MATH 342W / 642 / RM 742 Spring 2024 HW #4

Professor Adam Kapelner

Due 11:59PM April 14

(this document last updated 10:45pm on Wednesday 17th April, 2024)

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) and also we now have $f_{pr}, h^*_{pr}, g_{pr}, p_{th}$, etc from probabilistic classification as well as different types of validation schemes).

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 341/343.

(a) [easy] Why are flu fatalities hard to predict? Which type of error is most dominant in the models?

Predicting flu fatalities is challenging due to the unpredictable nature of the influenza virus, which can mutate rapidly, producing new strains each season with varying levels of contagion and severity. This unpredictability complicates the accuracy of prediction models. The type of error most dominant in flu prediction models is the Type I error, or false positives. This occurs when predictions forecast severe impacts, such as high fatality rates or significant outbreaks, that do not materialize. An example provided in the book is the 1976 swine flu incident, where fears of a pandemic akin to the deadly 1918 Spanish flu led to widespread vaccination and public panic, despite the severe pandemic not occurring. This example illustrates the prevalence of Type I errors in flu prediction models, highlighting the challenges in balancing the risks of underestimation and overestimation.

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

In Nate Silver's book, the term "extrapolation" is defined in the context of prediction errors that arise from overly simplistic assumptions that current trends will continue indefinitely into the future. Silver uses the term "extrapolation" to discuss its inherent dangers, especially in fields like population growth and disease where exponential growth can lead to significantly erroneous forecasts. For example, he mentions historical instances such as city planners worrying about horse manure piling up in the streets or incorrect population growth predictions based on past trends. The terminology conflict arises because in the context of the book, "extrapolation" is used to caution against over-simplistic models, while in academic settings, extrapolation might be employed more carefully with considerations of its limitations and potential margins of error. This distinction highlights a different emphasis on the term's practical implications in real-world scenarios versus its theoretical application in academic studies. Silver is particularly critical of the use of extrapolation without adequately accounting for the potential changes in trends or new variables that might affect future outcomes .

(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.

Two notable examples of prediction failures due to reckless extrapolations include the "Horse Manure Crisis" and Paul Ehrlich's "The Population Bomb." In the late 19th century, forecasters predicted that London's streets would be buried under nine feet of horse manure by the 1940s due to the rising use of horse-drawn carriages. This prediction did not come to pass, as the advent of the automobile drastically reduced the number of horses in the city. Similarly, Paul Ehrlich's 1968 book predicted mass starvation in the 1970s due to overpopulation, extrapolating from the high fertility rates of the 1960s. However, this prediction failed to account for advances in agriculture and changes in social structures, leading to a significant overestimation of the crisis. These examples highlight the dangers of extrapolating current trends without considering potential changes in technology, society, and environment.

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

In the context of the class, a self-fulfilling prophecy relates to over-fitting a model when doing K-Fold CV and opening the $\mathcal{D}$ test, when we open the "blackbox" for validation, we might just be bias of a model that will likely result in a better prediction, instead of letting our validation do its thing. I cannot relate a topic in class that would make me make a 0ne to one comparison to what a self-cancelling prediction" is, but a good example I find is out predictions on global warming, if people came to the realization that something most be done and change their ways to improve this issue, your model will fail, which is what we want, which can sound counter intuitive but ultametly is what we are aiming for.

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

The SIR model of infectious disease tends to underfit rather than overfit. This is primarily because the model is based on oversimplified assumptions that do not realistically represent the complexities of disease spread. The model assumes everyone has the same likelyhood to contract the disease, as well as likelyhood of recovery and many other factors which do not account for variations across different demographics, such as age, race, or behavior patterns.

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

What Norbert Weiner meant by "the best model of a cat is a cat" is that in a perfect, deterministic world, the best model we can achive is the actual phenomena. In reality the closest we can get are models.

(g) [easy] Not in the book but about Norbert Weiner. From Wikipedia:

In the context of this class what we mean by feedback mechanisms is evaluation different models by their accuracy metrics such as RMSE for OLS, when we create different methods and get the minimum, that is our feedback to keep improving them, as we know that by minimizing certain metrics we can get a better model.

> 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.

What gives Voulgaris and edge in his betting is that he was very specific insight into the workings of the phenomena (in his case basketball) that he is trying to predict, by hyperanalyzing he is able to create feautures that regular models would over look.

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

Many scientific studies struggle with reproducing their studies due to a variety of factors: insufficient detail in experimental procedures makes replication challenging; publication bias favors positive results, overshadowing equally important negative outcomes; statistical variability can lead to reproducible results by chance in underpowered studies; selective reporting may omit non-supportive data; financial and competitive pressures rush results and hinder thorough validation; and the complexity of some experimental conditions, sensitive to minor variations, complicates consistent replication. Addressing these issues requires robust experimental design, greater transparency, and a cultural shift to value replication as much as original discovery.

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

Fisher was bias towards tabaco companies as he was an avid smoker, between this and being an contracted advisor for such companies is easy to assume that his bias played a havy part to why he didnt believe that smoking caused lung cancer.

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

The world is moving on "Bayesianism", Bayesian methods are more flexible and are able to incorporate prior knowledge into their analysis.

(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?

Fisher had a very deep understanding for the game of chess, he had a different way of seeing it as others due to this profound understanding.

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

The Y that google uses to predict "usefulness" of the search result. The reason why is because when a search is queried, what google best expect is what the subject means. For example, for a query like "Basketball Jersey", do you mean most expensive collectable? NBA Store? Most Famous number? This approximation to the meaning is what google thought would be the best metric to predict.

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

In this class we call Googles theories in this class as $\bar{y}$, which are all the outputs for particular feature inputs, and the testing these theories is performing model validations.

(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?

The practical advice given on page 315 for aspiring data scientists stresses the importance of grasping the basics of model building and data analysis beyond just using automated tools. These easy-to-use tools often hide the essential details and principles needed for sound decision-making. Taking a class that covers these fundamentals gives you an advantage by allowing you to not only operate these tools more effectively but also understand their limitations and biases. This deeper understanding equips you to adjust algorithms to better fit specific situations, providing a significant edge over those who may only know how to use the tools without understanding the underlying processes.

(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.

Silver suggests that success often stems from the ability to think probabilistically. He emphasizes the importance of understanding that results are influenced by both skill and luck, and successful individuals tend to be those who can navigate this complexity effectively.

(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

Nate Silver argues that it may not be advisable to completely remove humans from the predictive process even after a good model has been built. He emphasizes that while models are powerful tools for prediction, they still require human oversight for interpretation, decision-making, and adjustments based on new data or unforeseen circumstances.

(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?

According to Fama, a mutual fund that performs spectacularly in one year but fails in subsequent years can be explained using the notion of randomness or "noise" in financial markets. Using the notation from this class, such performance (denoted as y_t for year t) can be viewed as an outlier or result of specific favorable conditions present in that year but not necessarily reproducible.

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

The Manic Momentum model did not validate effectively, as described in the text. This model, which attempted to predict stock market trends based on momentum, ultimately failed to produce consistently reliable or reproducible results when tested over extended periods and different market conditions.

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

Noticing stock market bubbles while we are in them is quite challenging. While in theory, signs of a bubble—such as extremely rapid price increases or high trading volumes in the absence of corresponding fundamental value—are observable, in practice, it's difficult to distinguish these signs from normal market behavior at the time.

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

Shiller's model, particularly his use of the cyclically adjusted price-to-earnings (CAPE) ratio, indicates that high CAPE ratios suggest overvalued markets with potentially lower future returns for long-term investors. Conversely, low CAPE ratios may signal undervalued conditions, offering better long-term investment opportunities. This model advises long-term investors to factor in broader economic cycles rather than just short-term market trends, aiming for a more strategic approach to investing based on historical valuation levels.

(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?

In the text, Silver's heuristics were "Buy stocks when there is blood in the streets." This heuristic works so well because in the market, in moments of economic fear, many people decide to sell due to fear of losing value. That's when there is blood in the water and can buy for a cheaper price. Which has been an effective heuristic to have.

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

Something that would prevent me from executing on it is that it would alarm people about a potential pop, and influence the model in the way he interprets self-fulfilling prophecies.

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

Heuristics can get us in trouble when we start applying them to more complex problems. When we ignore other factors that can be consider by focusing only in the heuristics, we set a limit to how much we can apply these models.

Problem 2

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?

We were trying to minimize misspecification error by fitting a more complex function to our dataset. The mathematical theory was to use interpolation to better map the trained data, The interpolant then experienced Runges phenomena at the end behaviors of the functions making it less than ideal of a solution.

(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.

The interpretations of b_1 is the coefficient for our feature x, and b_2 is the coefficient we determined for our feature x^2 through OLS. The interpretation of OLD coefficients is that when we are trying to minimize the errors between the data and the best model, some coefficients would being the function closer to those points, this is why we are minimizing the coefficient vectors by checking the RMSE and R^2s of the function.

(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?

It would not be beneficial at all to predict in such a small scale when it comes to polynomials, in our example we need to have the funcitons X's be .1 away from each other, neededa masive separation in both coeffiecents.

(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.

To show that a 10% increase in x_2 results in an approximate $0.1b_2$ change in the response $\hat{y}$, consider the model $\hat{y} = b_0 + b_1 x_1 + b_2 \ln(x_2)$. The coefficient b_2 multiplies the natural logarithm of x_2, indicating how changes in x_2 affect $\hat{y}$. When x_2 increases by 10%, the new value of x_2 is $1.1x_2$. The change in $\hat{y}$ can be approximated using the derivative with respect to x_2, which is b_2/x_2. Applying this derivative to the change in x_2 (which is $0.1x_2$), the change in $\hat{y}$ is approximately $0.1b_2$. Thus, a 10% increase in x_2 leads to a predicted increase of approximately $0.1b_2$ in $\hat{y}$. This is derived from the fact that the derivative b_2/x_2 captures the rate of change of $\hat{y}$ in response to changes in x_2, and multiplying this by the change in x_2 (10% or $0.1x_2$) gives the resultant change in $\hat{y}$.

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

We expect the previous approximation to work when we are dealing with 2 correlated features which are being mapped to a ln function, we would not be able to make this work when we are dealing with higher level functions.

(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.

The interpretation of this function is "transforming" the data to create a linear relation between the features and the outputs, when we map both to ln the both can increase or decrease monotonically. The approximate interpretation of b_2 is the coefficient with the most "weight" in the model.

(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.

Consider $\bar{y} = b_0 + b_1 ln(x)$ We denote the change in $\bar{y}$ as $(b_0 + b_1 ln(x_*) - (b_0 + b_1 ln(x_f))$, then factor out the b_1 and combine like terms eliminating the b_0 resulting in the following function $b_1(ln(x_f) - ln(x_0))$ we simplify this function with properties of ln to the following, $b_1(x_f/x_*) = b_1 \delta x$. If we normalize the $\bar{y}$ with ln, we get the following equation, $\hat{y} = e^{b_0 + b_1 x}$ then we can substitute these e's as m_0 and m_x and we get the resulting function $\hat{y} = m_0 m_1^x$.

Problem 3

These are some questions related to extrapolation.

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

Extrapolation is a method we use to predict on future data using current historical data to train our models, as an example, if we use training data from 2000 to 2010 to predict in 2020, this is extrapolating. The reason why this is net-negative is because we have no knowledge of the future. We also have no access to details like stationarity for the features which can dramatically change our predictive outcomes.

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

Linear, logistic, and polynomial regression models handle extrapolation in distinct ways due to their structural differences. Linear regression assumes a continuous linear relationship between variables, which can lead to unrealistic extrapolations if the true relationship deviates from linearity outside the observed range. Logistic regression, used for binary outcomes, typically plateaus towards its minimum or maximum class probability, making its extrapolations somewhat conservative but potentially misleading if new data points suggest different probabilities. Polynomial regression, capable of fitting more complex curves, may offer better fit within the data range but can behave erratically when extrapolating, especially with higher-degree polynomials, often leading to significant prediction errors as the model's response becomes highly sensitive to inputs outside the range seen during training. Each model's extrapolative behavior underscores the importance of understanding underlying assumptions and limitations when applying them to data outside the observed scope.

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

The reason why polynomial regression models suffer terribly from extrapolation is due to runges phenomena, this causes irratic end behavior on the ends of a closed interval for polynomials, which when applied to extrapolation would cause massive errors.

Problem 4

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

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

The fundamental problem of "model selection" is that we will never know the actual f function due to its complexity, this forces us to be limited and never able to achive a "perfect" model, but we can get really close.

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

When using 2 splits of data, if the data before splitting was randomized, we can have 2 models running on these subsets of data, and whichever metric we use to measure accuracy would help us select which model was best at predicting with this data.

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

The limitation of 2 splits can be the following, if the dataset is small it can lead to overfitting for both models as there is not enough data to predict on, another problem is that the subsets can by chance have some sort of unforeseen correlation, making their predictions bias to the data and not applicable with future data.

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

Slowly increasing the amount of "splits" seems to be beneficial as long as the dataset is large enough, it seems that the more splits we have the less relation that is experience b.

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

Nested cross-validation, employing both inner and outer folds, significantly enhances model selection and evaluation by clearly segregating model tuning and performance assessment. Inner folds are utilized for hyperparameter tuning and model selection, ensuring that these choices are optimized across various data subsets without compromising the integrity of the test data. Outer folds independently validate the model, providing a robust estimate of the generalization error and reducing the risk of overfitting. This layered approach ensures each data point is used both for training and as part of the test set, leading to more reliable and effectively validated models that are likely to perform well on new, unseen data. This method is particularly valuable in achieving an unbiased evaluation and obtaining models that generalize well in real-world scenarios.

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

The way that g_{final} is constructed is by the following procedure. First we fit all the m models in D train to obtain $g_1(\bar{x})$, . . ., $g_m(\bar{x})$, then we predict all models on D select to obtain S_e1, . . ., S_em, then we select $g^*(\hat{x})$ which has the best oos performance metric. Then predict on D test using $g^*(\hat{x})$ to get a conservative estimate of future predictions preformance and lastly, we build the $g_{final}(x)$ on all D the $g^*(\bar{x})$.

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

To use nested resampling for hyperparameter tuning in models, divide your dataset into three outer folds, and split the data not included in each outer fold into three inner folds. Employ hyperparameter tuning methods like grid or random search within these inner folds to train and validate different hyperparameter combinations. The best-performing set from the inner folds is tested against the outer fold to assess generalization. Repeat this for each outer fold, choosing the best hyperparameters based on average performance or reapply the tuning on the entire dataset. This procedure ensures that the chosen hyperparameters are robust, minimizing overfitting and optimizing the model's performance on unseen data.

(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 $p \gg n$.

If you square features $x_1, \ldots, x_{p_{raw}}$, you would make n^2 features when you ignore like terms.

(i) [easy] Describe the methodology from class 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.

The step-by-step procedure for CART algorithms is the following, first you consider all possible orthogonal - to axis splits, where $x_1 <= x_1(1)$ which is the minimum for x_1, $x_1 <= x_1(2)$ which is the second minimum for x .... For each split, there are two "daughter" nodes. Assign $\hat{y} = \bar{y}$ of the responses in the nodes. Calculate SSE in each node.

$$SSE = \sum_{i \in \text{Nodes}} (y_i - \hat{y_{\text{node}}})^2$$

Locate the "best split" by minimizing the following objective function.

$$SSE_{\text{weight}} := \frac{n_L SSE_L + n_R SSE_R}{n_L + n_R}$$

and create the split. Repeat steps 1-2 recursively for each daughter node until daughter node has $<= N_0$, then node size (hyper-parameter).

(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.

Instead of using the mean you can use the average for the "leaf assignments", they are both very similar metrics conceptually but in practice depending on the model it might be beneficial.

(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.

To optimally prune an overfit decision tree, begin with a fully grown tree where each leaf holds a single observation. Systematically check each inner node to see if both its children are leaves, indicating eligibility for pruning. Prune by selecting the inner *node that, when converted into a leaf with both its daughters removed, results in the smallest increase in sum of squared errors (SSE). Replace this inner node with a new leaf whose predicted value is the average of responses from the removed daughter nodes. Continue this process iteratively, pruning one node at a time until no eligible inner nodes remain or a predefined complexity criterion is met. This pruning algorithm reduces tree complexity, minimizes overfitting, and enhances the model's generalizability by simplifying the decision tree in a controlled, step-wise manner.*

(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).

You repeat the steps for Regression trees where the splits are measured by the "gini" metric instead of the SSE. To calcualte the Gini you take the average count of the "classes" that we are trying to classify.

(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 that i learned while researching for this question is Entropy. Entropy in the context of decision trees measures the disorder or randomness in the data. The formula for entropy for a classification problem with classes defined as the following.

$$H(S) = -\sum_{i=1}^{k} p_i \log_2 p_i$$

where:

- $H(S)$ is the entropy of set S,
- p_i is the proportion of the class C_i within set S,
- k is the number of classes.

We then want to reduce the "information gained" which is the difference between the original entropy and the weighted entropy of both subsets. We represent Information Gained as,

$$\text{Information Gain} = H(S) - \left(\frac{|S_1|}{|S|}H(S_1) + \frac{|S_2|}{|S|}H(S_2)\right)$$

where:

- $H(S)$ is the entropy of the entire dataset before the split,
- S_1 and S_2 are the subsets of S after the split,
- $|S|$, $|S_1|$, and $|S_2|$ are the sizes of the sets S, S_1, and S_2 respectively.