MATH 342W / 650.4 Spring 2022 Homework #4

Professor Adam Kapelner

Due 11:59PM April 14 by email

(this document last updated 1:03pm on Tuesday 5th April, 2022)

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 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?

(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.

(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?

(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.

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

(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 *approximate* interpretation of b_2? Although we didn't yet discuss the "true" interpretation of OLS coefficients, do your best with this.

(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.

Problem 2

These are some questions related to extrapolation.

(a) [easy] Define extrapolation and describe why it is a net-negative during prediction.

(b) [easy] Do models extrapolate differently? Explain.

(c) [easy] Why do polynomial regression models suffer terribly from extrapolation?

Problem 3

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 4

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

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

(b) [easy] Using two splits of the data, how would you select a model?

(c) [easy] Discuss the main limitation with using two splits to select a model.

(d) [easy] Using three splits of the data, how would you perform model selection?

(e) [easy] How does using both inner and outer folds in a double cross-validation nested resampling procedure improve the model selection procedure?

(f) [easy] Describe how g_{final} is constructed when using nested resampling on three splits of the data.

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

(h) [difficult] Given raw features $x_1, \ldots, x_{p_{raw}}$, produce the most expansive set of transformed p features you can think of so that $p \gg n$.

(i) [easy] Describe a methodology that can create a linear model on a subset of the transformed featuers (from the previous problem) that will not overfit.

Problem 5

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 δ 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. This should be short because you can reference the steps you wrote for the regression trees 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.