MATH 390.4 / 650.2 Spring 2018 Homework #1t

Bryan Lliguicota

Sunday 25th February, 2018

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?

Predict and *forecast* are largely used interchangeably today, but they originally had different meanings. A *prediction* was mostly associated with what a person who supposedly can foresee the future would tell you. While a *forecast* typically implied planning under conditions of uncertainty, involved wisdom, cautiousness and diligence.

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

John P. Ioannidis found that most successful predictions of medical hypothesis carried out in laboratories would fail when applies to the real world. This implies that even today with the amount of "big data" in our possession, our models of the real world are inaccurate.

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

Our most powerful defense lies in our ability to make quick decisions. In looking at collections of data we are able to find patterns and relationships. Note: He also linked this as being a negative saying that we tend to look for patterns in places where their are simple no patterns.

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

Our understanding of how to process it.

(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 is reality, we cannot model it. Its the true relation between the true casual inputs and the true outcome.

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

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

A hypothesis is not scientific unless it is feasible. Meaning that it can be tested in the real world by means of a 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 fault lied in the forecasters model of the world. The rating agencies had no historical data of the new highly novel securities. The default rates claimed by the $S\&P$ were not derived from historical data but instead where assumptions based on statistical models. Their $\mathcal{A}$ algorithm used to pick the best candidate function was flawed, they did not have a good enough $\mathbb{D}$ (data-set) to work with.

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

risk is somthing that is measurable, you know your odds and so you can plan ahead. Uncertainty on the other hand is risk that is hard to measure, you might have a vague idea of the risk but you can not quantify it.

(i) [difficult] How does Silver define *out of sample*? 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. WARNING: Silver defines *out of sample* completely differently than the literature (and differently than practitioners in industry). 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 assuming something based from false or non related preconceived notions. Going back his example of the driver who had 20,000 trips with only 2 minor accidents. Now the trips count as historical data, that can be used to perdict the outcome of a future trip. Each past trip has its own set if $x_{1\cdot}, \ldots, x_{n\cdot} \in \mathcal{X}$, but their is not a single features in the covariant space where he is drunk. Therefore if if think of all his trips as sample points, we can say that there are no sample points in the sample space(all past trips) that involve him being drunk, therefor this is an out of the sample situation.

(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.

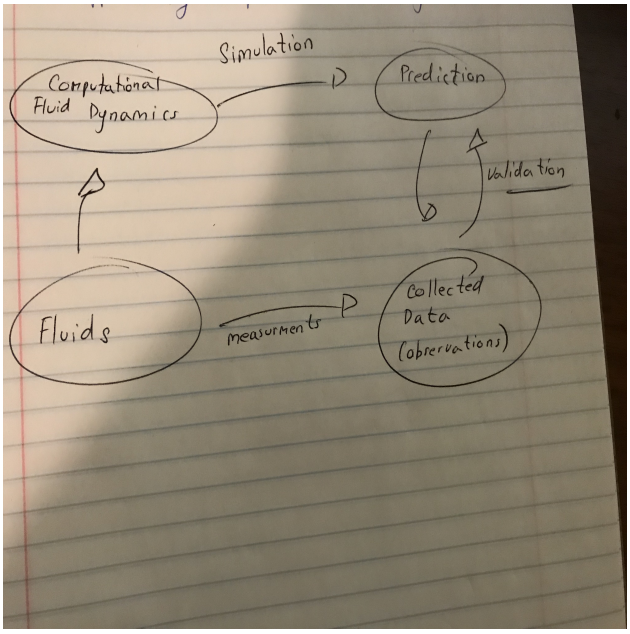
Variance can be linked to Silvers use of the term accuracy, Silver describes accuracy as the distance the bullet wholes are from its target. Accuracy can be measured as the sum of distances from the target (mean) divided by the number of bullet wholes. This reflects the definition of variance as the target can be tough of as the mean (μ) and the bullet whole is the realization of you random variable X. Precision on the other

hand is viewed in how close all you bullet holes with relation to each other and bias is the tendency to overestimate or underestimate a parameter, you model can be precise but the amount of bias will make it stray away from the optimal result.

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. In the top right quadrant, you should write "predictions" not "data" (this was my mistake in the notes). "Data / measurements" are reserved for the bottom right quadrant. The quadrants are connected with arrows. Label these arrows appropriately as well..



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

Data is what the reality "the world" gives us, we use it as a reference point for our predictions. The output data from our model will be compared to the data given to use from reality so we may measure its accuracy.

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

perdictions are what we obtain from our model when we input our data set $\mathbb{D}$. These predictions may or may not match the true outputs. That's where your measurements come into play to see how far you strayed away from the true output.

(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 casual inputs. Their will always be an error due in ignorance in out algorithms and independent variables, as well estimation error and misspecification error.

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

some models are useful b/c they fit to a acceptable degree the real world data we collected. In other words the data obtained from out model is close enough to the data from the real world.

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

The difference from a "good model" and "bad model", is basically how close our y_i is from $\hat{y}_i$. This can be due many factors such as out data set $\mathbb{D}$ and out chooses from the input space $\mathcal{X}$, they may stray away to much from the true causal inputs $\mathcal{Z}$

Problem 3

We are now going to investigate the aphorism "An apple a day keeps the doctor away". We will use this as springboard to ask more questions about the framework of modeling we introduced in this class.

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

This is a bad model because our set of independent variables are stray to much from the true casual inputs.

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

Yes it can be turned into a mathematical model based on how we give a numeric value to the outcome and inputs.

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

In the model the inputs are the consumption of apples in a day, or average in a range of days.

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

The output is whether or nor the person is healthy. How we will measure "health" is something else.

(e) [easy] Devise a means to measure the main input. Call this x_1 going forward.

Based on a range of days, lets say 5 on average how many apples did the individual eat.

(f) [easy] Devise a means to measure the main output. Call this y going forward.

The persons health, based on whether or not he/she saw the doctor.

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

1 - has not seen the doctor
0- has seen the doctor

(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 casual inputs, ant t is the true relation function b/w the true inputs and outputs.

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

p represents the dimension we are working in. Since we only observe one independent variable x_1, our p is only one.

$$x_1 \in \mathbb{R}^{p=1}$$

(j) [harder] From this point on, you only observe x_1 is in the model. 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 our input space, a countable infinity set of independent variables, some of which we can observe.

$$x_1 \in \mathcal{X}$$

Because we are dealing with mathematical models if x_1 is not numeric, it will have to be altered. This is where bias may come into play, if x_1 is categorical suppose n categories, than we can associate each category with a number $1...n$. This has its downside, b/c we may be giving one category more importance than the other.

(k) [harder] How did we term the functional relationship between y and x_1?

$$\begin{aligned} y &\approx f(x_1) \\ y &= f(x_1) + \delta \\ \text{where } \delta &\equiv t(z_1) - f(x_1) \text{ is error due to ignorance} \end{aligned}$$

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

Supervised learning is a method used to train our learning algorithm. We are given historical data and the true causal output for a set of x in the data. We use this data to alter out weights to obtain an optimal solution.

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

supervised learning is an empirical solution b/c we going through many iterations

to obtain an output based on a dataset. The output is gained through "observations/experiences", while a analytical solution involves proofs and a series of logical steps.

(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 composed of the training data, if we are only observing one independent variable x_1 we can have it be whether or not they had an apple, or how many apples they consumed on average during a given time period. Each y associated with an x_1 could be the persons health. There are many ways to measure a persons health, one could be the number of times he/she went to the doctor in the time frame specified above.

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

$\mathcal{H}$ is the set of candidate functions, because we are dealing with only one variable I think it should be composed of threshold parameters. Maybe eating a certain amount of apple a day for a given period on average really does keep you healthy. If that's the case, what we want to know is what is that range. To much of something is bad and so is to little, finding the sweet spot would mean have 2 threshold values to indicate your range.
$\mathcal{A}$ is the algorithm we use to find an optimal function given $\mathbb{D}$ and $\mathcal{H}$

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

Based on my response to the previous question, the domain $(\mathbb{D}, \mathcal{H})$ would be the training data and set of candidate functions. The range would be a tuple of two values indicating a low and high for the number of acceptable apples one can eat and be healthy(say away from the doctor as much as possible).

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

Yes, $g \in \mathcal{H}$. This doe not necessarily mean that we choose the best candidate function. The best possible function in $\mathcal{H}$ is $h^\star$, but it is unlikely we choose it.

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

No it not reasonable to assume $f \in \mathcal{H}$ since this would mean that we obtain the best possible function that approximates the real world. What is reasonable to assume is the $h^* \in \mathcal{H}$, where h^* is the best approximation of $f \in \mathcal{H}$.

(t) [easy] If $f \notin \mathcal{H}$, what are the three sources of error? Write their names and provide a sentence explanation of each. Note that I made a notational mistake in the notes

based on what is canonical in data science. The difference $t - g$ should be termed e as the term $\mathcal{E}$ is reserved for $t - h^*$.

When $f \notin \mathcal{H}$ the three possible errors are:
1. $(t(\vec{z}) - f(\vec{x}))$ Denoted as error due to ignorance. This error will always exist, and the best we can do is try to minimize it. It represents the difference b/w the true relation function t and the best approximation function f.
2. $(f(\vec{x}) - h^*(\vec{x}))$ Denoted as the misspecification error; this is the difference b/w out best approximation of the real world function f and the best candidate function $h^* \in \mathcal{H}$.
3. $(h^*(\vec{x}) - g(\vec{x}))$ Denoted as the estimation error, it says that even if the best candidate function is our set, their is still a good chance that we may not even pick it. Instead the candidate function we end up with is g. The estimation error is their to tell us how far we are from the best candidate function h^*.

(u) [harder] For each of the three source of error, provide a means of reducing the error. We discussed this in class.

- For error 1 (error due to ignorance) we can improve it by increasing the amount of data we have and improving or independent variables we choose from the input space.
- For error 2(misspecification error) This can be improved by increasing our set of candidate functions, as well as improving our algorithm.
- For error 3(estimation error) this error can be decreased by improving out algorithm $\mathcal{A}$

(v) [easy] Regardless of your answer to what $\mathcal{Y}$ was above, we now coerce $\mathcal{Y} = \{0, 1\}$. If we use a threshold model, what would $\mathcal{H}$ be? What would the parameter(s) be?

Demote the threshold as b:

$$\hat{y}_i = 1 \; iff x_i \geq b$$
$$\hat{y}_i = 0 \; iff x_i < b$$

We would have $\mathcal{H} = \{\mathbb{1}_{\vec{w} \cdot \vec{x} \geq b} : x \in \mathbb{R} \wedge w \in \mathbb{R}\}$

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

Lets say the best number of apples to eat in 5 days is 7. On average people who ate more than seven apples in the course of five days lived healthier than those who didn't. Therefore out threshold is 7, anything 7 or greater will be classified as 1, anything less than 7 will be a 0. Now if we introduce a person who has eaten 100 apples in 5 days (with an average of 20 per day) we will classify him/her as a 1 (healthy).

Problem 4

These are questions about the linear perceptron. This problem is not related to problem 3.

(a) [easy] For the linear perceptron model and the linear support vector machine model, what is $\mathcal{H}$? Use b as the bias term.

In both the linear perceptron model and the linear support vector machine $\mathcal{H}$ is considered the set ("pool") of candidate functions (linear classifiers) .For the perceptron model $\mathcal{H}$ is the set of indicator functions which give out a binary response of 0 or 1 based on the dot product of the weight vector $\vec{w}$ and independant variable $\vec{x}$ (the bias b is w_0 in the weight matrix). For the support vector machine $\mathcal{H}$ consists of linear functions that meet the condition $\forall_i \ (y_i - \frac{1}{2})(\vec{w} \cdot \vec{x}_i + b) \geq \frac{1}{2}$

(b) [harder] Rewrite the steps of the *perceptron learning algorithm* using b as the bias term.

Let $\mathcal{X} = \mathbb{R}^d$ be the input space, where d represents the dimension we are working in. The space $\mathcal{Y} = \{1, 0\}$, is the output space denoted by a binary decision. $x_{1.}, \ldots, x_{d.} \in \mathbb{R}^d$ correspond to features in our supervised learning algorithm(the perceptron).We will have a hypothesis set of candidate functions $\mathcal{H}$, as well as weights $w_{1.}, \ldots, w_{d.} \in \vec{w}$. Our threshold is another factor that is taken into account, the threshold is denoted as b (bias).

$$\hat{y}_i = 1 \ iff \sum_{i=1}^{d} w_i \cdot x_i > b$$

$$\hat{y}_i = 0 \ iff \sum_{i=1}^{d} w_i \cdot x_i < b$$

The bias can show how lenient to strict the model is. The perceptrons learning algorithm $\mathcal{A}$ will search $\mathcal{H}$ for weights and bias that perform well on the data set. The optimal weights and bias define the final hypothesis $g \in \mathcal{H}$. Suppose we are working in a 2 dimensional space, our parameters w_1, w_2, b correspond to different lines $w_1x_1 + w_2x_2 + b = 0$. We can simplify the notation of the perceptron formula by letting the bias b be the weight $w_0 = b$, we will need to increment the number of elements in $\vec{x}$ in order to do the dot product, let $x_0 = 1$. With all the notation done we now move to the perceptron learning algorithm PLA. The PLA determines $\vec{w}$ based on historical data we provide it with $\mathbb{D}$. Assuming the data is linear separable, we use a simple iterative method to determine the optimal $\vec{w}$. During the i_{th} iterative step the current weight vector is w_i. The algorithm picks a point from $\mathbb{D}$ lets say (x_i, y_i) and uses it to output a $\hat{y}_i$ that is compared to y_i,based on the difference w_{i+1} is modified.

$$w_{i+1} = w_i + (\hat{y}_i - y_i)x_i$$

(c) [easy] Illustrate the perceptron as a one-layer neural network with the Heaviside / binary step / indicator function activation function.

A threshold determines if a neuron fires. A threshold is a type of activation function, an activation function determines the range of values that will cause the neurons to fire. The threshold is a simple activation function, in our case an indicator function will be used as a activation function that fires either a 0 or a 1 based on input values.

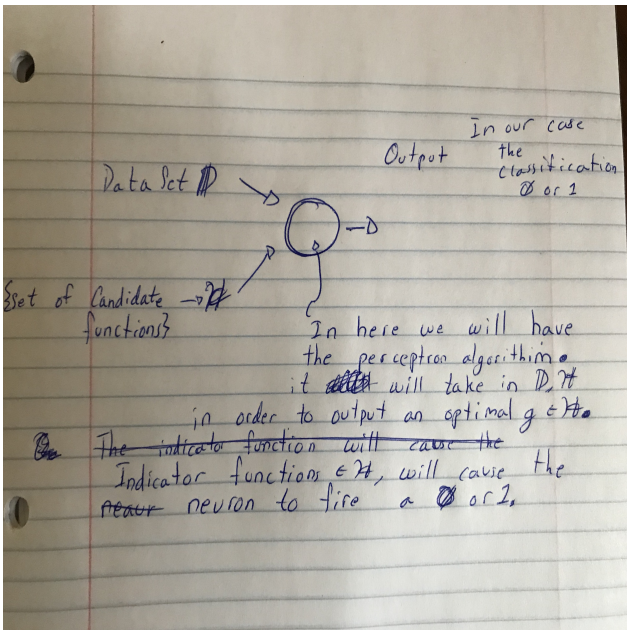


Figure 1: 1 layered N.N.

(d) [easy] Provide an illustration of a two-layer neural network. Be careful to indicate all pieces. If a mathematical object has a different value from another mathematical object, denote it differently.

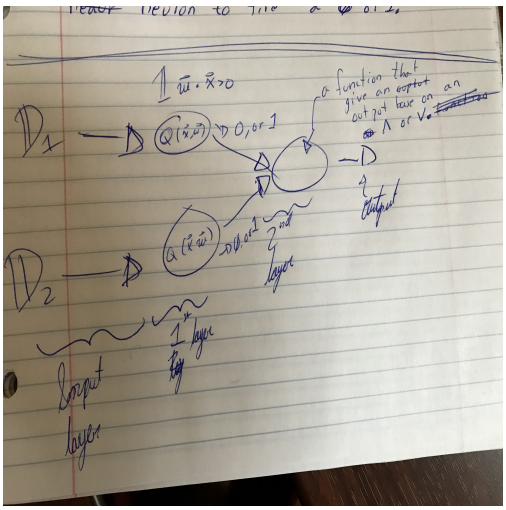


Figure 2: 2 layered N.N.