MATH 390.4 / 650.2 Spring 2018 Homework #1t

Angel Montero

Monday 26th 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?

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

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

(d) [easy] Information is increasing at a rapid pace, but what 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.

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

(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.

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

(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.

(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 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 the result of measuring a property of the objective reality. It is information that comes from the real world, not the constructs we create. It is the reality.

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

Predictions are the results of the models we create. In contrast to data, predictions are man-made and are supposed to approximate reality, but is not reality itself.

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

The idea is that there is only one objective truth, the world follows only one set of possible rules. This objective truth and all all its rules are so complex that we can never hope to describe it exactly. There can be an infinite set of models that closely approximate the truth, but only one actual truth. A simple way to put it would be that if a model was right, it would cease to be a model and would just be reality itself, which is why all models have to be wrong.

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

Some models can do a good job of closely representing reality. We can use the rules of this model to find explanations about real world phenoma or make predictions that fall close to real outcomes.

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

A model is considered good if the predictions it produces closely follow actualy reality. If it doesn't it's considered a poor 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?

If we define a model to be good if predictions are close to reality, this model would be a good fit if a person who eats an apple on a day does not see a doctor on that day. I think this would be a good model on this criteria alone, because a person on any given day is more likely to eat an apple than they are to see a doctor (unless they are really ill) So the days where an apple is eaten and there are no doctor visits would outnumber the days where an apple is eaten and there has been a doctor visit. The converse, if a person does not eat an apple every day, then they will see a doctor, is tricker because the adage does not make any specific claims about it. However let's assume it is true, and lets make the model more specific. If a person eats an apple or more a day, they will not see a doctor, otherwise, if a person eats less than one apple day, they will see a doctor on that day. This more specific definition has a poorer fit because we are predicting that one the days that the person does not eat an apple or more a day, they will see a doctor, which I don't think is an accurate presentation of reality.

Another issue with the model is that it claims that eating apples alone is the only causal input, or alone is the input that predicits wethere a person will see the other. There many other factors, such as previous illnesses, family history, traumatic injuries... etc.. that will determine if you see a doctor or not.

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

This is not a mathematical model. By definition, a mathematical model has numerical inputs and this statement does not. We can however attempt to transform the statement into a mathematical model in the following questions.

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

The apples that the person eats in a given day.

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

Wether the person saw a doctor that day.

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

Apples vary in their size and weight, in order to compare the data it might be best to measure the weight of apples (in grams) eaten per day, which will be more comparable from person to person regardless of how many individual apples they eat.

Let x be the mass weight (in grams) of apple eaton a day.

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

let the output, y be 1 if the person did *not* see a doctor that day. If the person did see a doctor that day, then $y = 0$

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

$y = g(x_1)$

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

$y = t(z_1, \ldots, z_t)$ represents the objective reality. It is the true description of the world.

$z_1, \ldots, z_t$ are the inputs that completely describe/ control y. The causal inputs, of which x_1 may or may not belong.

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

$p = dim([x_1]) = 1$

Our model consists of a single predictive feature, x_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.

If x_1 is the gram weight of apples eaten a day, then $x_1 \in \mathbb{R}$

and $\mathcal{X} = [x_1]$

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

There is a function, f, that given x_1, most closely approximates y.
$y \approx f(x_1)$.
$f(x_1)$ will never *equal* y but it will approximate it.

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

Supervised learning uses prior data (from the real world) in order to find a model that can best approximate or predict y.

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

Supervised learning uses empirical numerical data from the past to make a model that produces an empirical solution. An analytical solution is an exact relationship with variables, functions, which cannot be derived from numerical data.

(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 training data obtained from real world measurements.

$\mathbb{D} := < \vec{X_1}, \vec{Y_1} >, \ldots, < \vec{X_n}, \vec{Y_n} >$
where $\vec{X_i} = [x_1]$ and $\vec{Y_i} = [y]$

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

In supervised learning:
$\mathcal{H} =$ The set of all candidate functions
$\mathcal{A} =$ The algorithm that picks the best candidate function in $\mathcal{H}$

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

For $\mathbb{D}$:
$x_1 \in \mathbb{R}+; y \in 0, 1$

g has as its only argument x_1 and its output is y. So the domain of g is $\mathbb{R}+$ and its range $0, 1$

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

Yes. g is by definition the best possible approximation to f in $\mathcal{H}$, where f is the best possible function given available data.

(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 $\hat{y}^*$.

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

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

It is usually not reasonable. f is the function that best predicts y when you only have the feature x_1 available. A very many x_1 is needed to find the best f but out training data almost always is limited in size, so we would never be able to find f with the training data we have.

(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 the best function f that can explain y with x is NOT in the set of candidate functions, there are three sources of error.

Estimation Error: $h^*(x) - g(x)$

$h^*(x)$ is the best approximation of $f \in \mathcal{H}$. g is the function that is actually chosen. The estimation error measures the difference between the two.

Misspecification Error: $f(x) - h^*(x)$

Misspecification error measures how well you picked you model, or how well you picked the candidate functions. Are there any features that were left out or are not needed.

Error due to ignorance: $t(z) - f(x)$

How far off is our best possible estimate of reality $f(x)$ from actual reality $t(z)$

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

Estimation Error: $h^*(x) - g(x)$

Increase the number of samples.

Misspecification Error: $f(x) - h^*(x)$

Develop a better algorithm, or have a better $\mathcal{H}$ that has an h^* as close as possible to f

Error due to ignorance: $t(z) - f(x)$

Expand our attributes, $\vec{X}$

(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?

$\mathcal{H} = \{\mathbb{1}_x > x_t : x_t \in \mathbb{R}\}$

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

$$\mathbb{1}_x = \begin{cases} 1 \text{ if x >= 150} \\ 0 \text{ otherwise} \end{cases}$$

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.

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

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

1. Initialize $\vec{w}$ to 0
2. Calculate $\hat{y}_i = 1$ if $\vec{w} \cdot \vec{x} > 0$. Otherwise it is 0.
3. Update the weights $\vec{w}_j = \vec{w}_j + (y_i - \hat{y}) * \vec{x}j$
4. Repeat steps 2 and 3 for all $i \in \{1, ..., n\}$ 5. Repeat steps 2 through 4 until a treshold error is reached or until 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.