MATH 390.4 / 650.2 Spring 2020 Homework #5

Tyron Samaroo

Thursday 28th May, 2020

Problem 1

These are some questions related to the CART algorithms.

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

Step 0: Let the dataset be all the data

Step 1: Consider every possible orthognoal-to-axis split $x_j \leq \boldsymbol{x}_i$ where j = 1..p and i $\in 1...n-1$.Then compute SSE_l (SSE in the left node) and SSE_r(SSE in the right node).Select the rule where $\text{SSE}_{weighed} = \frac{n_L SSE_L + n_R SSE_R}{n_L + n_R}$ is the smallest. An inner node is created with that split rue and a left leaf node with $\hat{y} = \bar{y_L}$ and right left node with $\hat{y} = \bar{y_R}$

Step 2: Then if $n_L > N_o$ where N_o is a hyperparameter, you set the dataset on the left of the partition and run Step 1 on it. You then do the same thing for when $n_R > N_o$

(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.

The $\mathcal{H}$ for regression trees would be all of the different split rules that is created.

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

Instead of taking the average response in every node you can potentially take the median instead.

(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.

You would follow a similar approach to the tree algorithm but instead you will determine the error of the current tree, this will be the initial based metric. Then you will calculate the out of sample error for each pruning that is done to the tree and the lowest error will be the tree that is kept.

(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.

Since $f(\boldsymbol{x})$ will be the best possible model with some error due to ignorance. Vanilla OLS would not beat regression tree but may return the same result since regression tree would have a richer hypothesis set $\mathcal{H}$

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

It is almost the same for regression tree algorithm except that Step 1: Rather than computing SSE you will use a different objective function called Gini.
Select the rule where $Gini_{weighed} = \frac{n_L Gini_L + n_R Gini_R}{n_L + n_R}$ is the smallest.
Where $Gini_L := \sum_{K=1}^{K} \hat{p}(1-\hat{p})$

For $Gini_R$ *same applies just change* n_L *to* n_R
Where $\hat{p_k} := \sum_{i=1}^{n_L} \frac{I_{y_i = C_k}}{n_L}$
Then for each leaf assignment $\hat{y} = MODE[y's]$

(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.

Another objective function that makes sense besides Gini would be calculating the information that is gain at each spit.

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)?

Logistic regression is an example of a generalize linear model because it still consist of the same $\mathcal{H}$ but used a different link function.

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

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

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

It will be about 1. Logistic regression used the sigmoid function.

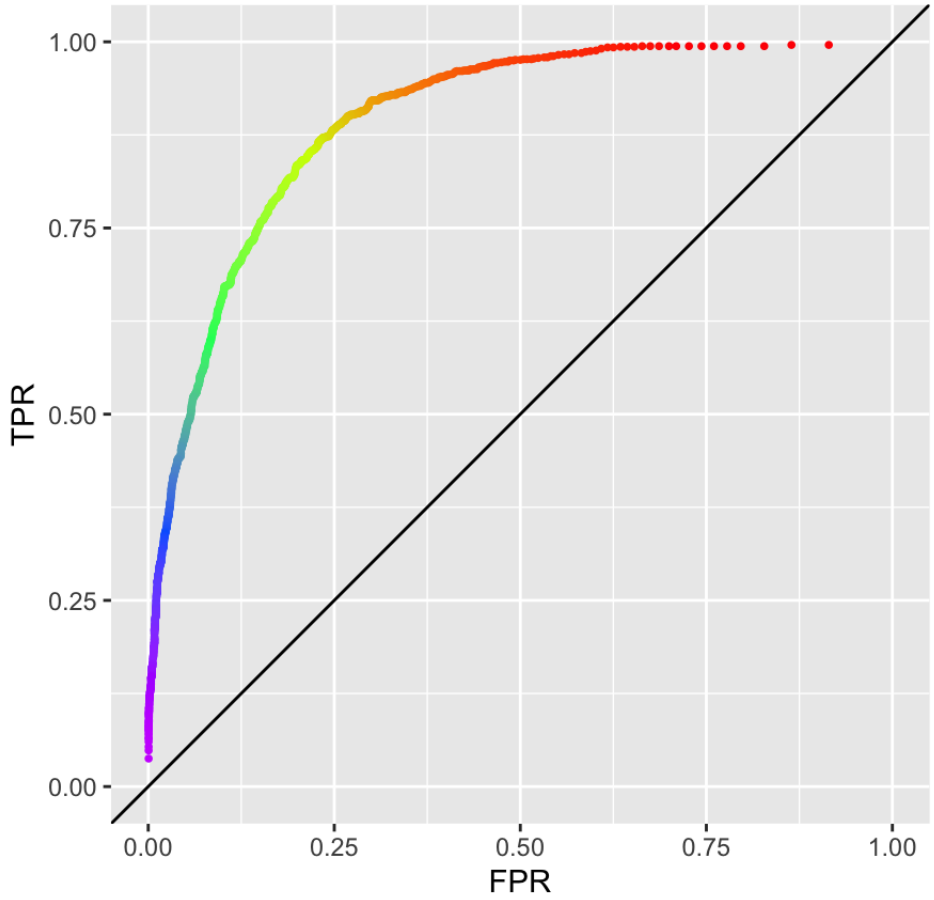
(d) [harder] What is $\mathcal{H}_{pr}$ for the probability estimation algorithm that employs the linear model in the covariates with cloglog link function?

$$\mathcal{H}_{pr} = \left\{ 1 - e^{-e^{\vec{w}\vec{x}}} : \vec{w} \in \mathbb{R}^{p+1} \right\}$$

(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.

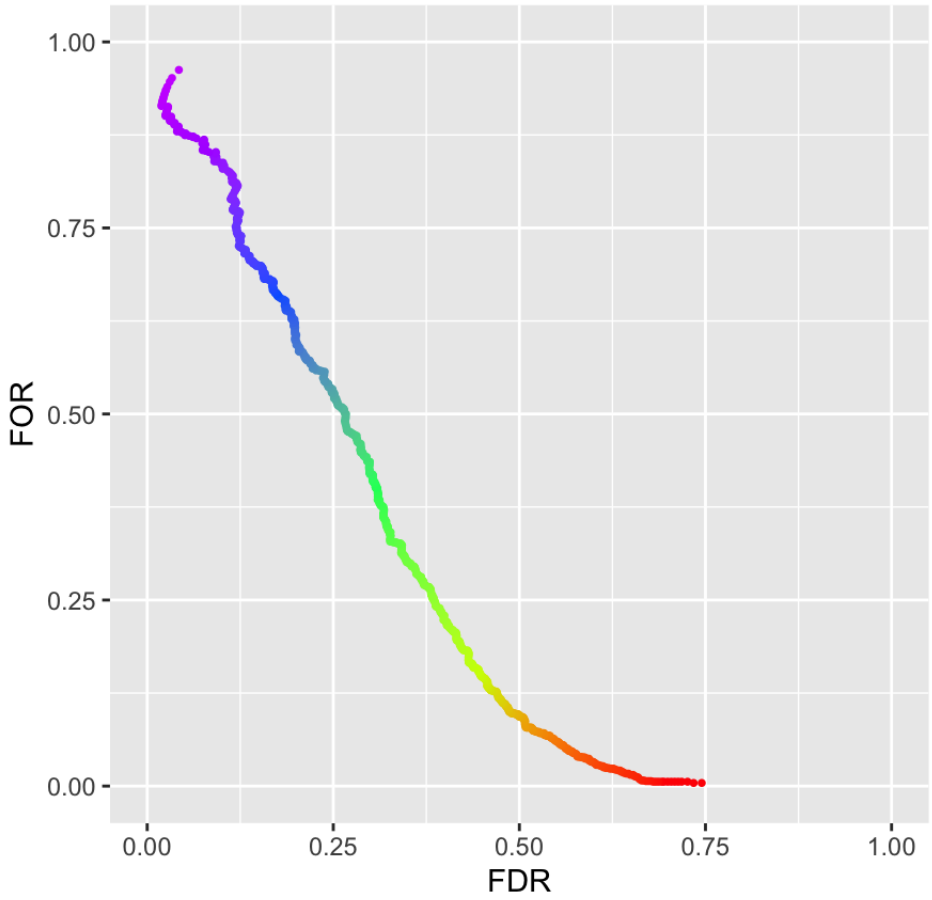


FPR calcuates the false positive rates. So it determine how many of the data was label as a false positive. TPR is the true postie rate that meanure how much of the data was label as truly postive.

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

If a point that is picked at TPR being close to one. This mean that we will make we will have more true positive but will make more error when it comes to FPR. We are trading one or the other depending on the situation or problem we are so solving

(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.

Picking FDR at .75 and FOR at 0.0 means that we care more abou the false discovery rate than the False Omission rate.

(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.

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

(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.

$$MSE(\boldsymbol{x}_*) = \sigma^2 + \text{Bias}[G]^2 + \mathbb{V}\text{ar}\left[G\right]$$

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

$$MSE = \sigma^2 + \mathbb{E}\left[\text{Bias}[G]^2\right] + \mathbb{E}\left[\mathbb{V}\text{ar}\left[G\right]\right]$$

(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?

This is underfit

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

This is overfit

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

Bagging reduces MSE because there is a richer H the hypothesis set, which will reduce the Bias to almost 0

(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.

RF builds from bagging but now reduced ρ even more. From bagging we have $MSE \approx \sigma^2 + \rho\mathbb{E}\left[\mathbb{V}\text{ar}\left[g_{(m)}\right]\right]$. The ρ is reduced which is the average correlation between the trees that was built with different bootstrap samples. It does this by splitting on a subset of features where $p_{try} < p$.

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

If N_o is set to be too low

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.

$$b = \underset{w \in \mathbb{R}^{p+1}}{\arg\min} \ \{SSE + \lambda ||\vec{w}||^2\}$$

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

$$b = \underset{w \in \mathbb{R}^{p+1}}{\arg\min} \ \{SSE + \lambda ||\vec{w}||\}$$

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

If $\lambda = 0$, Then $\boldsymbol{X}^T\boldsymbol{X}$ is not invertable.

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

Cant find it on wiki

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

$$b = \underset{w \in \mathbb{R}^{p+1}}{\arg\min} \ \{SSE + \lambda\big(\alpha ||w||_1 + (1-\alpha)||w||_2^2\big)\}$$

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

Restricting $\alpha \in (0, 1)$ will factor in both L1 and L2 Regularization which are ridge and 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).

Examples taken from Missingness Wiki

MCAR(Missing completely at random). Example would be random data that is missing by luck.

MAR(Missing at random). An example is that males are less likely to fill in a depression survey but this has nothing to do with their level of depression, after accounting for maleness.

NMAR(NOT MAR). Example would be if men failed to fill in a depression survey because of their level of depression.

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

Listwise Deletion is a terrible idea because a lot of your data may contain 1 column of missing data and if you do that you will lose a lot of data that would of helped in building a good predictive model. You can end up deleting 50 to 99 percent of your data.

(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?

It is good practice to augment $\mathbb{D}$ to include missingness dummies because it can provide insight why that data is missing and provide more predictive power.

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

A good strategy to impute values in $\mathbb{D}$ would be missForest because it is will always converge. It will work on the $\mathbb{D}$ until there is no missingness

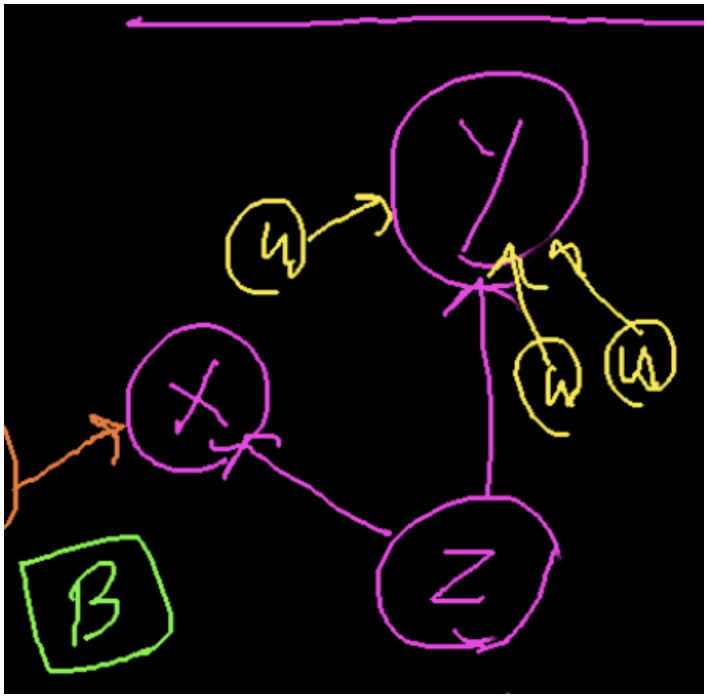
Problem 6

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.

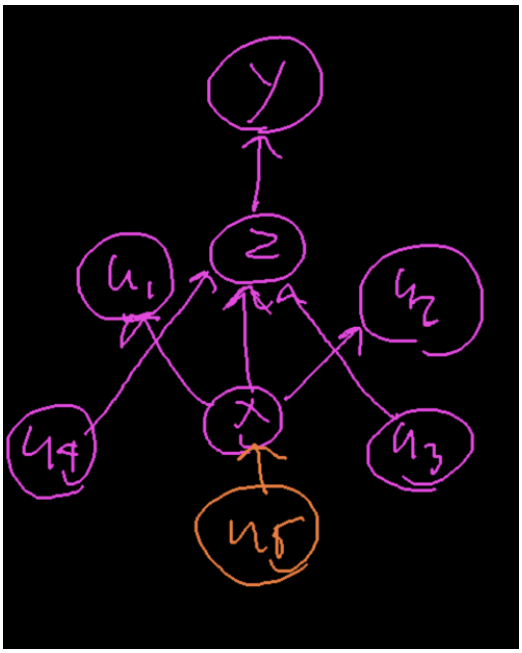
When comparing mutually observed features A and B sampled in the same fashion as features in the dataset, when A has an observation x value one unit larger than B's x measurement but both A and B share same measurements $x_1, x_2, ..., x_p$ then A is predicted to have a response y that differs by some unit b on average from 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.

(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.


y
z
x
u1
u2
u3
u4
u5