MATH 390.4 / 650.2 Spring 2020 Homework #5

Professor Adam Kapelner

Due Monday, May 18, 2020 11:59PM by email

(this document last updated 8:50am on Thursday 7th May, 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, you should finish Silver's book but I am not asking any questions on ch12, 13 and the conclusion. They are very interesting though! 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 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. 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 2

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 3

These are some questions related to bias-variance decomposition. Assume the two assumptions from the notes about the random variable model that produces the δ 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 4

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 λ as the hyperparameter.

(b) [easy] Write down the objective function to be minimized for lasso. Use λ 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) [difficult] Assume $\boldsymbol{X}$ is orthonormal. Derive $\boldsymbol{b}_{\text{lasso}}$ in closed form. Try to do this yourself. The answer is on wikipedia if you get stuck.

(f) [harder] Write down the objective function to be minimized for the elastic net. Use α and λ as the hyperparameters.

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

Problem 5

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 6

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.