# MATH 342W / 650.4 Spring 2021 Homework #4

Professor Adam Kapelner

NOT DUE

(this document last updated 10:11pm on Wednesday 28th April, 2021)

### 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*. 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 and write in your answers. I do not accept homeworks which are *not* on this printout. Keep this first page printed for your records.

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?

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

(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.

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

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

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

(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?

(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.

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

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

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

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

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

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

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

(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?

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

(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.

(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

(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?

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

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

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

(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?

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

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

## 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?

(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.

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

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

## Problem 3

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

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

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

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

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

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

## Problem 4

These are some questions related to the CART algorithms.

(a) [easy] Write down the step-by-step $\mathcal{A}$ for regression trees.

(b) [difficult] Describe $\mathcal{H}$ for regression trees. This is very difficult but doable. If you can't get it in mathematical form, describe it as best as you can in English.

(c) [harder] Think of another "leaf assignment" rule besides the average of the responses in the node that makes sense.

(d) [harder] Assume the $y$ values are unique in $\mathbb{D}$. Imagine if $N_0 = 1$ so that each leaf gets one observation and its $\hat{\boldsymbol{y}} = y_i$ (where $i$ denotes the number of the observation that lands in the leaf) and thus it's very overfit and needs to be "regularized". Write up an algorithm that finds the optimal tree by pruning one node at a time iteratively. "Prune" means to identify an inner node whose daughter nodes are both leaves and deleting both daughter nodes and converting the inner node into a leaf whose $\hat{\boldsymbol{y}}$ becomes the average of the responses in the observations that were in the deleted daughter nodes. This is an example of a "backwards stepwise procedure" i.e. the iterations transition from more complex to less complex models.

(e) [difficult] Provide an example of an $f(\boldsymbol{x})$ relationship with medium noise $\delta$ where vanilla OLS would beat regression trees in oos predictive accuracy. Hint: this is a trick question.

(f) [easy] Write down the step-by-step $\mathcal{A}$ for classification trees. Feel free to reference steps in (a).

(g) [difficult] Think of another objective function that makes sense besides the Gini that can be used to compare the "quality" of splits within inner nodes of a classification tree.

## Problem 5

These are some questions related to probability estimation modeling and asymmetric cost modeling.

(a) [easy] Why is logistic regression an example of a "generalized linear model" (glm)?

(b) [easy] What is $\mathcal{H}_{pr}$ for the probability estimation algorithm that employs the linear model in the covariates with logistic link function?

(c) [easy] If logistic regression predicts 3.1415 for a new $\boldsymbol{x}_*$, what is the probability estimate that $y = 1$ for this $\boldsymbol{x}_*$?

(d) [harder] What is $\mathcal{H}_{pr}$ for the probability estimation algorithm that employs the linear model in the covariates with cloglog link function?

(e) [difficult] Generalize linear probability estimation to the case where $\mathcal{Y} = \{C_1, C_2, C_3\}$. Use the logistic link function like in logistic regression. Write down the objective function that you would numerically maximize. This objective function is one that is argmax'd over the parameters (you define what these parameters are — that is part of the question).

Once you get the answer you can see how this easily goes to $K > 3$ response categories. The algorithm for general $K$ is known as "multinomial logistic regression", "polytomous LR", "multiclass LR", "softmax regression", "multinomial logit" (mlogit), the "maximum entropy" (MaxEnt) classifier, and the "conditional maximum entropy model". You can inflate your resume with lots of jazz by doing this one question!

(f) [easy] Graph a canonical ROC and label the axes. In your drawing estimate AUC. Explain very clearly what is measured by the $x$ axis and the $y$ axis.

(g) [easy] Pick one point on your ROC curve from the previous question. Explain a situation why you would employ this model.

(h) [easy] Graph a canonical DET curve and label the axes. Explain very clearly what is measured by the $x$ axis and the $y$ axis.

(i) [easy] Pick one point on your DET curve from the previous question. Explain a situation why you would employ this model.

(j) [difficult] The line of random guessing on the ROC curve is the diagonal line with slope one extending from the origin. What is the corresponding line of random guessing in the DET curve? This is not easy...

## Problem 6

These are some questions related to bias-variance decomposition. Assume the two assumptions from the notes about the random variable model that produces the $\delta$ values, the error due to ignorance.

(a) [easy] Write down (do not derive) the decomposition of MSE for a given $\boldsymbol{x}_*$ where $\mathbb{D}$ is assumed fixed but the response associated with $\boldsymbol{x}_*$ is assumed random.

(b) [easy] Write down (do not derive) the decomposition of MSE for a given $\boldsymbol{x}_*$ where the responses in $\mathbb{D}$ is random but the $\boldsymbol{X}$ matrix is assumed fixed and the response associated with $\boldsymbol{x}_*$ is assumed random like previously.

(c) [easy] Write down (do not derive) the decomposition of MSE for general predictions of a phenomenon where all quantities are considered random.

(d) [difficult] Why is it in (a) there is only a "bias" but no "variance" term? Why did the additional source of randomness in (b) spawn the variance term, a new source of error?

(e) [harder] A high bias / low variance algorithm is underfit or overfit?

(f) [harder] A low bias / high variance algorithm is underfit or overfit?

(g) [harder] Explain why bagging reduces MSE for "free" regardless of the algorithm employed.

(h) [harder] Explain why RF reduces MSE atop bagging $M$ trees and specifically mention the target that it attacks in the MSE decomposition formula and why it's able to reduce that target.

(i) [difficult] When can RF lose to bagging $M$ trees? Hint: setting this critical hyperparameter too low will do the trick.

## Problem 7

These are some questions related to correlation-causation and interpretation of OLS coefficients.

(a) [easy] Consider a fitted OLS model for y with features $x_1$, $x_2$, ..., $x_p$. Provide the most correct interpretation of the quantity $b_1$ you can.

(b) [easy] If $x$ and $y$ are correlated but their relationship isn't causal, draw a diagram below that includes $z$.

(c) [easy] To show that $x$ is causal for $y$, what specifically has to be demonstrated? Answer with a couple of sentences.

(d) [harder] If we fit a model for y using $x_1$, $x_2$, ..., $x_7$, provide an example real-world illustration of the causal diagram for $y$ including the $z_1$, $z_2$, $z_3$.

## Problem 8

These are some questions related to missingness.

(a) [easy] What are the three missing data mechanisms? Provide an example when each occurs (i.e., a real world situation).

(b) [easy] Why is listwise-deletion a terrible idea to employ in your $\mathbb{D}$ when doing supervised learning?

(c) [easy] Why is it good practice to augment $\mathbb{D}$ to include missingness dummies? In other words, why would this increase oos predictive accuracy?

(d) [easy] To impute missing values in $\mathbb{D}$, what is a good default strategy and why?

## Problem 9

These are some questions related to lasso, ridge and the elastic net.

(a) [easy] Write down the objective function to be minimized for ridge. Use $\lambda$ as the hyperparameter.

(b) [easy] Write down the objective function to be minimized for lasso. Use $\lambda$ as the hyperparameter.

(c) [easy] We spoke in class about when ridge and lasso are employed. Based on this discussion, why should we restrict $\lambda > 0$?

(d) [harder] Why is lasso sometimes used a preprocessing step to remove variables that likely are not important in predicting the response?

(e) [easy] Assume $\boldsymbol{X}$ is orthonormal. One can derive $\boldsymbol{b}_{\text{lasso}}$ in closed form. Copy the answer from the wikipedia page. Compare $\boldsymbol{b}_{\text{lasso}}$ to $\boldsymbol{b}_{\text{OLS}}$.

(f) [harder] Write down the objective function to be minimized for the elastic net. Use $\alpha$ and $\lambda$ as the hyperparameters.

(g) [easy] We spoke in class about the concept of the elastic net. Based on this discussion, why should we restrict $\alpha \in (0, 1)$?