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 6:33pm 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: Yoon Shin

Problem 1

N_0 is a hyper-parameter

These are some questions related to the CART algorithms.

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

Step 1 allow all the data be in a dataset

Step 2 Get all possible othogonal-to-axis split $x_1, \ldots x_p$ $j = 1 \ldots p$ & $i \in 1, \ldots n-1$ then calculate SSE_l & SSE_R the SSEs in putative left node & right node then

$$SSE_{weighted} = \frac{n_l SSE_l + n_r SSE_r}{n_l + n_r}$$

n_l = # left data
n_r = # right data

find the smallest i.e. then create a inner node that split rule and a left leaf and with $\hat{y} = \bar{y}_l$ and a right with $\hat{y} = \bar{y}_r$

3 If $n_l > N_0$ then set dataset = left partition and run step 2
if $n_r > N_0$ then set dataset = right partition and run step 2

(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.

$\mathcal{H} = \{ \vec{w}_1 \mathbb{1}_{x \in [\ ]} + \vec{w}_2 \mathbb{1}_{x \in [\ ]} + \ldots + \vec{w}_B \mathbb{1}_{x \in [\ ]}, \ldots$

where $\vec{w} \in \mathbb{R}^B$ where B is the # of Bins. Each indicator function will belong to a given interval that depends on the size and for each bin you get the avg.

Every Bin has different avg's.

The Bin size might give underfit if its too big, while if the bin is smaller the model can be better.

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

use mode to assign leaf values

(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.

Missclassification error function

$$\text{argmin} \left| \sum_{i=1}^{n} \mathbb{1}_{l_i + C_k} + \sum_{i=1}^{n} \mathbb{1}_{r_i \neq C_k} \right|$$

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)?

Because it uses logistic link function to predict

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

$$\mathcal{H}_{pr} = \left\{ \frac{1}{1+e^{-\vec{w}\cdot\vec{x}}} \,\middle|\, \vec{w} \in \mathbb{R}^{p+1} \right\}$$

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

The ~~Probability~~ estimate would be close to 1

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

: Complementary log-log : $\phi(v) = 1 - e^{-e^v}$

$\rightarrow \mathcal{H}_{\mathcal{P}} = \{1 - e^{-e^{(\vec{w} \cdot \vec{x})}} \mid \vec{w} \in \mathbb{R}^{p+1}\}$

(e) [difficult] Generaline 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).

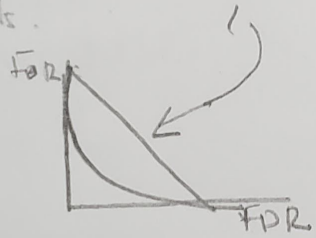
Once you get the answer you can see how this easily goes to $K > 3$ response categories. The algorithm for general K is known as "multinomial logistic regression", "polytomous LR", "multiclass LR", "softmax regression", "multinomial logit" (mlogit), the "maximum entropy" (MaxEnt) classifier, and the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

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

FOR is not to high thus 0's & 1's are more accurate

(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...

DET curve is the opposite of ROC so the line should be sloping downwards.



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.

$MSE(\vec{x}_*) = E_{\mathbb{D}}[E^2|\vec{x}_*] = E[(y_* - g(\vec{x}_*))^2|\vec{x}_*]$

$= \sigma^2 + (f(\vec{x}_*) - g(\vec{x}_*))^2 \geq \sigma^2$

(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.

$MSE(\vec{x}_*) = E_{\Delta_1 \ldots \Delta_n, \Delta_*}[(y_* - g(\vec{x}_*))^2|\vec{x}_*] =$

$= \sigma^2 + Bias[g(\vec{x}_*)]^2 + Var[g(\vec{x}_*)]$

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

$MSE = E_x[MSE(\vec{x}_*)] = \sigma^2 + E_x[Bias(g(\vec{x}_*))^2] + E_x[var(g(\vec{x}_*))]$

(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?

Since g is fixed there is a bias in (a) for (b) there different sets of TD and g is different for each TD thus more variance in MSE.

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

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

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

It takes the avg of all models and variance is lowered when we have large # of models

(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.

It shrinks the variance terms as trees become more decorrelated, this is done by allowing randomness to splits through the means of picking the right subset of features that are split randomly. Take the avg of all the models via bagging

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

When P_{try} is too small due to the gain of p term being low, it won't outweigh the bias term, eventually this leads to an ~~under~~ fit model

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.

$$\underset{\vec{w} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \{SSE + \lambda \|\vec{w}\|^2\}$$

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

$$\operatorname{argmin} \{SSE + \lambda \|w\|_1\}$$

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

λ is restricted to > 0 because when $\lambda = 0$ we're with Ridge, when $\lambda = 1$ it is lasso and λ having something between the 2 it is elastic net.

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

To remove variables which are not impactful as lasso picks variables which are impactful it sets the junk variables to 0

(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 $\boldsymbol{b}_{\text{OLS}}$.

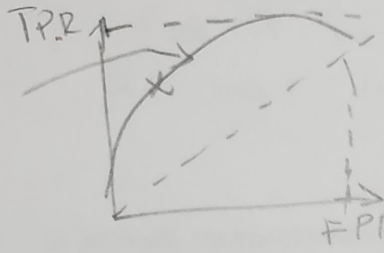
$$\hat{\beta}_j = S_{N\lambda}(\hat{\beta}_j^{OLS}) = \hat{\beta}_j^{OLS} \max\left(0, 1 - \frac{N\lambda}{|\hat{\beta}_j^{OLS}|}\right)$$

Comparing it to the function which minimizes in ridge we get $\hat{\beta}_j = (1+N\lambda)^{-1}\hat{\beta}_j^{OLS}$ so this shrinks all coefficients by a uniform factor of $(1+N\lambda)$ and doesnot set any coefficients to zero. Also comparing it to regression with best subset selection, in which the goal is to minimize $\min_{\beta \in \mathbb{R}^p}\{SSE + \lambda\|\beta\|_0\}$ where $\|\cdot\|_0$ is the "ℓ^0" norm, which is defined as $\|z\|_0 = m$ if exactly m components of z are non-zero. In this case it can be shown that $\hat{\beta}_j = H_{\sqrt{N\lambda}}(\hat{\beta}_j^{OLS}) = \hat{\beta}_j^{OLS} I(|\hat{\beta}_j^{OLS}| \geq \sqrt{N\lambda})$ where H_α is the so-called hard thresholding function and I is an indicator function

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

$$\hat{b}_{en} = \text{argmin}\left\{SSE + \lambda\left(\alpha\|w\|_1 + (1-\alpha)\|\vec{w}\|_2^2\right)\right\}$$

(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.



ROC

AUC is about .6

x-axis is the false positive rates

y-axis is the true positive rate

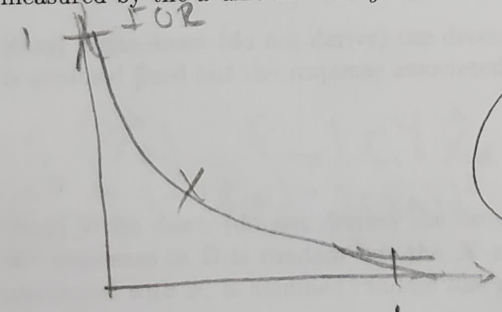
$$\frac{n(1-p+n)(1-\bar{y})}{n(1-\bar{y})} = 1-p+n$$

$$\frac{n(1-p+n)\bar{y}}{n\bar{y}} = 1-p+n$$

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

TPR is higher than FPR thus employ the 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.



x-axis is false discovery rate

y-axis is the false omission rate

FDR

$\frac{FP}{PP} = 1 - \text{precision}$

$\frac{FN}{PN}$

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

α is restricted to be (0,1) because $\alpha = 0$ when it is ridge & $\alpha = 1$ when it is lasso

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).

(1) MCAR: seeding emails and some are lost in junk and spam.

(2) MAR: People who don't vote because they don't care

(3) NMAR: People who don't talk about their problems due to being embrassed of it

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

Because it would increase variance and MSE increase

(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?

We build a prediction for X_j's and fill in the missing data with the predictions

(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 $\hat{y} = y_i$ (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 $\hat{y}$ 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.

keep steps 1 & 2 from (a)
then add another step Step 3 : if $n_l > N_0 = 1$
then data set = left partition and do step 2
if $n_r > N_0 = 1$ then do the same except
dataset = right partition
Step 4: if you get 2 leaf nodes, then take
the avg of 2 and replacing the
node

(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. There are none

(f) [easy] Write down the step-by-step $\mathcal{A}$ for classification trees. Feel free to reference steps in (a). Same as step 1 & 2 as (a)
except in step 2 you calculate $Gini_l$ & $Gini_r$
where $Gini_l = \sum_{k=1}^{K} \hat{p}_k(1-\hat{p}_k)$ & $\hat{p}_k = \frac{1}{n_l}\sum_{i=1}^{n} \mathbb{1}_{y_i=k}$
and same for $Gini_r$ except change
l to r and select the rule where
$Gini_{weighted} = \frac{n_l gini_l + n_r Gini_r}{n_l + n_r}$ and
follow the rest of step 2
and same step 3, 3

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

MissForest because it imputes missing values in $\mathbb{D}$. It fills in missing values with $\bar{x}_j$ and fit $\vec{x}_i \sim RF(\vec{x}_{-i})$ where $\vec{x}_i$ was present in the original $\mathbb{D}$, then set missing values of $\vec{x}_i$ to be predictions from the RF. Do this for all values up to p. This will be done until "convergence" meaning that the imputed values do not change from iteration to iteration thus giving a set $\mathbb{D}$ with no missing values

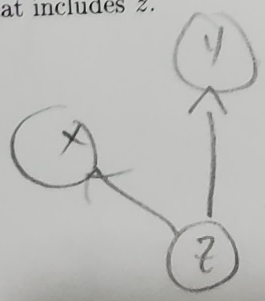
Problem 6

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

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

when comparing two "mutually observed" observations (A) & (B) sampled in the same function as observations in $\mathbb{D}$, when (A) has an x_1 measurement one unit larger than (B)'s x_1 but $x_2, x_3, \ldots, x_p$ then (A) is predicted to have a response y that differs by b_1 units on average from the response y of (B) assuming the linear model is true

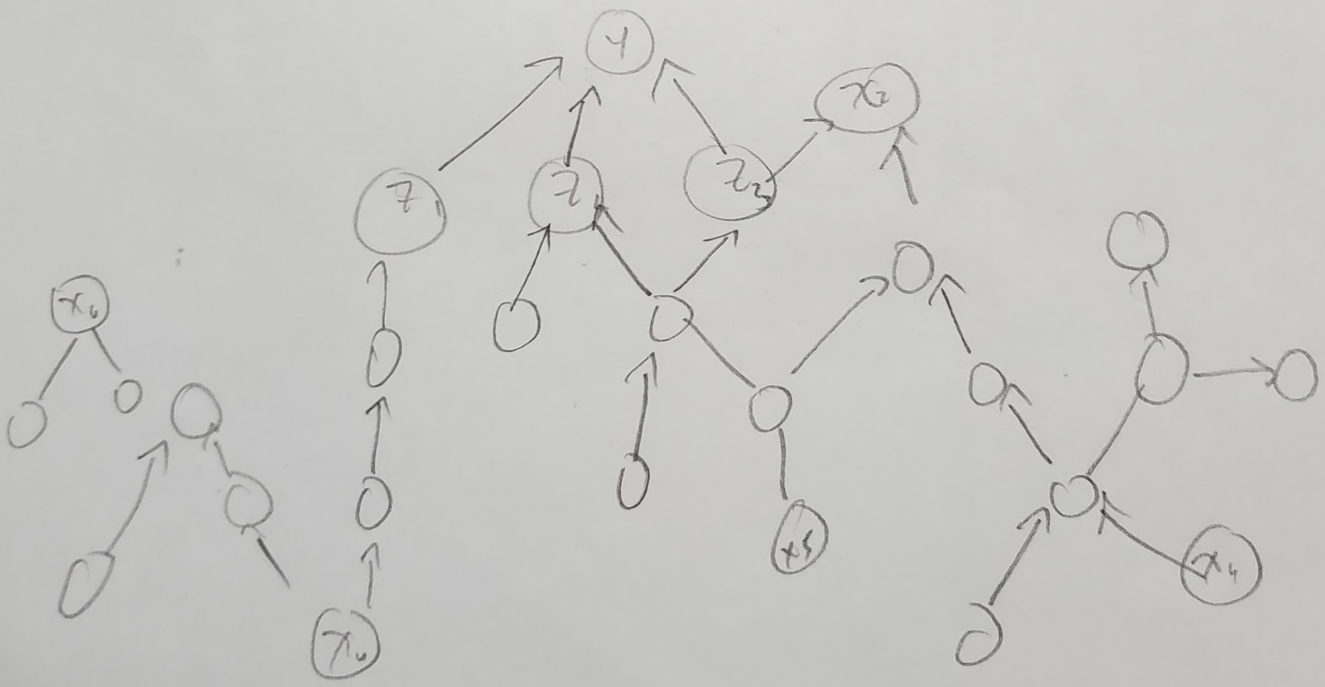
(b) [easy] If x and y are correlated but their relationship isn't causal, draw a diagram below that includes z.



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

x is causal for y if we have to show that x has a connection to y whether it is indirect or direct Also that if x is manipulated in any form then y changes

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



$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots b_n x_n$$