MATH 390.4 / 650.2 Spring 2020 Homework #4

Tyron Samaroo

Monday 18th May, 2020

Problem 1

These are questions about the rest of Silver's book, chapters 7–11. You can skim chapter 10 as it is not so relevant for the class. For all parts in this question, answer using notation from class (i.e. $t, f, g, h^*, \delta, \epsilon, e, t, z_1, \ldots, z_t, \mathbb{D}, \mathcal{H}, \mathcal{A}, \mathcal{X}, \mathcal{Y}, X, y, n, p, x_{\cdot 1}, \ldots, x_{\cdot p}$, $x_{1\cdot}, \ldots, x_{n\cdot}$, etc.) as well as in-class concepts (e.g. simulation, validation, overfitting, etc.)

Note: I will not ask questions in this assignment about Bayesian calculations and modeling (a large chunk of Chapters 8 and 10) as this is the subject of Math 341. It is obviously important in Data Science (that's why Math 341 is a required course in the data science and statistics major).

(a) [easy] Why are flu fatalities hard to predict? Which type of error is most dominant in the models?

The reason why flu fatalities are hard to predict is because there is not enough $\mathbb{D}$ on the flu. Hence, we will have a difficult time to accurately predict the flu so the type of error that is most dominant would be estimation error.

(b) [easy] In what context does Silver define extrapolation and what term did he use? Why does his terminology conflict with our terminology?

Sliver define extrapolation as basic method of prediction that is too basic. Extrapolation assumes that the current trend will continue indefinitely, into the future. I would say his terminology is closely related to ours since we defined extrapolation to be predicting on data that is outside of our data $\mathbb{D}$ range. Just because we see a trend doesn't mean we can predict correctly when its outside our range. The only difference is that he mention just future, future data could be in our range so thats the only difference.

(c) [easy] Give a couple examples of extraordinary prediction failures (by vey famous people who were considered heavy-hitting experts of their time) that were due to reckless extrapolations.

Example 1: A writer in Times of London predicted that by the 1940's that every street in London would be buried under nine feet of horse manure.

Example 2: English economist Sir William Petty assumed that things remain the same way and predicted that the global population will be just over 700 million in 2020.

Example 3: Stanford biologist Paul R. Ehrlich and his wife Anne predicted that hundreds of million of people would die of starvation in the 1970;s

(d) [easy] Using the notation from class, define "self-fulfilling prophecy" and "self-canceling prediction".

A self-fulfilling prophecy is when our prediction correctly predicts on future data.
A self-canceling prediction is when our prediction incorrectly predicts on future data.

(e) [easy] Is the SIR model of infectious disease under or overfit? Why?

The SIR model underfits.SIR model assumes that everyone behaves the same way. It assumes that everyone is equally likely to be susceptible to a disease,vaccinated. It doesn't take into consideration things like race,gender,age,religion,sexual orientation.

(f) [easy] What did the famous mathematician Norbert Weiner mean by "the best model of a cat is a cat"?

Mathematicion Norbert Weiner means that to truly understand any phenomenon is know the entirety of the phenomenon itself.In our terms we will never know the true casual inputs $z_1, \ldots, z_t$ for the true function t.Models don't know the entirety of a phenomenon so it could only approximate the $z_1, \ldots, z_t$

(g) [easy] Not in the book but about Norbert Weiner. From Wikipedia:

> Norbert Wiener is credited as being one of the first to theorize that all intelligent behavior was the result of feedback mechanisms, that could possibly be simulated by machines and was an important early step towards the development of modern artificial intelligence.

What do we mean by "feedback mechanisms" in the context of this class?

The feedback mechanisms would be the $z_1, \ldots, z_t$ the causal inputs.

(h) [easy] I'm not going to both asking about the bet that gave Bob Voulgaris his start. But what gives Voulgaris an edge (p239)? Frame it in terms of the concepts in this class.

Whats gives Voulgaris an edge is that he combines his knowledge of statistics with his knowledge of basketball in order to identify meaningful relationships in the data. In our terms Voulgaris is able to find $x_1, \ldots, x_t$ that best approximates the $z_1, \ldots, z_t$

(i) [easy] Why do you think a lot of science is not reproducible?

A lot of science is not reproducible because its difficult to find the $x_1, \ldots, x_t$ that best approximates the $z_1, \ldots, z_t$. I also think its very difficult to narrow down the hypothesis set in order to reduce miss-specification error.

(j) [easy] Why do you think Fisher did not believe that smoking causes lung cancer?

Fisher believed that smoking and lung cancer were correlated and not a causation relationship. The reason why is that there was not enough information that supported the casual relationship between lung cancer and smoking.It was the 1950's so there wasn't enough information yet to determine the casual relationship.

(k) [easy] Is the world moving more in the direction of Fisher's Frequentism or Bayesianism?

The world is moving more towards Bayesianism. A lot of people are taking into consideration the odds of something happening.

(l) [easy] How did Kasparov defeat Deep Blue? Can you put this into the context of over and underfiting?

Kasparov was able to defeat Deep because he made a move that Deep Blue statistically assumed that moved he made was lower than another moved he could of made. Deep blue was overfit its only based its prediction with what it already knows to be statistically advantageous.Kasparov made a move that Deep Blue did not have enough examples of so it was not able to accurately predict the correct moved to make.

(m) [easy] Why was Fischer able to make such bold and daring moves?

Fischer was able to make such a move because he is able to think unlike Deep Blue. Deep Blue is only making statistical moves with the highest percentage. Fischer is able to understand what to do in scenario he never been in whereas Deep Blue had no idea how to correctly handle the situation.

(n) [easy] What metric y is Google predicting when it returns search results to you? Why did they choose this metric?

Google prediction is returning a search result usefulness. They chose this term because just because what you may search for might related to many different things.For example if searching for a Mexican restaurant doesn't mean you are looking for one in Mexico. It figures out you might be looking for one near you. If not it might be a good idea to make a better search query.

(o) [easy] What do we call Google's "theories" in this class? And what do we call "testing" of those theories?

We call Google theories models and we call the testing of those theories validation.

(p) [easy] p315 give some very practical advice for an aspiring data scientist. There are a lot of push-button tools that exist that automatically fit models. What is your edge from taking this class that you have over people who are well-versed in those tools?

Our edge from taking this class is that we are able to make better predictions.We think more analytical that society who do things out of bad habits or blind adherence to tradition.

(q) [easy] Create your own 2×2 luck-skill matrix (Fig. 10-10) with your own examples (not the ones used in the book).

	Low Luck	High Luck
Low Skill	Inheriting money from parents	Winning the lottery
High Skill	Having a Career or Job	Billionaire by studying and investing in stock market

(r) [easy] [EC] Why do you think Billing's algorithms (and other algorithms like his) are not very good at no-limit hold em? I can think of a couple reasons why this would be.

Algorithms aren't able to capture human behavior properly in these scenarios. In no limit hold em even though there is a low probability that a person has a better hand that you it still possible.

(s) [easy] Do you agree with Silver's description of what makes people successful (pp326-327)? Explain.

I agree with Silver description of what makes people successful. We seen many examples of hard work, talent, person's opportunities and environment. These are useful signals that can be consider as x's that approximate the z's for the phenomenon success.

(t) [easy] Silver brings up an interesting idea on p328. Should we remove humans from the predictive enterprise completely after a good model has been built? Explain

If we remove humans from predictive enterprise after a good model has been built we are ignoring a signal. We need as much features that relate to the phenomenon to accurately build a model or else we will underfit.

(u) [easy] According to Fama, using the notation from this class, how would explain a mutual fund that performs spectacularly in a single year but fails to perform that well in subsequent years?

Fama found that there was no consistence in how a fund did even over a five year increments. Fama is having a problem of trying to interpolation. Hes trying to predict on something that isnt in his range. He also is looking for correlation which may not imply causation. I believe Fama is not getting enough x's to approximate the z's for his predictions.

(v) [easy] Did the Manic Momentum model validate? Explain.

The Manic Momentum model did not validate. It failed during the early 2000 since the pattern did not hold true during this time.

(w) [easy] Are stock market bubbles noticable while we're in them? Explain.

Stock market bubbles difficult to notice while we are in them.Because its not very predictable short term, as it tells you almost nothing about the market. The market prediction lies in how traders behave so its difficult to get this information during short term period.

(x) [easy] What is the implication of Shiller's model for a long-term investor in stocks?

The implication of Shiller model is that stocks price pattern are predictive long term and are almost meaningless short term.

(y) [easy] In lecture one, we spoke about "heuristics" which are simple models with high error but extremely easy to learn and live by. What is the heuristic Silver quotes on p358 and why does it work so well?

The heuristic Silver quotes "follow the crowd, especially when you don't know any better" works so well because there's evidence that it works.More often whats working for majority often works more than when it does not.

(z) [easy] Even if your model at predicting bubbles turned out to be good, what would prevent you from executing on it?

Even if my model is good at predicting bubbles I would be cautious to execute it because there still a possibility that it can be wrong and I could lose equity that I might invest.

(aa) [easy] How can heuristics get us into trouble?

Heuristics can get us in to trouble because they aren't always true in ever single circumstance. Heuristics are too general so its best to use as a last resort.

Problem 2

These are some questions related to validation.

(a) [easy] Assume you are doing one train-test split where you build the model on the training set and validate on the test set. What does the constant K control? And what is its tradeoff?

The constant K control the proportion of how we split the train and test. The trade off with doing this is reducing the possibility of over fitting on our data.

(b) [harder] Assume you are doing one train-test split where you build the model on the training set and validate on the test set. If n was very large so that there would be trivial misspecification error even when using $K = 2$, would there be any benefit at all to increasing K if your objective was to estimate generalization error? Explain.

If n is very large and we are able to reduce miss specification there might not be any benefit in increasing K if our objective was to reduce estimation error. The reason why is that having such a large n means there is a low chance that we will over-fit.

(c) [easy] What problem does K-fold CV try to solve?

K-fold CV try to solve the problem of finding the best train-test split. It is trying to find the best split that reduce estimation error.

(d) [E.C.] Theoretically, how does K-fold CV solve it?

K-fold CV try to solve it by preforming K folds on the data it then holds one of the K folds it does as a validation and repeats this until it validates every fold with the rest of its fold. For example if its split the data into 10 parts it will take 1/10 of the data to validate the other 9/10. Then it would use a different 1/10 to validate another 9/10

Problem 3

These are some questions related to polynomial-derived features and logarithm-derived features in use in OLS regression.

(a) [harder] What was the overarching problem we were trying to solve when we started to introduce polynomial terms into $\mathcal{H}$? What was the mathematical theory that justified this solution? Did this turn out to be a good solution? Why / why not?

The problem we were trying to solve is what to do when the data isn't linear. Since our $\mathcal{H}$ only consist of linear models it did not do well with data that looked like a sine curve. So our currrent $\mathcal{H}$ set will not produce a good model. It was the Weierstrass Approximation Theorem that states that every continuous function defined on a closed interval [a, b] can be uniformly approximated as closely as desired by a polynomial function. So based on this we can approximate any phenomenon as a polynomial function. It able to capture all the curves and even work on linear models. It might be a good solution but not the best solution as if we make our $\mathcal{H}$ to contain complex polynomial function it can over fit.

(b) [harder] We fit the following model: $\hat{\boldsymbol{y}} = b_0 + b_1 x + b_2 x^2$. What is the interpretation of b_1? What is the interpretation of b_2? Although we didn't yet discuss the "true" interpretation of OLS coefficients, do your best with this.

b_1 is a constant that weights on x that scales it linearly. b_2 is another constant that scale x^2.$The b's$

(c) [difficult] Assuming the model from the previous question, if $x \in \mathcal{X} = [10.0, 10.1]$, do you expect to "trust" the estimates b_1 and b_2? Why or why not?

Given $x \in \mathcal{X} = [10.0, 10.1]$ it will be very difficult to extrapolate outside that range. Unless that range is justify for whatever the phenomenon is it will be difficult for the b's to properly weight the x's.

(d) [difficult] We fit the following model: $\hat{\boldsymbol{y}} = b_0 + b_1 x_1 + b_2 \ln(x_2)$. We spoke about in class that b_1 represents loosely the predicted change in response for a proportional movement in x_2. So e.g. if x_2 increases by 10%, the response is predicted to increase by $0.1b_2$. Prove this approximation from first principles.

In class we discussed that $\ln(x+1) = x - \frac{x^2}{2} + \frac{x^3}{3} ... \frac{x^n}{n} \approx x \; if \; x \; is \; small$

$$Note \; x_f = x_2 + 0.1x_2 \quad x_o = x_2$$

$$\begin{aligned} b_2 \Delta \ln(x) &= b_2(\ln(x_f) - \ln(x_o)) \\ &= b_2(\ln(x_2 + 0.1x_2) - \ln(x_2)) \\ &= b_2 ln\left(\frac{x_2 + 0.1x_2}{x_2}\right) \\ &= b_2 ln(1.1) \\ if \; \frac{x_f}{x_0} &\approx 1 \; then \; \approx b2(\frac{x_f}{x_0} - 1) \\ &\approx b_2(1.1 - 1) = 0.1b_2 \end{aligned}$$

(e) [easy] When does the approximation from the previous question work? When do you expect the approximation from the previous question not to work?

It only works with a lot of assumptions such that if x is small. It will not work where there is a large difference between x_2 so approximation will fail.

(f) [harder] We fit the following model: $\ln(\hat{\boldsymbol{y}}) = b_0 + b_1 x_1 + b_2 \ln(x_2)$. What is the interpretation of b_1? What is the interpretation of b_2? Although we didn't yet discuss the "true" interpretation of OLS coefficients, do your best with this.

b_1 is the proportional change in our prediction $\hat{y}$. b_2 is the $\ln(\hat{y})$ change for each unit of x_2

(g) [easy] Show that the model from the previous question is equal to $\hat{\boldsymbol{y}} = m_0 m_1^{x_1} x_2^{b_2}$ and interpret m_1.

$$\begin{aligned} \ln(\hat{\boldsymbol{y}}) &= b_0 + b_1 x_1 .... + b_p \ln(x_p) \\ &= e^{b_0 + b_1 x_1 .... + b_p \ln(x_p)} \\ &= \underbrace{e^{b_0}}_{m_0} \underbrace{e^{b_1 x_1}}_{m_1} \underbrace{e^{b_p \ln(x_p)}}_{m_p} \end{aligned}$$

$$p \; is \; 2 \; in \; this \; case \; so$$

$$\hat{y} = e^{b_0} e^{b_1 x_1} e^{b_2 \ln(x_2)}$$

$$\hat{y} = m_o m_1^{x_1} x_2^{b_2}$$

$$m_1 \; is \; the \; changed \; of \; e^b 1 \; for \; each \; x_1 \; for \; \bar{y}$$

Problem 4

These are some questions related to the model selection procedure discussed in lecture.

(a) [easy] Define the fundamental problem of "model selection".

There are many different models to choose from so its difficult to choose the model that is the best possible one. There are too many algorithms $\mathcal{A}$ and many hypothesis sets $\mathcal{H}$

(b) [easy] Describe the first procedure we introduced to solve it.

We try to solve this by trying to fit finite number of models m and do honest validation on each $g_1....g_m$. We fit an $\mathcal{A}$ and $\mathcal{H}$ on D_{train} and compute S_e on D_{test}. We then select g_m* which has the lowest S_e.

(c) [easy] Discuss possible problems with this procedure.

One problem we might have is that S_e will have high variance. So we need pick a reasonable K. Another problem is that we aren't doing honest validation since we use D_test multiple times.

(d) [easy] Describe how you would use this model selection procedure to find hyperparameter values in algorithms that require hyperparameters.

I would follow the same procedure picking a finite number of models with finite number of hyper-parameters. I will instead hold out a D_test for the end for the final g_m* which has the lowest S_e. So for each $g_1....g_m$ I would fit an $\mathcal{A}$ and $\mathcal{H}$ on D_{train} and compute S_e on D_{select} another partition of the $\mathcal{D}$. D_{select} is another partition of data that we will keep using to compute S_e. This solve issue of honest validation.

(e) [easy] Does using both inner and outer folds in a double cross-validation procedure solve some of these problems?

Inner fold solve some of these problems since we finding best way partition our D_{train} and D_{select} to reduce S_e. Outer fold does not solve the problem of honest validation since we are folding on out D_{test} many times making our validation dishonest.