MATH 390.4 / 650.2 Spring 2020 Homework #4

Pizon Shetu

Due Monday, April 20, 2020 11:59PM by email

(this document last updated 2:35am on Tuesday 21st April, 2020)

Instructions and Philosophy

The path to success in this class is to do many problems. Unlike other courses, exclusively doing reading(s) will not help. Coming to lecture is akin to watching workout videos; thinking about and solving problems on your own is the actual "working out." Feel free to "work out" with others; **I want you to work on this in groups.**

Reading is still *required*. For this homework set, read Chapters 7-11 of Silver's book. You should be googling and reading about all the concepts introduced in class online. This is your responsibility to supplement in-class with *your own* readings.

The problems below are color coded: green problems are considered *easy* and marked "[easy]"; yellow problems are considered *intermediate* and marked "[harder]", red problems are considered *difficult* and marked "[difficult]" and purple problems are extra credit. The *easy* problems are intended to be "giveaways" if you went to class. Do as much as you can of the others; I expect you to at least attempt the *difficult* problems.

This homework is worth 100 points but the point distribution will not be determined until after the due date. See syllabus for the policy on late homework.

Up to 7 points are given as a bonus if the homework is typed using LaTeX. Links to instaling LaTeX and program for compiling LaTeX is found on the syllabus. You are encouraged to use `overleaf.com`. If you are handing in homework this way, read the comments in the code; there are two lines to comment out and you should replace my name with yours and write your section. The easiest way to use overleaf is to copy the raw text from hwxx.tex and preamble.tex into two new overleaf tex files with the same name. If you are asked to make drawings, you can take a picture of your handwritten drawing and insert them as figures or leave space using the "\vspace" command and draw them in after printing or attach them stapled.

The document is available with spaces for you to write your answers. If not using LaTeX, print this document *including this first page* and write in your answers. I do not accept homeworks which are *not* on this printout.

NAME: _____

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?
Since extrapolation is difficult with small data thus making predicting for flu fatalities hard

(b) [easy] In what context does Silver define extrapolation and what term did he use? Why does his terminology conflict with our terminology?
Our terminology of extrapolation is an prediction outside the range of our data whereas Silver's is an assumption made which will continue indefinitely

(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.
Sir William Petty"s prediction of the growth of global population and Paul R. prediction of the amount of people dying from starvation

(d) [easy] Using the notation from class, define "self-fulfilling prophecy" and "self-canceling prediction".
The prediction Y differ from both where self-fulfilling prophecy is when a prediction occurs due to inherent bias which leads to expected results and self-canceling is when an inherent bias which leads to cancellation of expected results which otherwise should occur

(e) [easy] Is the SIR model of infectious disease under or overfit? Why?
It is underfit due to the assumption everyone will behave the same in a given population

(f) [easy] What did the famous mathematician Norbert Weiner mean by "the best model of a cat is a cat"? He meant if you're going to model after something your details should come from that something as in you can only get all the details from a cat if you're modeling for a cat rather than getting it from a dog or other animals.

(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 mechanism is a performance tester of a model to see how well it does with new data or unseen data

(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.
He took into an account for multiple features which gave him far greater insights into his bets thus giving him an edge

(i) [easy] Why do you think a lot of science is not reproducible?
Because it is impossible to replicate every little detail and nuisance of science, it'd be almost like knowing every little detail of a particle atom and etc to be able to get a perfect reconstruction of science

(j) [easy] Why do you think Fisher did not believe that smoking causes lung cancer?
Because he believed that it was a correlation rather than causation when it came to smoking and lung cancer

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

(l) [easy] How did Kasparov defeat Deep Blue? Can you put this into the context of over and underfiting?
Deep blue did not have all possible data available to him where Kasparov was able to find and maneuver to a victory due to his great experience at the game, this is a results of underfiting for Deep blue

(m) [easy] Why was Fischer able to make such bold and daring moves?
Because he had a unrelenting confidence in his skills which he believed could not fail him, and his imagination allowed him to think of many possibilities

(n) [easy] What metric y is Google predicting when it returns search results to you? Why did they choose this metric?
Google's y which is predicting which results will you find the most useful, they use this metric because Google can constantly test and experiment with others who have searched the same results and see which ones they clicked thus they constantly update the rankings on which webpages are the most revelevant.

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

Understanding what truly influences (features) an prediction thus devising a better model allowing for more accurate predictions

(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: coin flip | lottery high skill: Chess | 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.

It lacks many strategies which it cannot take account for.

(s) [easy] Do you agree with Silver's description of what makes people successful (pp326-327)? Explain. I agree with Silver, as certain features and qualities of a person will always allow them to succeed in any time of an era, as well as luck being at the right place and at the right time has made many their fortunes.

(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
We should not remove humans, as when uncertainties occur and new unseen data has entered the playing field for which models cannot account for we need humans to analyze.

(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?
This can be compared to a training vs test set split where a model does well to data it was trained on but fails on test data which it is not accustomed to.

(v) [easy] Did the Manic Momentum model validate? Explain.
It did not validate as it only used past data to making historical predictions thus when new data deviated from historical data the model collapsed.

(w) [easy] Are stock market bubbles noticable while we're in them? Explain.
No because if that was the case we would have predicted and stop many of the past historical bubbles that caused the market to drop and collapse, like the dotcom bubble or the housing bubble and etc. Most of the time once a bubble is noticed it is too late.

(x) [easy] What is the implication of Shiller's model for a long-term investor in stocks?
That a stock price will follow a trend which in the long run will be predictable due to how traders and the stock market behave

(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?
Because following others who are not failing at it is better than guessing and doing something random when you don't know anything about it.

(z) [easy] Even if your model at predicting bubbles turned out to be good, what would prevent you from executing on it?
It is very difficult to change the trend of masses even if a bubble is being predicted, to change the opinions and decisions of millions who affect the market which is highly volatile is a difficult if not impossible task.

(aa) [easy] How can heuristics get us into trouble?
For times which it cannot account for by that I mean when the Heuristic works for the general population but fails for special cases such as eating an apple a day keeps the doctor away, but what if a person has allergies to an apple and etc.

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?
K controls the size of the training and test set, which in turn effects the trade of bias vs 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.
There is no inherent benefit as increasing K would lead to a higher variance of the generalization error

(c) [easy] What problem does K-fold CV try to solve?
It tries to solve high variance due to train-test split

(d) [E.C.] Theoretically, how does K-fold CV solve it?
It does so by computing the out of sample residuals for each fold and runs and runs the metric on all n, thus it rotates the train-test split which allows each observation to be part of the test set once, this done K times and also validating K times.

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 with higher dimension and high polynomial terms in H is that it becomes hard to interpret and eventually it over-fits. The problem we are trying to solve is to minimize misspecification error and the mathematical theory that justifies this solution is Weierstrass-Approximation.

(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 slope of x where as b_2 is the slope of x_2.

(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?
Since the values are not from a overfit model I can expect to trust the b_1 and b_2 estimates

(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 $0.1 b_2$. Prove this approximation from first principles.
$yhat = b_0 + b_1 x_1 + b_2 ln(x_2)$
$ln(x_2 + 1) = x_2 - x_2^2/2 + x_2^3/3 - x_2^4/4 + ...$
$= b_1((x_{2f} - x_{2n}ot)/x_{2n}ot)$

(e) [easy] When does the approximation from the previous question work? When do you expect the approximation from the previous question not to work?
The approximation works when x is near 1 where as it won't work for any other values

(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 represents the percentage change for $\ln(\hat{\boldsymbol{y}})$ in x_1
b_2 represent the proportional change in $\ln(\hat{\boldsymbol{y}})$ in 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.
$\ln(\hat{\boldsymbol{y}}) = b_0 + b_1 x_1 + b_2 \ln(x_2) = \hat{\boldsymbol{y}} = e^{b_0 + b_1 x_1 + b_2 \ln(x_2)}$
$\hat{\boldsymbol{y}} = e^{b_0} e^{b_1^{x_1}} x_2^{b_2} = \hat{\boldsymbol{y}} = m_0 m_1^{x_1} x_2^{b_2}$

Problem 4

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

(a) [easy] Define the fundamental problem of "model selection".
The fundamental problem is selecting which model to choose within the infinite number of model available

(b) [easy] Describe the first procedure we introduced to solve it.
g_1 is the fit with A_1, H_1 on $D_t rainandSe_1 iscomputedon D_t est.$
g_2 is the fit with A_2, H_2 on $D_t rainandSe_2 iscomputedon D_t est.$
g_m is the fit with A_m, H_m on $D_t rainandSe_m iscomputedon D_t est. Select g_m*$ which has the lowest Se

(c) [easy] Discuss possible problems with this procedure.
Having a high variance and the integrity of honest validation

(d) [easy] Describe how you would use this model selection procedure to find hyperparameter values in algorithms that require hyperparameters.
Have many λ to find the best model where each λ will give us different models

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