MATH 390.4 Spring 2020 Homework #1

Tziporah Horowitz

Tuesday 11th February, 2020

1 Silver's Book, Introduction and Chapter 1

(a) [easy] What is the difference between *predict* and *forecast*? Are these two terms used interchangeably today?

Today, people use the terms predict and forecast interchangeably. However if traced back to their roots, the two words are used to describe different ways of foreseeing the future. According to Silver, forecasting typically involved planning under conditions of uncertainty, much like foresight, while prediction was based on fatalism and superstition.

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

In *Why Most Published Research Findings are False*, John P. Ioannidis confirms that many of the positive findings in peer reviewed laboratory experiments were likely to fail when applied in the real world. This implies that there is error in data validation.

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

The human being's most powerful defense is his ability to create models by observing features of reality and recognizing patterns that help him interpret his surroundings.

(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.

(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 describes a phenomenon with no error. In class we denoted it as:

$$y = t(z_1, z_2, \ldots, z_t)$$

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

According to Popper, theories can be considered scientific if they are falsifiable; meaning, they can be tested in the real world by way of 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 rating agencies failed to accurately predict the likelihood of a CDO defaulting because they used a crude model. They did not use the proper training data because they ignored the features that were hardest to model.

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

According to Silver, both risk and uncertainty are when the outcome depends on the assumptions and approximations you choose. The distinction between the two is in asserting the probability the outcome. Risk is something that you can put a price on while uncertainty is much harder to measure.

(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.

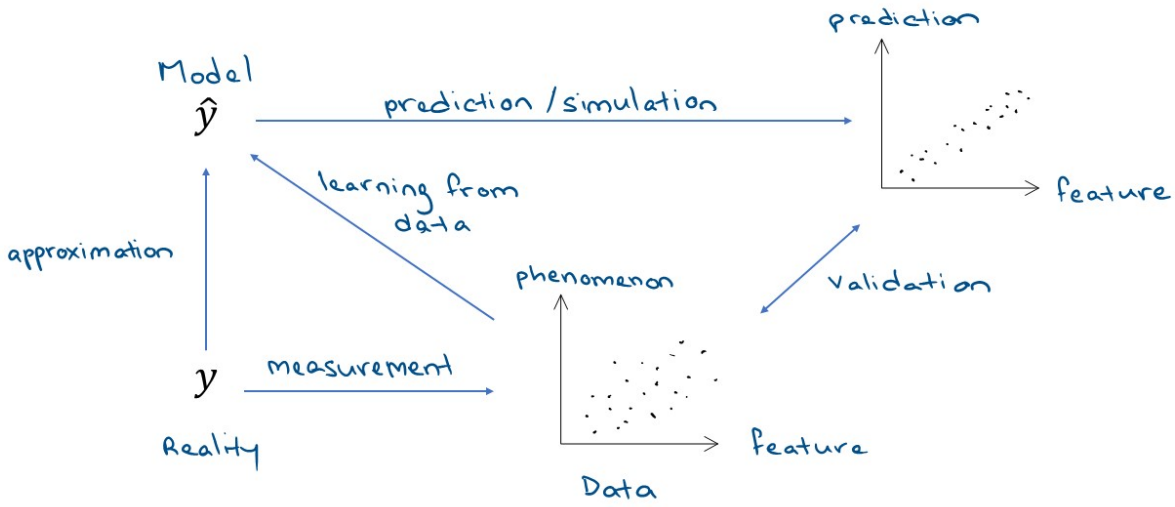
Silver defines *out of sample* as the misspecification of the training data, $\mathbb{D}$. If features, $x_k, \ldots, x_j$ are left out of the model, there will be a higher error due to ignorance, δ.

(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.

Statistical bias occurs when the expected value of the results differs from the true underlying quantitative parameter being estimated, i.e. the prediction is inaccurate. Variance is the measure of how far a set of numbers are spread out from their average value. The higher the variance, the less precise the prediction.

2 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 is the recorded result of the phenomenon based on the features measured.

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

Predictions are what will happen in the real world, given the conditions of the data.

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

All models are wrong because they are approximations of reality. The closer the approximation, the more useful the model.

(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 can explain and predict phenomena.

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

A good model uses metrics: it can successfully capture the phenomenon, is easily readable and unambiguous, has a good resolution, and is monotonic. A bad model is missing some or all of the metrics, making it hard to explain or predict the phenomenon.

3 The Framework of Modeling: "An Apple a Day Keeps the Doctor Away"

(a) [easy] Is this a mathematical model? Yes / no and why.

This is not a mathematical model. Mathematical models are ideas and abstractions that are formulated with measurements, not physical entities.

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

The input is eating an apple a day.

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

The output is remaining healthy.

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

This is not a good model because there are many other factors that health depends on.

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

x_1: ate 1 or more apples in 24 hours (0:00-23:59) $\in \{0, 1\}$

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

y: number of doctor's visits per year $\in \mathbb{N}_0$

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

$\mathbb{N}_0$

(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 the true causal inputs of phenomenon y.

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

Since there is only 1 feature, $p = 1$.

(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 ate at least 1 apple is 1 and ate no apples is 0.

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

Approximate relationship:

$$y = f(x_1) + \delta$$

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

Supervised learning is obtaining f by learning from data.

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

Supervised learning is an empirical solution because it approximates the model based on data. There is no analytical solution that can be used.

(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 the set of n historical examples of the simultaneous occurrence of x_1 and y such that:

$$\mathbb{D} = \langle \mathbf{X}, \mathbf{y} \rangle$$

Where $\mathbf{X}$ is the matrix of inputs and $\mathbf{y}$ is the vector of outputs.

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

$\mathcal{H}$ is the set of candidate functions for f such that $f : \mathbf{X} \rightarrow \mathbf{y}$ and $\mathcal{A}$ is an algorithm that takes in data $\mathbb{D}$ and set $\mathcal{H}$, and produces a model g such that $g = \mathcal{A}(\mathbb{D}, \mathcal{H})$.

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

The domain is $\mathbb{R}$ and the range is $\mathbb{N}_0$.

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

$g \in \mathcal{H}$ because g is the best candidate function that $\mathcal{A}$ can produce and $\mathcal{H}$ is the set of all candidate functions for the model.

(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 $\hat{y}^*$.

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

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

No, because there is error in modeling.

(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.

$$y = g(\vec{x}) + \big(h^*(\vec{x}) - g(\vec{x})\big) + \big(f(\vec{x}) - h^*(\vec{x})\big) + \big(t(\vec{z}) - f(\vec{x})\big)$$

(a) **Error due to ignorance**: $\delta = t(\vec{z}) - f(\vec{x})$, the difference between reality and the approximation of f.

(b) **Misspecification Error**: $f(\vec{x}) - h^*(\vec{x})$, the difference between the best function and the closest possible model.

- $\mathcal{E} = \big(f(\vec{x}) - h^*(\vec{x})\big) + \big(t(\vec{z}) - f(\vec{x})\big)$

(c) **Estimation error**: $h^*(\vec{x}) - g(\vec{x})$, the difference between the closest possible model and the best model that algorithm $\mathcal{A}$ finds.

- $e = \big(h^*(\vec{x}) - g(\vec{x})\big) + \big(f(\vec{x}) - h^*(\vec{x})\big) + \big(t(\vec{z}) - f(\vec{x})\big)$

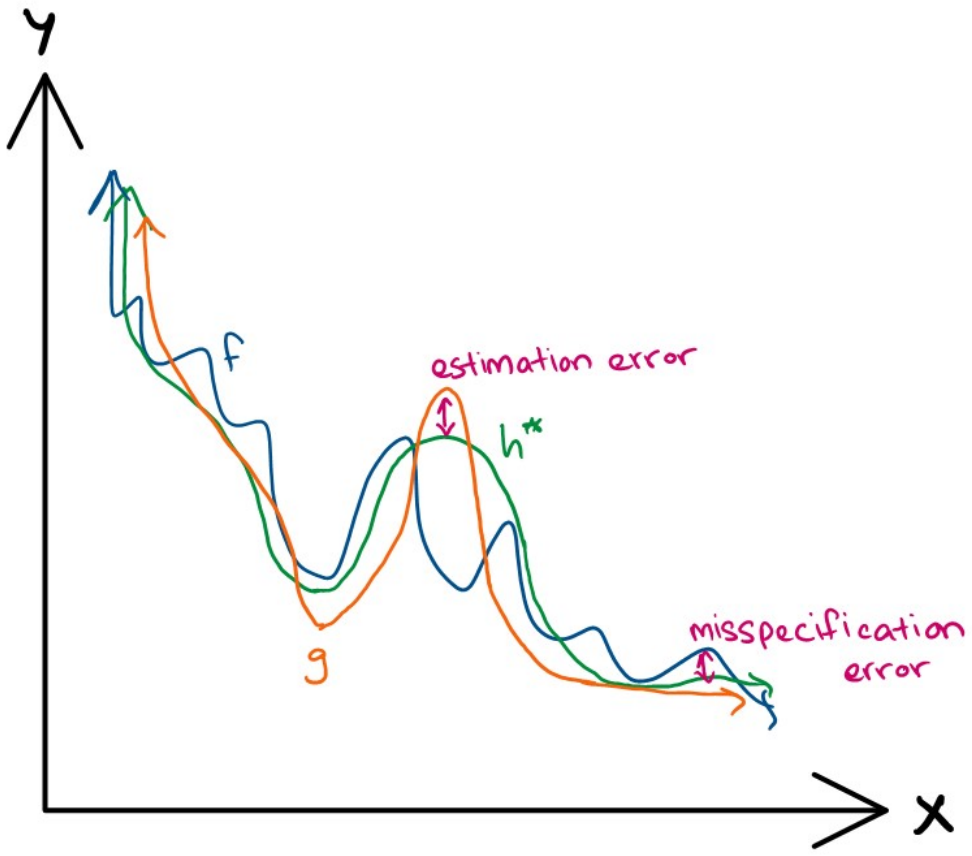
(u) [easy] In the general modeling setup, for each of the three source of error, explain what you would do to reduce the source of error as best as you can.

(a) **Error due to ignorance**: measure more x_j's of the units that contain information about $\vec{z}$.

(b) **Misspecification Error**: expand $\mathcal{H}$ to include more complicated functions.

(c) **Estimation error**: increase the 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?



The error missing from the picture is the error due to ignorance. This is because the true function, $y = t(z_1, \ldots, z_t)$ is unknown.