be measured quantitatively (with a 'price') with accuracy, Silver gives an example from poker with odds of winning / losing a particular hand.

Uncertainty is risk which is more difficult to measure, you may be aware of a problem but your best guess could be off by a factor of 1000.

(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* as a situation which you have not seen before, where you cannot use extensive past experience to forecast risk, he gives the example of driving for 30 years with no real issues and then deciding if it is a good idea to drive home drunk from a holiday party. The individual has a lot of time behind the wheel but never tried driving under the influence, so you cannot use the historical data (previous time driving not under the influence) to make a prediction about whether they will get into an accident.

In notation from class we have a phenomenon y which we are trying to model with a function g but the data we have to learn from $\mathbb{D}$ is not rich enough, or does not contain enough relevant measurements $< \vec{x_i}, y >$ to generate a prediction $\hat{y}$

(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.

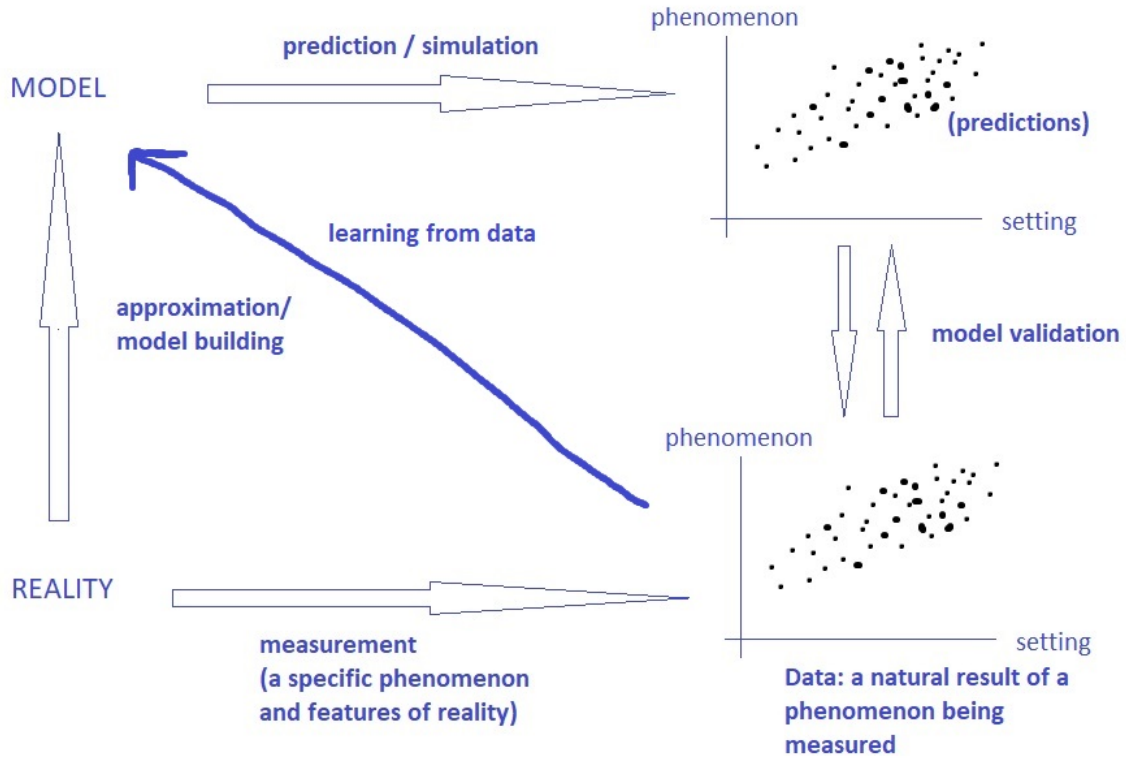
Based on Silvers use of the terms and the definitions of statistical bias and variance we have the following relationship:

accuracy = bias and precision = variance

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 table-top 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 are the measurements being taken which occur as a natural result of the phenomenon of interest.

(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 our model for the phenomenon/a of interest.

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

All models are wrong because we can never know the true "function" and true casual inputs which generate the phenomenon of interest, the models we construct are at best approximations of things which are unknowable, therefore they all contain errors which cannot be overcome.

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

Some models are useful because their errors are low enough such that we can use them to more accurately make predictions i.e. their predictions agree with nature or experiment to some degree of accuracy.

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

That would depend on what you are trying to accomplish with your model. At the start of class we said there are two potential goals: prediction and understanding causality. If our goal is prediction, which is what the class will focus on then a good model is one whose predictions agree with nature or our experiments.

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 we have not completely quantified the statement, there are still some ambiguities - for instance what do we mean by "keeps the doctor away"? Less than once a month, once a year, never? The desired result is unclear.

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

The input to the model would be daily apply consumption.

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

The output would be frequency of visits with a doctor.

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

It is a bad model because it is not fully mathematical and ambiguous in what the desired outcome is / should be.

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

x_1 = number of apples consumed per day

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

y = number of doctor visits per unit time

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

$\mathcal{Y}$ is the set of all possible outputs for the metric y. In this case it would be the set of natural numbers $\mathbb{N}$ including 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*).

$z_1, \ldots, z_t$ are unknowable true causal inputs which determine the response or outcome y. They could include the genetics of the person in question, what their overall diet is, level of physical activity, occupation etc.

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

p represents the total number of features/regressors (the $x's$) we are using to obtain information related to the $z's$ which are unknowable to us. In this case $p = 1$.

(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.

$\mathcal{X}$ is known as the covariate or input space and it is the set of all possible measurements we could obtain based on our selection of features.

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

a true equality between y and x_1 can be represented by:

$y = g(x_1) + (h^* - g) + (f - h^*) + (t - f)$

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

Supervised learning is a process where an algorithm is employed to find the best way to combine our chosen collection of features based off data we have collected.

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

It is an empirical solution because you are using data collected from some kind of experiment or survey to find the best way to combine your features vs. a deductive approach like a mathematical derivation to arrive at a closed form solution.

(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.

$\mathbb{D}$ is our set of training data. In the case of the phrase "An apple a day ..." it would be a matrix with two columns and n rows, where n would represent the total number of subjects in our survey. The first column would record the $x_{i,1}$'s the number of apples consumed per day while the second column, the $y_{i,1}$'s, would have the total number of doctor visits for that same subject over the course of the unit time we agree on.

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

$\mathcal{H}$ is the set of all candidate functions h we would use to approximate f. While $\mathcal{A}$ is the algorithm applied to our training data to search for g the best approximation to h^* where h^* is the best candidate 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 of g should be $\mathbb{D}$ and the range should be $\mathcal{Y}$.

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

Yes g is in $\mathcal{H}$ it is a possible function used to combine our features to approximate the function f.

(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 $\hat{y}^*$.

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

(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?

No it is not reasonable to assume $f \in \mathcal{H}$. The function f can be arbitrarily complicated and the set $\mathcal{H}$ is the set of all functions which could approximate it, by definition the set is a collection of much more "simple" functions which could be found by using an algorithm $\mathcal{A}$ applied to our training data $\mathbb{D}$.

(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.

$$y = g(x_1) + \overbrace{(h^* - g)}^{\text{est error}} + \underbrace{(f - h^*) + \underbrace{(t - f)}_{\delta}}_{\epsilon} \Bigg\}_{e}$$

δ is our "error due to ignorance" which arises from the fact that the inputs ($z_i's$) to and the function t and the function t itself are unknowable to us and must be approximated by our features the $x_i's$ being input to f.

ϵ is equal to the sum of both δ and $(f - h^*)$ where $(f - h^*)$ is model miss-specification error which results from a selection of $\mathcal{H}$ that is too strict. That is $\mathcal{H}$ does not include enough possible approximations to the function f.

"est error" short of estimation error is a result of a inefficient algorithm or too little data for the algorithm to learn from.

Finally, e is known as the "residuals" or sum of all the possible error.

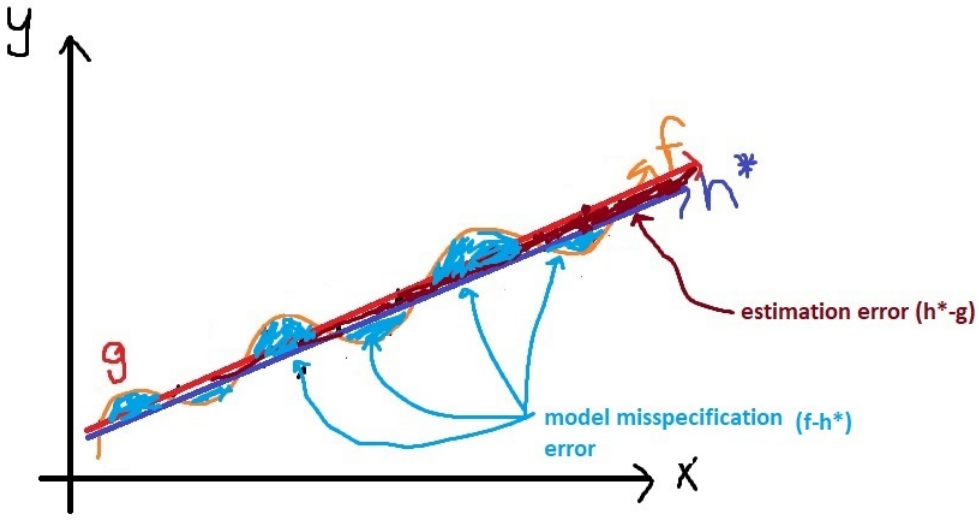
(u) [easy] In the general modeling setup, for each of the three source of error, explain what you would do to reduce the source of error as best as you can.

To reduce δ you would need to collect more relevant $x's$ features that could approximate the true casual inputs $z's$.

To reduce model miss-specification one would need to expand $\mathcal{H}$ the set of possible candidate functions to approximate f to include more complicated functions.

To reduce estimation error $h^* - g$ you could find a better algorithm $\mathcal{A}$ or collect more observations (increase amount of data 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?



The first source of error we discussed δ is missing from the picture because the true function t is unknowable and therefore so are its outputs so we cannot measure the distance (f-t).

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

$g_0 = \mathcal{A}(\vec{y}, \mathcal{H})$

It is a model that does not have any features - it is the "simplest stupidest" model you can come up with based on the collected outputs of the phenomena of interest.

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

A parameter in $\mathcal{H}$ would be a potential weight for one of our selected features.

(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 model would be $g_0 = \text{Mode}[\vec{y}]$. This would be the null model because it is the simplest way to determine "doctor visits" in the context of problem 3.

(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?

If we select a threshold model $\mathcal{H}$ would be all possible lines we could draw to separate the $0's$ from the $1's$ if we were to geometrically represent our model in the plane. The parameter would be the weight w_1 we are assinging to our one feature x_1 as well as the bias b.

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

$g(x) = \mathbb{1}_{x \geq 10}$

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)

I'm not sure about this. It seems the same to me. Is the difference because Feynman says to get started with the scientific method you are guessing as to what the relationship is as opposed to learning from data which starts from a proposed relationship between features and outputs?

(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)

"Compute consequences" would correspond to using our model to generate predictions, so the "consequences" according to Feynman would be our predicitions.

(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)

Our word from class which we used for "experiment" would be data.

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

When we do the model validation, that is compare our model predictions to the data.

(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?

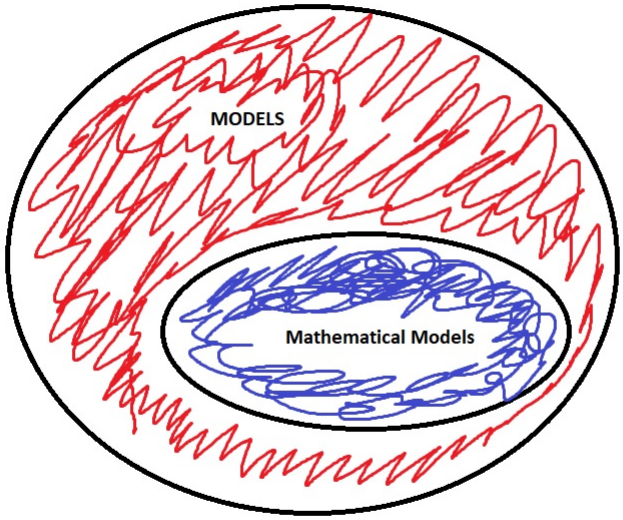
Yes a data scientist would agree if you construct a model and compare your models predictions to what actually happens in nature, and the model is off dramatically it can't be a good model. As a consequence your ideas about the relevant features and approximation function (represented by the model) are wrong in some way. They may further comment about

(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?

In the context of our class it means once we have a model we can always compare its predictions to nature or experiment and evaluate (validate) it's performance. If the model is bad at making predictions, it is just like if the "computed consequences" don't agree with experiment, that is it would be a bad model. If however the model is good at making predictions it does not necessarily mean our ideas have been proven correct, just that they are not wrong for the time being. It could happen that at a later point in time the model could produce bad predictions and would need to be re-evaluated.

(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?

Mathematical models and metrics are required for the scientific method to be performed. A "vague theory" according to Feynman would likely correspond to one which falls in the red shaded region



Since a vague theory has not been quantified using mathematics there are ambiguities which exist that prevent it from being evaluated against data coming from the real world.

(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 defining precisely (quantifying) what is meant by too much or too little "love" so that tests could be performed.

(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?

You can't claim to know anything in this case because it is impossible to state in mathematical terms what is meant by too much or too little love and therefore no experiments can be done. The theory and method of "computing its consequences" remains vague and as such any results can always be twisted to conform to what the theory about being loved by your mother as a child would predict. It can never be proven wrong.