MATH 390.4 / 650.2 Spring 2020 Homework #4

Tziporah Horowitz

Monday 20th April, 2020

1 Silver's Book Chapters 7-11

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?

Flu fatalities are hard to predict because there is not enough data flu pandemics. Each strain of the flu has its own fatality ratio, so when extrapolated, flu fatality cannot be predicted accurately. The majority of the error comes from the *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?

Silver defines extrapolation as "the assumption that the current trend will continue indefinitely, into the future" in the context of exponential growth. In our terms, extrapolation is predicting on data that is outside the range of the training set $\mathbb{D}$. While both definitions assume that the trend can change in the future, our definition does not necessarily imply that the model will fail under extrapolation.

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

- In 1894, a writer for the Times of London predicted that by the 1940s, every street in London would be buried under nine feet of the manure.
- In 1682, English economist, Sir William Petty predicted that the human population growth rate would be slow.

- In 1968, Paul and Anne Ehrlich predicted that hundreds of millions of people would die from starvation in the 1970's.

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

A *self-fulfilling prophecy* occurs when a prediction ($\hat{y}$) causes an event (y_{future}) to take place while a *self-canceling prediction* tends to undermine itself.

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

SIR models underfit because they assume that everyone in a population behaves the same way. They do not account for differences in susceptibility, vaccinations, demographics, or social interactions.

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

Wiener meant that because models are merely approximations of $t(z_1, \ldots, z_t)$, there will always be some error in explaining/predicting y. The only way to get an exact estimate is to see the phenomenon itself.

(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?

Feedback mechanisms are the true causal inputs, $z_1, \ldots, z_t$.

(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.

Voulgaris' advantage is that he combines his knowledge of statistics with his knowledge of basketball to identify meaningful relationships in the data, i.e. he uses *a priori* information.

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

Much of science is not reproducible because the hypothesis sets are too large, so laboratory conditions do not accurately mimic real phenomena and in turn may produce random predictions.

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

Fisher believed that smoking is correlated with lung cancer, but that the relationship is not causal because there was no prior information suggesting it at the time.

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

The world is moving more toward Bayesianism.

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

Kasparov defeated Deep Blue by playing moves that were statistically unlikely. The computer's algorithm was overfit, being too precise in predicting the most likely move for Kasparov. So, when Kasparov moved to a position that only occurred one time in a master-level chess competition, the computer lost.

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

Fischer was able to make such moved because he thought out of the scope of traditional chess heuristics.

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

Google is trying to measure the "usefulness" of web-pages so that the searcher can find the optimal results with a simple query.

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

Google's theories are models and they are tested with 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?

We are starting without any bad habits.

(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	inherit money	win the lottery
High Skill	get a job	play the 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.

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

Silver defines success as a combination of hard work, natural talent, and a person's opportunities and environment. I agree with his definition because success is often attributed to opportunities but is difficult to sustain without hard work or natural talent.

(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 you remove humans from the predictive enterprise, the model would only be able to predict well under perfect circumstances, i.e. it would overfit.

(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's claim could not predict well over 5 years because he did not take into account other factors like the type of product a company sells or if the company made a profit or a loss. His model therefore had large estimation error.

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

The Manic Momentum model was not validated and would therefore have the adverse effect if it was used in the early 2000's.

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

Stock market bubbles are hard to notice while we're in them because there they're not predictable in the short-term. Nevertheless, there are economic indicators pointing toward them.

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

While stocks are unpredictable in the short-term, dramatic changes can often be predicted in the long-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?

Silver quotes the heuristic, "Follow the crowd, especially when you don't know any better." This works well because in many situations, following the crowd leads to the most probable outcome.

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

Real-world constraints on trading and capital can prevent executing on a good prediction.

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

Heuristics can get us into trouble when they don't align with circumstance.

2 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?

K controls the proportion of test observations used for validation. It is the tradeoff between bias and variance.

(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.

Following the *Central Limit Theorem*, the variance will be approximately the same with any K when n is large. So, there would be no benefit in increasing K.

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

K-fold CV tries to find the best train-test split for validation.

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

3 Polynomial-Derived and Logarithm-Derived Features

(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?

If the data is non-linear, $\mathcal{H} =$ *linear models* will not be able to fit a good model. The *Weierstrauss Approximation Theorem* states that for any continuous function f, there exists a polynomial function p that is close to f. Using the *Weierstrauss Approximation Theorem*, you can use a polynomial function that approximates a linear function to fit a model using *OLS*. By *Vandermonde's Matrix Theorem*, $\boldsymbol{X}$ will still be full-rank.

(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 the constant change in response to a unit of x while b_2 will change by x in response to a unit of x.

(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?

No, because the model will fail under extrapolation.

(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_2 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.

$$\begin{aligned}\ln(x+1) &= x - \frac{x^2}{2} + \frac{x^3}{3} - \ldots + \ldots \\ &\approx x \qquad \text{if } x \text{ is small}\end{aligned}$$

Let $\Delta \ln(x_2) = \ln(x_2 + 0.1x_2) - \ln(x_2)$,

$$\begin{aligned}b_2 \Delta \ln(x_2) &= b_2\big(\ln(x_2 + 0.1x_2) - \ln(x_2)\big) \\ &= b_2 \ln\left(\frac{x_2 + 0.1x_2}{x_2}\right) \\ &= b_2(1.1) \\ &\approx (1.1 - 1)b_2 = 0.1b_2\end{aligned}$$

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

This approximation works when $\frac{x_2'}{x_2} \approx 1$. When there is a large difference between x_2' and x_2, the approximation will likely 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 predicted change in $\ln(\hat{\boldsymbol{y}})$ for each unit of x_2 when x_2 is held constant. b_2 is the predicted change in response for a proportional movement in x_2 when x_1 is held constant.

(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.

$$\begin{aligned}\ln(\hat{\boldsymbol{y}}) &= b_0 + b_1 x_1 + b_2 \ln(x_2) \\ &= e^{b_0 + b_1 x_1 + b_2 \ln(x_2)} \\ &= e^{b_0} e^{b_1 x_1} e^{b_2 \ln(x_2)} \\ &= m_0 m_1^{x_1} x_2^{b_2}\end{aligned}$$

m_1 is the predicted change of e^{b_1} in $\hat{\boldsymbol{y}}$ for each unit of x_1 when x_2 is held constant.

4 Model Selection

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

There are many models to choose from, so how do we know that we're choosing the right one?

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

Fit m models and perform honest validation on $g_1, \ldots, g_m$. Then select g_{m_*} with the lowest S_e.

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

- How do you know you're picking the right model and how well does g_{m_*} perform?
- The validation is not honest since $\mathbb{D}_{test}$ is used many times.

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

You can create a model g and fit it m times using m possible values for hyperparameters. Then perform honest validation and select the model g_{m_*} with the lowest S_e.

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

Yes.