MATH 390.4 / 650.2 Spring 2018 Homework #1t

Rebecca Strauss

Tuesday 22nd May, 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?

Although terms 'predict' and 'forecast' are used interchangeably today, in earlier centuries, the two were not synonymous. Predictions were associated with fortune telling and other whimsical practices. Forecasting involved using knowledge to plan ahead for the future.

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

John P. Ioannidis's found that most experiments conducting in a laboratory setting, would fail if recreated in the real world. This implies that there is something fundamentally wrong with the way that "certified" scientists run their experiments and/or interpret results.

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

Human being's most power defense is our mind, according to Silver. As humans, we have the ability to notice patterns in the world around us.

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

The amount of useful information is not increasing.

(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.

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

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

Science is anything that can be tested in the real world with predictions.

(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 made the wrong assumptions. They assumed that an individual defaulting on his mortgage is uncorrelated to others defaulting on their loan. The combination of the housing bubble burst and a bad economy created a common factor among mortgage holders and made them all more likely to default. Additionally, the error in the CDO models were nonlinear, making the error compound very quickly once a mistake was made.

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

Risk is something you can calculate and assign a number or probability to. Uncertainty is unquantifiable.

(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 an event that is not included in your historical data set.

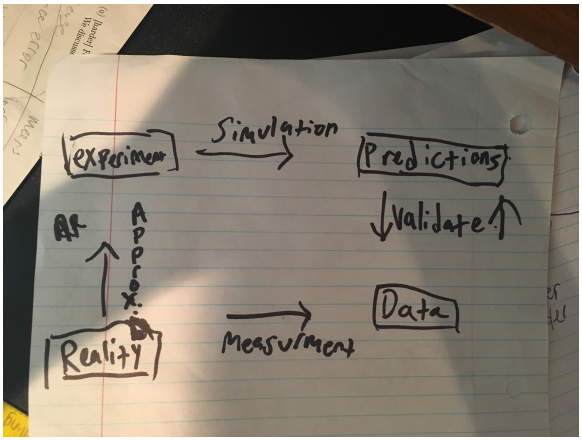
$$\mathcal{X} \not\subset \mathbb{D}$$

(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.

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..



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

Data is the measured response to a phenomenon.

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

Predictions are the response to the model we build g,

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

We can never know the "truth"

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

We can approximate the truth and derive useful results.

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

When our predictions are close to the data we measured as the response to the phenomenon, we have a good 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?

A bad model since it is extremely vague, like this answer.

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

Technically yes, since we could put it in mathematical terms.

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

If a person has eaten an apple in the last 24 hours(1) or not(0).

(d) [easy] What is(are) the output(s) in this model?
Either the doctor stays away(1) or he doesn't(0)

(e) [easy] Devise a means to measure the main input. Call this x_1 going forward.
Let x_1=apples eaten/24 hours

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

Need to define a time period for the apple effect, see if person visited doctor during that amount of time following the eating of an apple

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

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

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

$\vec{z}$ are the unobservable aspects producing the phenomenon, for example the strength of a persons immune system.

(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.

$\mathcal{X}$=apples eaten/24 hours

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

$$y = f(x_1)$$

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

Supervised learning is using historical records/responses to model a phenomenon.

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

Supervised learning is not an analytical solution because there are no formulas to solve. We are using the data to create an answer, hence empirical.

(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.

$$\mathbb{D} = \{< x_1, y_1 >, \cdots, < x_n, y_n >\}$$

Each $< x_i, y_i >$ represents how many apples a person ate in 24 hours and whether or not the same person visited the doctor within a specific amount of time following their apple consummation.

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

$\mathcal{H}$ is the set of all candidate functions that approximate the phenomenom we are trying model. $\mathcal{A}$ takes in $\mathbb{D}, \mathcal{H}$ and returns the "best" choice in $\mathcal{H}$.

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

The domain of g would be all $\mathbb{R}$ and the range would be $0, 1$

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

Yes, $g \in \mathcal{H}$. $\mathcal{A}$ returns a function from $\mathcal{H}$ and since $g = \mathcal{A}(\mathbb{D}, \mathcal{H})$, g must be an element in $\mathcal{H}$

(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 is not reasonable because we don't even know what f is.

(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 $t - h^*$.

(a) Parameter Estimation Error $h^*(\vec{x}) - g(\vec{x})$: we don't always pick the best candidate model from the set.

(b) Misspecification Error $f(\vec{x}) - h^*(\vec{x})$ $f \notin \mathcal{H}$, if f is a curve and we choose $\mathcal{H}$ to be lines, h^* is going to differ from f.

(c) Ignorance Error $t(\vec{x}) - f(\vec{x})$ we can never know the truth.

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

(a) Parameter estimation error can be decreased with a larger n

(b) Misspecification Error can be decreased by a better $\mathcal{A}$

(c) Ignorance Error can be decreased by adding features p.

(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?

$$\mathcal{H} = \{\mathbb{I}_{x \geq x_T}\}$$

The parameters of $\mathcal{H}$ are x_T.

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

$$g(\vec{x}) = \mathbb{I}_{x \geq 1}$$

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.

$$\mathcal{H} = \{\mathbb{I}_{b \cdot \vec{x} > 0} \, \vec{b} \in \mathbb{R}^{p+1}\}$$

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

(a) Initiate the weights to be $\vec{b} = 0$ or random

(b) Calculate $\hat{\boldsymbol{y}}_i = \mathbb{I}_{b_0 \cdot \vec{x_i} > 0}$

(c) update all weights $j = 0, \cdots, p$

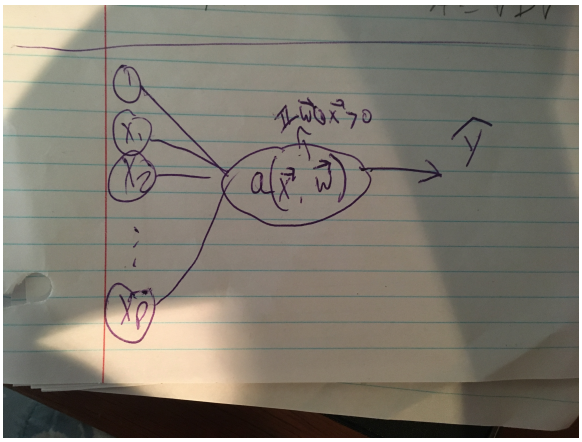
$$\begin{aligned} b_0^{t=1} &= b_0^{t=0} + (y_i - \hat{y}_i) && (1) \\ b_1^{t=1} &= b_1^{t=0} + (y_i - \hat{y}_i) && (x_{i,1}) \\ &\vdots \\ b_p^{t=1} &= b_p^{t=0} + (y_i - \hat{y}_i) && (x_i, p) \end{aligned}$$

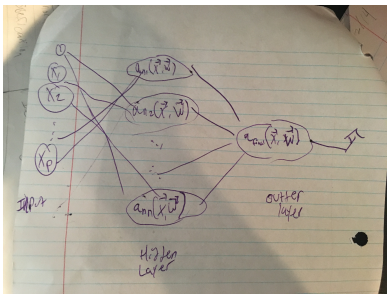
(d) Repeat steps (b) and (c) $\forall i \in \{1, \cdots, n\}$

(e) Repeat setps (b)-(d) until error hits threshold or a prespecfied number of iterations are run.

(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.


1
x_1
x_2
x_p
$a(\vec{x}, \vec{w})$
$\vec{w} \cdot \vec{x} > 0$
$\hat{y}$


x_1
x_2
x_p
$a_{out}(\vec{x}, \vec{w})$
outter layer
Hidden Layer