MATH 390.4 / 650.2 Spring 2020 Homework #1

Professor Adam Kapelner

Due 11:59PM Tuesday, February 11, 2020 under the door of KY604

(this document last updated 12:43am on Wednesday 12th February, 2020)

Instructions and Philosophy

The path to success in this class is to do many problems. Unlike other courses, exclusively doing reading(s) will not help. Coming to lecture is akin to watching workout videos; thinking about and solving problems on your own is the actual "working out." Feel free to "work out" with others; **I want you to work on this in groups.**

Reading is still *required*. For this homework set, read the first chapter of "Learning from Data" and the introduction and Chapter 1 of Silver's book. Of course, you should be googling and reading about all the concepts introduced in class online. This is your responsibility to supplement in-class with your own readings.

The problems below are color coded: green problems are considered *easy* and marked "[easy]"; yellow problems are considered *intermediate* and marked "[harder]", red problems are considered *difficult* and marked "[difficult]" and purple problems are extra credit. The *easy* problems are intended to be "giveaways" if you went to class. Do as much as you can of the others; I expect you to at least attempt the *difficult* problems.

This homework is worth 100 points but the point distribution will not be determined until after the due date. See syllabus for the policy on late homework.

Up to 7 points are given as a bonus if the homework is typed using LaTeX. Links to instaling LaTeX and program for compiling LaTeX is found on the syllabus. You are encouraged to use `overleaf.com`. If you are handing in homework this way, read the comments in the code; there are two lines to comment out and you should replace my name with yours and write your section. The easiest way to use overleaf is to copy the raw text from hwxx.tex and preamble.tex into two new overleaf tex files with the same name. If you are asked to make drawings, you can take a picture of your handwritten drawing and insert them as figures or leave space using the "\vspace" command and draw them in after printing or attach them stapled.

The document is available with spaces for you to write your answers. If not using LaTeX, print this document and write in your answers. I do not accept homeworks which are *not* on this printout. Keep this first page printed for your records.

NAME: ______________**PIZON SHETU** ______________

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?

According to Silver forecast and predict are used interchangeably today, but this was not always the case prior. In Shakespeare's time a prediction was what the soothsayer told you; a forecast was something more like Cassius's idea. Making a forecast typically implied planning under conditions of uncertainty. It suggested having prudence, wisdom, and industriousness, more like the way we now use the word foresight. Whereas a prediction was a vague event or something that might happen in the future.

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

His published a paper called "Why Most Published Research Findings Are False." The paper studied positive findings documented in peer-reviewed journals: descriptions of successful predictions of medical hypotheses carried out in laboratory experiments. It concluded that most of these findings were likely to fail when applied in the real world. Bayer Laboratories recently confirmed Ioannidis's hypothesis. They could not replicate about two-thirds of the positive findings claimed in medical journals when they attempted the experiments themselves. This implies that not all experiments translate into to the real-world the same can be said when it comes to modeling, while in theory some models might look great but in application it might not yield the same results

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

According to Silver, as humans its our pattern recognition. Our ability to recognize and assess pattern in difficult situation allowed us to propel. This is crucial in data recognition, being able to depict a pattern from data and making something out and being able to devise a model which might yield to results that benefit reality is key.

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

While quantity of information is increasing, the amount of useful information is not. According to Silver most of it is just noise, and the noise is increasing faster than the signal. There are so many hypotheses to test, so many data sets to mine—but a relatively constant amount of objective truth.

(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.

$$Y = t(Z1, Z2, ...Zt)$$

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

For something to be scientific, it needed to be falsifiable, it other words it had be to able to be tested in the real world by means of a prediction. Otherwise Popper would not view a hypothesis scientific.

(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 ratings agency created a model which predicted that only 0.12 percent of the CDO would default but unfortunately reality was different, the rating agency used a poor model which did not tell the true picture of the mortgage loans at the time. Thus their model was at fault but not the CDO's itself. According to Silver "In the instance of CDOs, the ratings agencies had no track record at all: these were new and highly novel securities, and the default rates claimed by SP were not derived from historical data but instead were assumptions based on a faulty statistical model." and the ratings agencies made a out of sample mistake. Moody's estimated the extent to which mortgage defaults were correlated with one another by building a model from past data—specifically, they looked at American housing data going back to about the 1980s. 101 The problem is that from the 1980s through the mid-2000s, home prices were always steady or increasing in the United States. Under these circumstances, the assumption that one homeowner's mortgage has little relationship to another's was probably good enough. But nothing in that past data would have described what happened when home prices began to decline in tandem. The housing collapse was an out-of-sample event, and their models were worthless for evaluating default risk under those conditions.

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

A risk is a calculated idea where you're odds of playing a hank in poker, e.g you have probability of winning 3/11 hands but risk of losing 8/11 hands. Whereas a uncertainly is the lack of knowledge where you just lack the information to even calculate your odds. According to Silver "Risk greases the wheels of a free-market economy; uncertainty grinds them to a halt."

(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 defines as in a person uses data $\mathbb{D}$ to predict something but the data $\mathbb{D}$ is not under the same conditions as needed for the prediction thus it was an out of sample problem. In other words his model $g = \mathcal{A}(\mathbb{D}, \mathcal{H})$ is using a set of $\mathbb{D}$ to predict an outcome that are under different circumstances than that of the $\mathbb{D}$

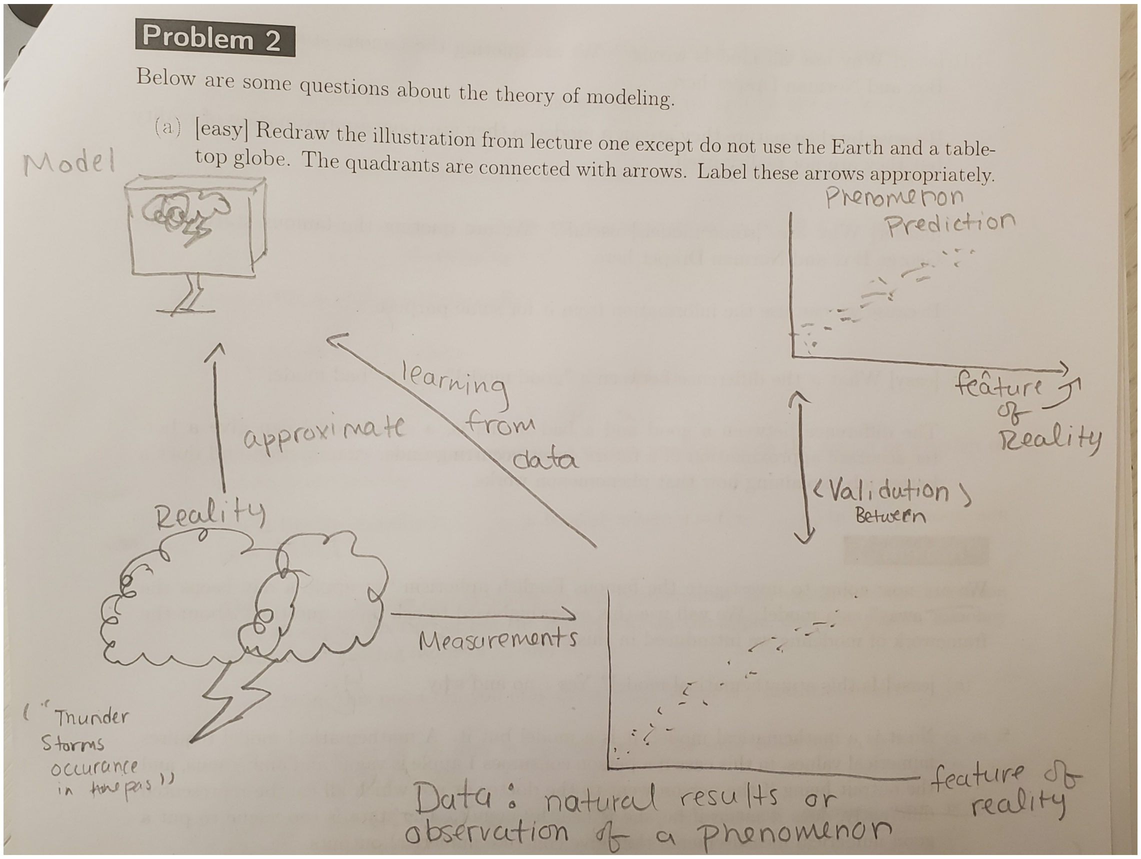
(j) [harder] Look up *bias* and *variance* online or in a statistics textbook. Connect these concepts to Silver's terms *accuracy* and *precision*. This is another example of Silver using non-standard terminology.

Silver makes a point by saying history has shown failures of prediction—stem from the false sense of confidence from a precise forecast to justify their actions and mistake it for being an accurate prediction by this he means just because when you're trying to shoot a target and you hit the same spot consistently does not mean it was close to the target which is accurately. Bias refers to the tendency of a measurement process to over- or under-estimate the value of a population parameter. The variance is a numerical value used to indicate how widely individuals in a group vary. If individual observations vary greatly from the group mean, the variance is big; and vice versa. I think the connection can made when people use a precise prediction and it becomes a bias one rather than an accurate one. Whereas the variance will tell us how far it really is from our actual target.

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.



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 can be defined as measurements of reality or historical examples or even something that has happened already.

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

Predictions can be defined as it the ability to say or see if something will occur in the future with certain odds. In other words prediction is the process of determining the magnitude of statistical variates at some future point of time

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

Because by their nature they are an a model so they are some approximation of reality but they are not reality itself.

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

Because we can use the information from it for some purpose.

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

The difference between a good and a bad model is, a good model can give a better accurate approximation of a future event occurring under examination and does a better job explaining how that phenomenon works.

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 it is a mathematical model, it is a model but it. A mathematical model requires numerical values, in this case if a person consumes 1 apple is vague and ambiguous, and the output being if the person went to the doctor or not which all can be represented differently, does it mean if he/she is healthy or not. The state is too vague to put a good numerical measurements that give concrete numerical outputs.

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

If a person ate an apple at least once a day

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

If the person went to the doctor or not, or you can say if he/she is healthy or not because if they are healthy they would not need to go to the doctor's

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

This is a bad model because the quote is ambiguous and it takes in a vague criteria to justify in this case good health where in reality you could be going to the doctor even if you are healthy and eating apples does not correlate with health.

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

Whether or not a person ate an apple where $x_1 \in (0, 1)$ if he/she ate an apple $x_1 = 1$ otherwise $x_1 = 0$

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

Whether or not a person went to the doctor's where $x_1 \in (0, 1)$ if he/she went to the doctors $x_1 = 1$ otherwise $x_1 = 0$

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

$\mathcal{Y}$ is the set of all y_i where

$$y \in Y = \{0, 1\}$$

$$Y = t(z_1, z_2, ...z_t)$$

Y is going be the prediction whether a person needs to go to the doctor or whether or not they are healthy.

(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 $z_1, \ldots, z_t$ are the casual inputs, in this *phenomenon* they would be does the person have an apple, whether or not the person will eat the apple, does the person have any intention of visiting a doctor,... and etc.

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

p is the number of features in our case since we only observed x_1 $p = 1$ or in a general case it is $p = (x_1, \ldots, x_i)$

(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 the input space matrix which has all the x_i where $\mathcal{X} = [x_1, x_2, ...x_i]$ where $x_i \in (0, 1)$

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

$y \neq x_1$ but x_1 helps us get a good approximation to our casual drivers which are the z_i which predicts y thus $(y = t(z_1, z_2...z_t)) \neq (f = (x_1, x_2...x_p))$

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

Supervised learning is simply a process of learning algorithm from the training dataset. Supervised learning is where you have input variables and an output variable and you use an algorithm to learn the mapping function from the input to the output. The aim is to approximate the mapping function so that when we have new input data we can predict the output variables for that data

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

Because supervised learning supported by experiment, observation and data but not necessarily supported by theory, whereas a analytical solution would based on logical steps that can be followed and verified as correct

(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 actual training data set so

$$\mathbb{D} =< X, Y >$$

where

$$X = (x_1, x_2, ...x_n), Y = (y_1, y_2, ...y_n)\ x_i \in X, yi \in Y$$

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

$\mathcal{H}$ is the set of all candidate functions so in this case its the set of all functions that can take in our inputs and spit out an appropriate output. $\mathcal{A}$ is our Algorithm which takes in $\mathcal{D}$ and $\mathcal{H}$ which produces a model $g = \mathcal{A}(\mathcal{D}, \mathcal{H})$ in this case A is our algorithm which will try to find the best $h^* \in \mathcal{H}$ which will provide us with a model to predict whether or not the person visit's the doctor or whether or not if a person is going to be healthy.

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

The domain of g is going to be the set of all different pairs of $\mathbb{D}$ and $\mathcal{H}$ and the range is going to the set of all possible models corresponding to the pairs in our domain.

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

$g \in \mathcal{H}$ because we get g from $\mathcal{A}$ which finds the best $h^* \in \mathcal{H}$ and use's $\mathcal{D}$ to get a g

(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 $\hat{y}^*$.

Given that we have g we can use this with x^* to get $\hat{y}^*$ so our formula would be $\hat{y}^* = g(x^*)$

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

No it is not f is $\notin$H$because$h* is an approximation of f which is $\in$Hbutf$\neq$h*

(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.

$$Y = g(x) + (h^*(x) - g(x)) + (f(x) - h^*(x)) + (t(z) - f(x))$$

where g(x) is your model
$(h^*(x) - g(x))$ is your estimation error
$\epsilon = (f(x) - h^*(x)) + (t(z) - f(x))$ is your epsilon error where
$\delta = (t(z) - f(x))$ is your delta error and
$(f(x) - h^*(x))$ is your mis-specification error
$y* = g(x) where y*$ is the prediction of y in setting x
$e = y - y*$ residual if x element of D (training data) otherwise they are unknown

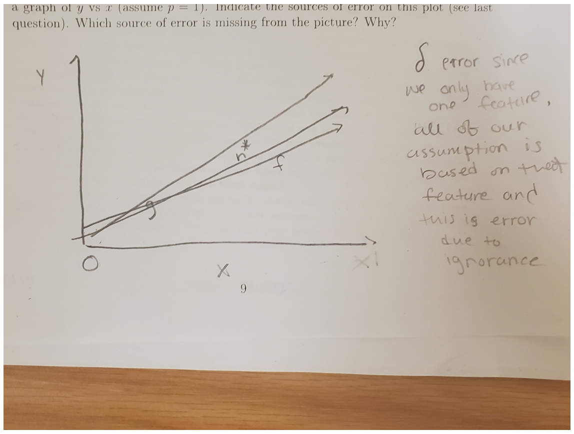
(u) [easy] In the general modeling setup, for each of the three source of error, explain what you would do to reduce the source of error as best as you can.

(a) δ, ignorance error can be reduced by measuring more $X_j's$ (features) of the units that contain information about Z.

(b) Misspecification error can be reduced by expanding $\mathcal{H}$ to include more complicated functions

(c) Estimation error can be reduced by increasing sample size

(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?


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?

δ error since we only have one feature, all of our assumption is based on that feature and this is error due to ignorance