MATH 390.4 / 650.2 Spring 2018 Homework #1t

Darshan Patel

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 interchangeably today?

Predict has to do with taking a sign of something and interpreting them to gain some advantage. On the other hand, *forecast* deals with making plans under what's certain and uncertain. These two terms are used interchangeably today but not back in Shakespeare's time.

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

Ionnidis found that the positive findings found by successful predictions for medical hypotheses did not replicate itself in the real world. Whatever was predicted was done in a closed lab environment. The implication is that predictions cannot be accurate each time around.

(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, according to Silver, is its ability to survive using its minds. We use prior knowledge to create responses for new data, or threats.

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

The amount of useful information is not increasing. Whatever new information is earned is mostly noise.

(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.

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

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

Science is a way of testing predictions in the real world.

(g) [harder] Why did the ratings agencies say the probability of a CDO defaulting was 0.12% instead of the 28% that actually occurred? Answer using concepts from class.

The rating agencies downplayed the probability of a CDO defaulting because it was using default rates from assumptions from a bad model. In other words, the model they used did not capture the true phenomenon, even come close to.

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

According to Silver's definitions, *risk* is some probability of an event you can calculate of occurring. On the other hand, *uncertainty* is an unmeasurable risk. It is unknown as to what is the chance of something occurring or even if it will or will not occur at all.

(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* to be something that happens unpredictably. In other words, *out of sample* is a parameter estimation error where $g \neq h^*$, or $h^*(\vec{x}) - g(\vec{x})$.

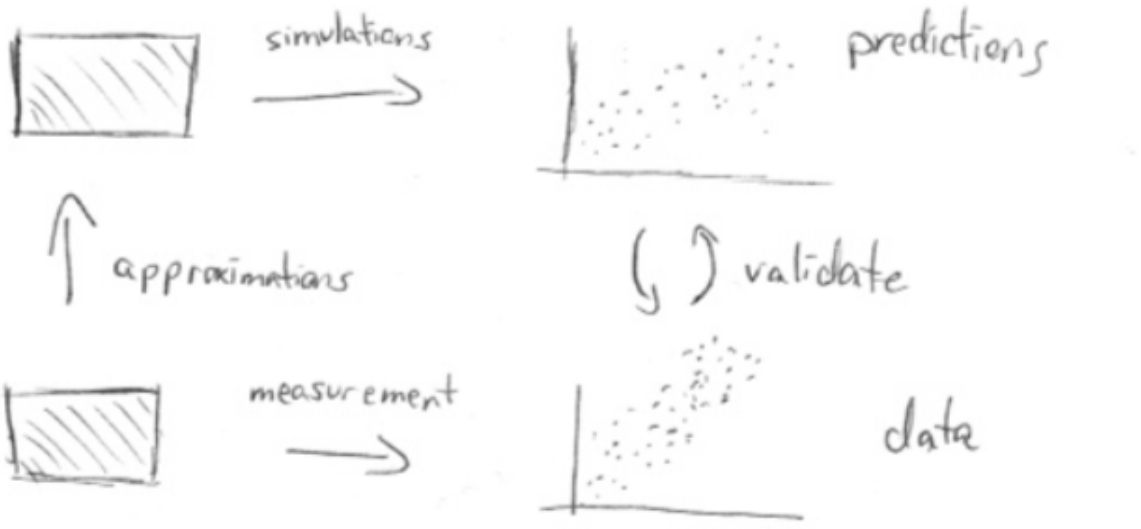
(j) [harder] Look up *bias* and *variance* online or in a statistics textbook. Connect these concepts to Silver's terms *accuracy* and *precision*. This is another example of Silver using non-standard terminology.

According to *Wikipedia*, *variance* measures the spread of (random) numbers from their average value whereas *bias* is the difference of the expected value of the captured results and the parameter being estimated. This connects to Silver's terminology of *accuracy* and *precision*. He demonstrates *accuracy*, using dartboards, showing how all darts are close to the center. This ties to the term *bias*, where all values seem to be closely estimated to one certain parameter. Silver also demonstrates *precision*, showing how all darts hit the dartboard in small circle albeit not at the center. This ties to the term *variance* where values captured do not differ much amongst its average value.

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 defined to be the normal "measured" result of a phenomenon.

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

Predictions is defined to be a probable outcome to a certain situation.

(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're not reality.

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

"All models are wrong but some are useful" because they serve as some useful function even though it's not reality.

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

A "good model" is one that creates predictions that are close to the data. A "bad model" is one where the predictions created is far from the data.

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 model is "bad" because it does not take into consideration of other factors that can keep the doctor away or closer. An apple a day may have no role at all.

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

This is a mathematical model because both the input and outputs can be measured numerically.

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

The input of this model is how many apples a person consumes.

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

The output of this model is whether or not a person has seen a doctor recently.

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

x_1 can be measured by indicating how many apples a certain person consumes in a day.

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

y can be measured by indicating whether a certain person has seen a doctor in the past 1 month.

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

$$\mathcal{Y} = \{0, 1\}$$

(h) [easy] Briefly describe $z_1, \ldots, z_t$ in English where $y = t(z_1, \ldots, z_t)$ in this *phenomenon* (not *model*).

$z_1, \ldots, z_t$ is the causal input information that can create a model of y by some function t.

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

$$p = 1$$

(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.

$$\mathcal{X} = x_1$$

where x_1 is the number of apples a person eats in a day.

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

The functional relationship between y and x_1 is f, where $y = f(x_1)$.

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

Supervised learning is learning that uses 3 things. The first thing is training data, a set of input values that can give an insight into the output. The second thing is a family of functions that can approximate the functional relationship. Using this can help to narrow down f from an infinite number of function to a certain type, such as step functions or linear functions. The third thing is an algorithm to help pick out that function.

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

Supervised learning is an *empirical solution* because it creates a model that can be deduced from inputs. It is not an *analytical solution* because there is no way a function can be attained by minimizing or maximizing the inputs.

(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.

$$\mathcal{D} = \{\langle x_1, y_1\rangle, \langle x_2, y_2\rangle, \ldots, \langle x_n, y_n\rangle\}$$

This is a set of input and outputs where the x's represent how many apples person i ate and y_i indicates whether or not the person went to the doctor in the past 1 month.

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

The role of $\mathcal{H}$ is to narrow down the infinite number of math functions to a certain family that can be used to approximate f. The role of $\mathcal{A}$ is to enable a mechanism of finding that special function f.

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

The domain of g is $\mathbb{R}$ and the range of g is $\{0, 1\}$, where 1 represents a person not seeing a doctor in the past month and 0 represents a person seeing a doctor in the past month.

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

$g \in \mathcal{H}$; it is a type of threshold function that represents how apple eaten affects the occurrence of medical visits .

(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?

It is reasonable to assume $f \in \mathcal{H}$ because it would mean the appropriate model was classified correctly.

(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 $t - h^*$.

Sources of Error:

(a) Error due to ignorance: this is created from ignoring values when estimating t (the true relation) by estimating f

(b) Error due to misspecification: this is created when $f \notin \mathcal{H}$ thus making inaccuracies in predicted values from the model and the true values

(c) Error due to parameter estimation: this is created from the random chance that occurs when selecting $h^* \in \mathcal{H}$

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

Reducing Errors:

(a) Error due to ignorance: no solution. $f(x)$ is the best we can do and it will differ from $t(z)$, the true relation

(b) Error due to misspecification: use another algorithm $\mathcal{A}$

(c) Error due to parameter estimation: increase size of population

(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?

$\mathcal{H}$ would be a family of step functions where the parameter x_i represent how many apples a person ate and $f(x_i)$ will be whether he/she/they met a doctor or not.

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

An example of g is:

$$g(x_i) = \mathbb{1}_{x_i > 1}$$

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.

$$\mathcal{H} = \left\{ \mathbb{1}_{\vec{w} \cdot \vec{x}_i > 0} : \vec{w} \in \mathbb{R}^{p+1} \right\}$$

This is a set of indicator functions that is derived using the dot product of $\vec{w}$ (the estimated parameters) and $\vec{x}_i = \begin{bmatrix} b & x_{i1} & x_{i2} & \dots & x_{ip} \end{bmatrix}$, the set of values to approximate y. Note that b is the bias term.

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

Steps of the Perceptron Learning Algorithm:

(a) Initialize $\vec{w}^{t=0} = \vec{0}$ or random

(b) Calculate $\hat{y}_i = \mathbb{1}_{\vec{w}^t \cdot \vec{x} > 0}$

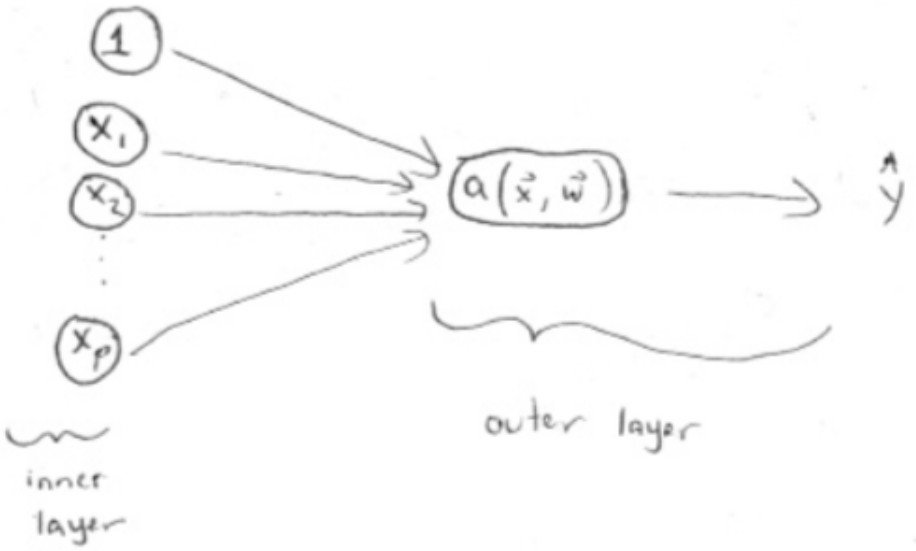
(c) Update all weights from $i = 1, \dots, p+1$

$$\begin{aligned} w_1^{t=1} &= w_1^{t=0} + (y_i - \hat{y}_i) b \\ w_2^{t=1} &= w_2^{t=0} + (y_i - \hat{y}_i) x_{0,1} \\ &\vdots \\ w_{p+1}^{t=1} &= w_{p+1}^{t=0} + (y_i - \hat{y}_i) x_{p+1,1} \end{aligned}$$

(d) Repeat steps b and c for all $i \in \{1, \dots, n\}$

(e) Repeat steps b through d until a threshold error is reached or a maximum number of iterations.

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



(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.