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

Laasya Indrakanti

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 interchangeably today?

Forecasts are predictions on the future. The terms are used interchangeably today.

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

Ioannidis found that most positive findings in laboratory experiments are actually false positives. This implies that there is a very high need for improving research practices and replicating experiments so that the results are reliable and reproducible.

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

According to Silver, the human being's most powerful defense is our wit and pattern detection.

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

The amount of useful information is not increasing as fast.

(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.

$f(x_1, \ldots, x_p)$

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

Karl Popper's definition of science is falsifiability. He believed that hypotheses aren't scientific if they are not falsifiable.

(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 said the probability was 0.12% because they used a bad model that had metrics that were too optimistic and the model lacked other important measures, resulting in very optimistic approximations.

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

Risk is something that can be measured and can be costly. Uncertainty is risk that is hard to measure. There is a vague awareness of potential risks but its magnitude and timing is unknown.

(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.

Out of sample is defined as a situation where an event or data that requires prediction ($\hat{\boldsymbol{y}}$) is tested by a predictive model (g) that is constructed from data ($\mathbb{D}$) that differs from the new event.

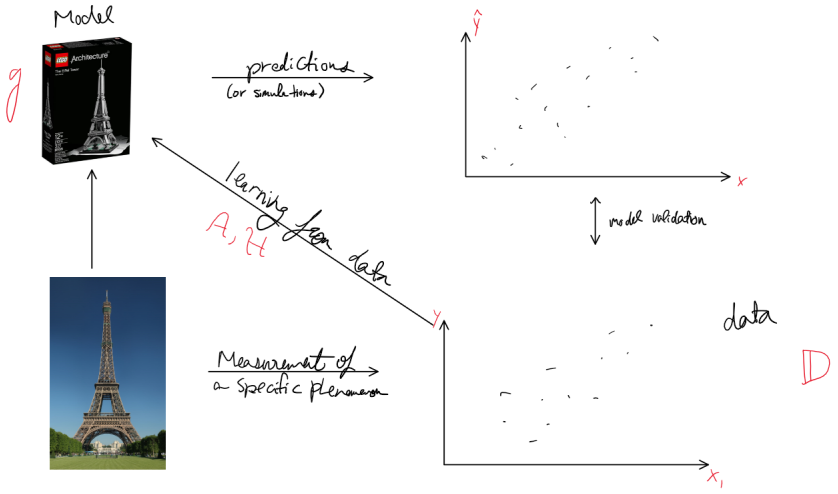
(j) [harder] Look up *bias* and *variance* online or in a statistics textbook. Connect these concepts to Silver's terms *accuracy* and *precision*. This is another example of Silver using non-standard terminology.

Variance measures how far data points are from the mean, which can be represented as Silver's definition of low precision. Bias is how much an estimator varies from the true value, as is Silver's definition of accuracy.

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 are values measured from reality.

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

Predictions are simulations based on previous data.

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

Models are by definition approximations, so they cannot be completely accurate, making all models "wrong".

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

They are useful because they can still be used to predict or explain reality.

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

A good model makes accurate predictions and generalizes the training data well, bad models do not.

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 because there's settings (eating an apple a day) that result in a phenomena (keeping the doctor away).

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

Eating an apple a day (diet)

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

Keeping the doctor away (health)

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

I think this is a bad model because there are many more factors affecting health than just a diet.

(e) [easy] Devise a metric for gauging the main input. Call this x_1 going forward.
x_1 is the diet measured by frequency of apple consumption.

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

y is health, measured by the number of doctor visits.

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

Output space. $\mathcal{Y} = \{0, 1\}$ where $y = 0$ means there were no doctor visits and $y = 1$ means there were doctor visits.

(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 properties of apples that benefit health.

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

$p = 1$

(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.

$\mathcal{X} = \{0, 1\}$ where $x_1 = 0$ means no apples are eaten that day and $x_1 = 1$ means that apples were eaten that day.

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

$y = f(x_1)$

It is approximate so it is better represented by $y = f(x_1) + \delta$

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

Supervised learning is when a model learns patterns in data and maps inputs to outputs so that it can make predictions on unseen data.

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

Supervised learning is empirical because it uses optimization to understand patterns instead of using an exact formula to make predictions.

(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.

$\mathbb{D} =< X, \vec{y} >$ where X is a 1 x n matrix populated by n measurements of x_1 and $\vec{y}$ is a vector of the y measurements.

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

$\mathcal{H}$ is the hypothesis space, it is the set of functions that may describe the relationship between x_1 and y. $\mathcal{A}$ selects the element of $\mathcal{H}$ that best fits the relationship.

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

The domain is all the datasets of $\mathbb{D}$ and the range is all possible models produced using $\mathcal{H}$.

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

Yes because $\mathcal{A}$ is a function that spits out the "best" element of $\mathcal{H}$, which is then assigned to 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 $\hat{y}^*$.

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

(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?

This is not a reasonable assumption because $\mathcal{H}$ assumes that the true relationship is perfectly represented by the models but this is not true because there are errors.

(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.

Estimation error $(h^*(x_1,\ldots,x_p)-g(x_1,\ldots,x_p))$, misspecification error $(f(x_1,\ldots,x_p)-h^*(x_1,\ldots,x_p))$, and ignorance error (δ). The sum of these three is called residual, or total error, represented by e. $\mathcal{E}$ is the sum of misspecification error and ignorance error.

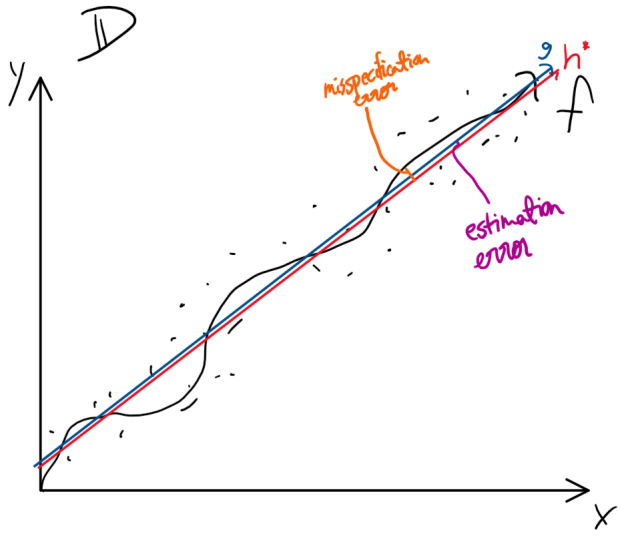
Estimation error is how well the algorithm accurately approximates the phenomenon. Misspecification error is error due to using a model that is different than the true relationship. Ignorance error is error due to missing information in the dataset.

$$y = g(x_1,\ldots,x_p)+\underbrace{(h^*(x_1,\ldots,x_p)-g(x_1,\ldots,x_p))+\underbrace{(f(x_1,\ldots,x_p)-h^*(x_1,\ldots,x_p))+\delta}_{\mathcal{E}}}_{e}$$

(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.

Reduce misspecification error by redefining $\mathcal{H}$ to be more expansive. Reduce estimation error by increasing n. Reduce ignorance error by gaining a better understanding of the phenomenon.

(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?
δ is missing because we don't know what we don't know.



(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 is a model based only on an output. It is a baseline and is used as a diagnostic tool to determine if a g is good since it must always beat g_0. It uses $\vec{y}$ and doesn't use any x's.

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

θ

(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?

$g_0 = \textbf{Mode}[\vec{y}] \in \{0, 1\}$

Since the outcome is binary, we want the null model to reflect the most frequent outcome (the mode).

(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?

$\mathcal{H} = \{\mathbb{1}_{x \geq \theta} : \theta \in \mathcal{X}\}$

The model parameter is θ

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

$$g = \begin{cases} 1 & \text{if } x \geq 0.5 \\ 0 & \text{if } x < 0.5 \end{cases} \tag{1}$$

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 says the scientific method is to make a guess, compute consequences, and then compare the model to nature/experiments. This is different than learning from data to build models because it doesn't use "nature" to inform the "guess".

(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)

Predictions/simulations

(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)

Reality

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

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?

A data scientist would agree, further commenting that there may be some missing metrics or lack of data that is causing this discrepancy.

(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?

We can't prove that a model is correct, we can only say that it is close enough to predicting reality. However, a model can be proven wrong if more data is collected that disproves predictions made by a model.

(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?

A vague mathematical model with metrics that aren't clearly defined can't be proven wrong since it is able to capture all possible measures and outliers. However, this isn't very helpful because it wouldn't be able to draw specific predictions of new data.

(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)?

Define exactly how much love is not enough and how much is overindulgent.

(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 know anything about a metric if it is not clearly defined because assumptions made would be too vague to be useful.

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.

The model is simple, I don't think it encompasses all or enough factors to determine behaviors and goals.

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

He says that his theory is wrong at least 25% of the time, and these limits are due to theory being a simplification of a complicated reality, since important factors may sometimes be left out. In the context of this class, he is saying that even the best models (h^*) are wrong because they are approximations to reality. These limits are caused by ignorance error (δ) when necessary metrics are overlooked.