MATH 390.4 / 650.2 Spring 2018 Homework #1t

Joseph Peltroche

Tuesday 27th 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?

Solution. The difference between *predict* and *forecast* lies in their meaning. Forecast typically implies planning under conditions of uncertainty, suggesting prudence, wisdom, and industriousness. On the other hand, prediction was what something a soothsayer would say, a person able to foresee the future; It does not originate from a sense of planning, but has already been set in stone. Today, they are used interchangeably. □

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

Solution. Joan P. Ioannidis's findings are that most of the positive findings in journals, the descriptions of successful predictions of medical hypotheses carried out in laboratory experiments, were likely to fail in the real world. This implied that, even with "Big Data", there are bound to be pitfalls and we can regressfor some time, but that "Big Data" will eventually produce progress. □

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

Solution. As human beings, our most powerful defense is our ablilty to model the reality around us. We observe features of our reality and develop an algorithm to recognize patterns and develop an abstraction of the world around us that may or may not be close to the actual truth. □

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

Solution. Despite information increasing at a rapid rate, the amount of useful information barely changes. □

(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.

Solution. The objective truth, using the notation from class, is represented by y. Our predictions, $\hat{y}$, are what we compare to the truth. The objective truth, y is described by the "true model", including true causal inputs:

$$y = t(z_1, z_2, z_3, \ldots, z_n)$$

□

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

Solution. Science can be tested in the real world by means of a prediction, and such theories that are not falsifiable are therefore unscientific or that they lack any value. □

(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.

Solution. The ratings agencies depended on a faulty model to create wildly off predictions. The deafult rates claimed by S∧P were not derived from a trainning data, but on assumptions by this faulty model which were far from the truth. These rating agencies also had no successful experiences in this realm, they did not have a track record, so there was even more pressure on this model that produced default rates more than two hundred times higher than S∧ predicted. □

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

Solution. *Risk* is the something you can put a price on, so the probablity and the odds of a certain situation. However, *uncertainty* is a vague awareness of in asserting the real odds of a situation. □

(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.

Solution. Silver defines *out of sample* in the following way:
Obtaining historical data and extracting the features $x_1, \ldots, x_n$ to be used in the training data $\mathbb{D}$. The problem is, the features $x_1, \ldots, x_n \notin \mathbb{D}$, that is, the features we have collected from the past data are irrevelant to the features we actually need and therefore cannot be used in the training data to create predictions. This effectively brings our $p = 0$ since there are no features $x_1, \ldots, x_n$ to use in our training data (no sample, $n = 0$). □

(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.

Solution. Bias is connected to accuracy and precision related to variance. Bias, being the tendency of a measurement process to over- or under- estimate the value of a population parameter, is related inversely related to the accuracy of a prediction. Variance, measuring how far a set of points are spread out from their average value, is related to to the precision of the amount of points (on a dartboard, as Silver illustrates on page 46) □

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. 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..

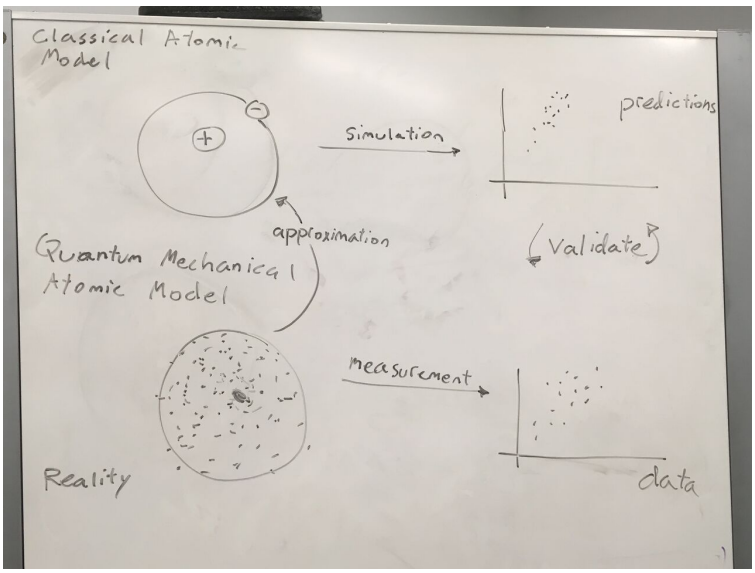


Figure 1: The figure depicts the classical and quantum representation of the atomic model. The classical atomic model is studied at an early age but is not the entire picture, it really is just an approximation to the reality of the quantum mechanical atomic model.

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

Solution. "Data" is the result of the system or phenomena which is" measured" □

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

Solution. "Predictions" is what will happen in the real world under a set a conditions. □

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

Solution. All models are wrong because they are merely an abstraction or an approximation to the reality/truth. They simply aren't exact. □

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

Solution. Although no models are exact, some are useful by the explanations they provide and the predictions that can be formulated by it. □

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

Solution. If the predictions from a model are "close" when compared to the measured data then the model is referred to as a "good model". If comparisons are far-off then this would be a "bad model" □

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?

Solution. In my opinion, the model is a very bad one. The model predicts an apple a day will keep the doctor away, implying good health. However, by looking back on my personal history and the history of the impact of medicine, the reality is that an apple has never been sufficiently enough to constitute good health. The inputs and outputs are also unobservable for a general public. □

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

Solution. This is not a mathematical model. It does not have any numbers or numerical measurements. The idea of health is vague in this model and it has no numerical representation. □

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

Solution. The input to this model is an apple per day (1 apple per 24 hours) □

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

Solution. The output is good health. □

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

Solution. In order to measure the main input, we will observe the amount of apples bought per year (per person): x_1. □

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

Solution. To measure good health, a controversially hard thing to measure, we will measure the amount of visits to a hosital per year (for health-related issues),y. □

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

Solution. $\mathcal{Y}$ is the set of all real numbers, $\mathbb{R}$. □

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

Solution. The inputs $z_1, \ldots, z_t$ are known as the "true" causal inputs, which are ideal inputs to incorporate for the obtaining the truth in the "true model" or phenomena $y = t(z_1, \ldots, z_t)$. Unfortunately, these inputs are usually unobservable and inaccessible. □

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

Solution. Since x_1 is the only feature we are observing, then $\vec{x}$ has one dimension. Hence, $p = 1$ □

(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.

Solution. The feature x_1 a value of the real numbers, neither categorical or binary. This feature is contained within the set of the real numbers, therefore $\mathcal{X} = \mathbb{R}$. □

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

Solution. The functional relationship between y and x_1 is known as the approximation to y, or the model to y. □

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

Solution. Supervised learning is a method of obtaining f or a function close to f by using and learning from historical data and records. □

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

Solution. Supervised learning is an empirical solution because it relies on previously recorded data, rather than an analytical expression. For such modeling, there exists no analytical solution. □

(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.

Solution. The trainning data $\mathbb{D}$ would be equal to $< X, \vec{y} >$, where X is a $n \times 1$ matrix (really a vector) consisting of the feature in consideration (amount of apples bought per year) for n-people and $\vec{y}$ would be a vector of length n consisting of the "true" value for every nth person (the true amount of hospital visits per year). □

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

Solution. $\mathbb{H}$ is the set of all candidate functions for f. From $\mathbb{H}$ a best candidate function will be chosen using our algorithm $\mathbb{A}$ as a function and $\mathbb{D}$ as another input. $\mathbb{A}$ will be in charge taking in the trainning data $\mathbb{D}$ and $\mathbb{H}$ to produce a function g. □

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

Solution. The domain and range of g would be the set of real numbers $\mathbb{R}$, since both the feature and the output space is the $\mathbb{R}$ □

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

Solution. g is within $\mathbb{H}$ since $\mathbb{H}$ is the set of all candidate functions and g is the best candidate function the model can produce. □

(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 $\hat{y}^*$.

Solution.

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

□

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

Solution. It is unreasonable to assume $f \in \mathcal{H}$ because there is a strong likelihood for a soruce of error, whether it may be the algorithim or the population size. □

(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 $t - h^*$.

Solution. The three sources of error are as follows:

(a) Estimation error $(h^*(\vec{x}) - g(\vec{x}))$: This error is due to the difference from the best approximation to f that is within $\mathcal{H}$, h^*, and the produced function g.

(b) Misspecification error $(f(\vec{x}) - h^*(\vec{x}))$: Error due to the difference between the best appoximation to f in $\mathcal{H}$ and the actual function f itself.

(c) Error due to ignorance $(t(\vec{z} - f(\vec{x}))$: Error due to the difference between the approximation f to the reality t.

□

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

Solution. For estimation error, it is best to increase the population size to reduce the difference (just like the normal approximations in math 241). For misspecification error, there must be improvements to the algorithm. Finally, for error due to ignnorance, a solution can be to obtain better data. □

(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?

Solution. $\mathcal{H}$ would be the set of all indicator functions and there would only be one parameter belonging to the set of real numbers x_T, the threshold (if we use the threshold model). □

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

Solution.

$$g = argmin \left\{ \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \right\}$$

□

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.

Solution.

$$\mathcal{H} = \left\{ \mathbb{1}_{x_2 > a + bx_1} : \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R} \right\}$$

□

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

Solution.

$$\begin{aligned} \mathcal{H} &= \left\{ \mathbb{1}_{x_2 > a + bx_1} : \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R} \right\} \\ &= \left\{ \mathbb{1}_{-ax_2 + x_2 > 0} : \begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R} \right\} \\ &= \left\{ \mathbb{1}_{w_0 + w_1 x_1 + w_2 x_> 0} : \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} \in \mathbb{R} \right\} \end{aligned}$$

$$= \left\{ \mathbb{1}_{w_0 + \vec{w} \cdot \vec{x} > 0} : w_0 \in \mathbb{R}, \vec{w} \in \mathbb{R}^2 \right\} \qquad \text{(Let } \vec{x} = [1, x_1, x_2]\text{)}$$
$$= \left\{ \mathbb{1}_{\vec{w} \cdot \vec{x}} : \vec{w} \in \mathbb{R}^3 \right\}$$

From this model of using weights $\vec{w}$, we can proceed to explain the precetron learning algorithm. The perceptron learning algorithm is as follows:

(a) Intialize $\vec{w} = 0$ or a random quantity (random line).

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

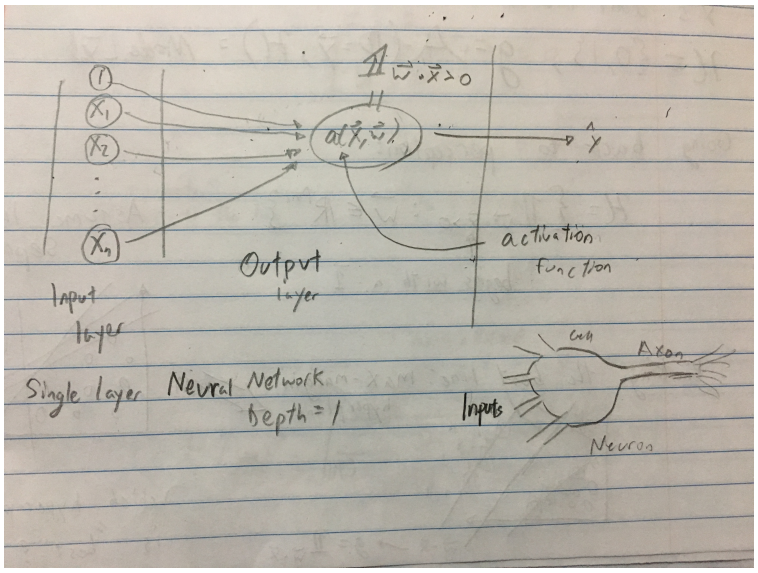
(c) Update all neighbrs $j = 1, \ldots, p+1$ (weights/parameters)

(d) Repeat for $i = 1, ..., n$

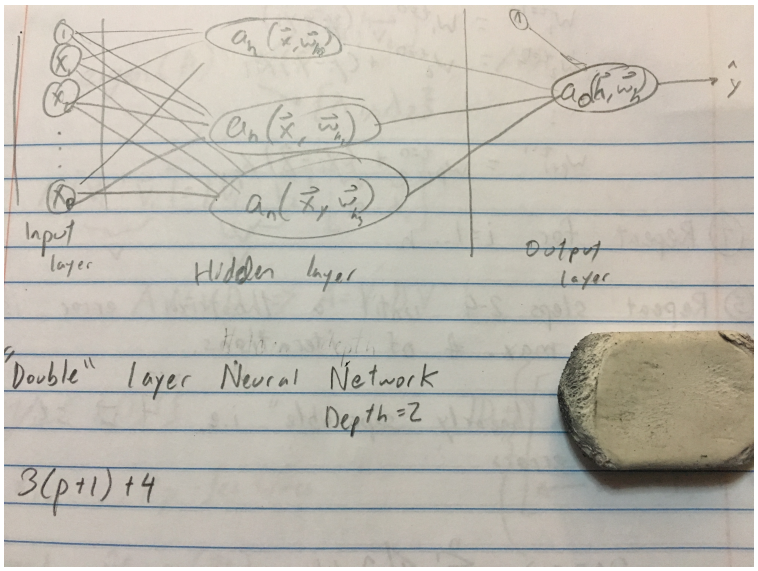
(e) Repeat steps 2-4 until a theoretical errors is reached or until a maximum amount of iterations is reached.

□

(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.


Input layer
Hidden layer
Output layer
"Double" layer Neural Network
Depth = 2
3(p+1)+4