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

Loyd Flores

Monday 15th April, 2024

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 hard because of the unpredictable nature of the flu virus itself, which can vary significantly from year to year in terms of severity and methods of transmission. Additionally the data available during a flu outbreak can sometimes be inaccurate or incomplete which complicates the task of making reliable predictions. What makes things worse is that flu symptoms are relatively common, being found in other illnesses, in this case there is a lot of noise, leading to potential misdiagnoses.

The models predicting flu outbreaks typically suffer from both systematic and random errors. There is a high bias that does not align well with the actual dynamics of the flu spread. There is also high variance or noise in the data which introduces unpredictability that is difficult to account for the models when making accurate predictions.

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

In the book, Nate Silver describes extrapolation as a way of predicting the future by assuming that the current events will just continue occurring, in simpler terms Silver extrapolation assumes stationarity. For example it's like assuming that a person will keep growing taller the same rate as they did as a kid, which is not really realistic. In our class we described extrapolation as making predictions based on information

from a certain set of data but applying it beyond that specific set, or the predictions being outside of the sample set.

The only conflict I saw was both agreed that predictions that extrapolate can sometimes be tricky and inaccurate because things change in ways that the original data can't accoutn for.

(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.
A good example of failure of predictions due to extrapolation is *Thomas Watson's* , chairman of IBM, quoted that "There is a world market for maybe five computers." Watson extrapolated from the large size, high cost , and complexity of early computers, failing to foresee the rapid advancements in technology that would lead to widespread development in techonology making them cheaper and more accessible to everyday consumers.

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

Based on the notation used in class, Self-fulfilling prophecy is like bias. It occurs when assumptions made by a model do not fully represent the underlying data or when a model is overly simplified. Self canceling prediction could be linked to variance. Variance refers to the amount which a model's predictions would change if different training data were used. High variance usually suggests that the model is overly complex, capturing the noise rather than the signal in the data set , leading to poor performance on new, unseen data. For example if there is a model that predicts the market crash and that model is overly complex with high variance, meaning it is sensitive to its input data, it could over estimate a slight change of the market and misinterpret it as a market crash allowing people to respond before hand cancelling out the potential effect of the crash hence making the model wrong.

(e) [easy] Is the SIR model of infectious disease under or overfit? Why? The SIR model which stands for Susceptible, Infected, Recovered, is a basic mathematical model that was used to describe how disease spreads within a population. It did so by grouping people into three bins. People who are susceptible to catch the disease, infected individuals that could spread the disease even more, and people who have recovered and are immune. Immediately we can see that this an insane simplification of the problem. It is Severely under fit due to its simplicity. The model fails to consider people who get re-infected.

(f) [easy] What did the famous mathematician Norbert Weiner mean by "the best model of a cat is a cat"?
Norbert Weiner elaborates the fundamental truth about modeling. No matter how sophisticated or well-developed they are, it is still merely an abstraction or simplification of reality. Models still have limits and is best only as a reference or a guide. The only thing that comes close to a cat is a real cat.

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

> 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?

I think the quote highlights the importance of validation in modeling. It played a big role in the development of statistical modeling and theory and is still a cruicial step in the machine learning pipeline. Great machines produce great outcomes when they are validated. The better score you attain after validation the better it will perform in the real world. In the context of our class the best way we can attain the best model is using validation techniques such as k-fold-cross validation which offer honest metrics, leading us to believe that our model can *reproduce* its results on the validation and test set on the real world phenomena.

(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.
Voulgaris is a prime example of the quote that I'll coin, ***Data is KING!*** The edge Voulgaris has built against opposing sports bettors in the NBA is through his data-driven decision making. He utilizes detailed game data and statistics which he continously collects and integrates. He also has knowledge regarding Game theory and the intricacies that randomness presents as well as the intricate rules of what makes basketball a game filled with technicalities. Voulgaris doesn't stop there he also feature engineers new metrics that could be of importance such as the effect of the coaches and how they manage game clocks.

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

There could be a lot of reasons but here are the ones that stand out to me the most:

- Complexity of Experimental Design / Modeling : Some scientific experiments involve complex designs and specific conditions that are hard to replicate. Small differences in the process or setup could lead to drastically different results.
- Sample Size : Studies that have low sample sizes could provide results that are not generalize or just happened due to chance.
- Selective reporting / Bias : maybe scientist only publish work they want to release because they attained results that were in favor of their personal biases.

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

Fisher was a strong believer of experimental design and rigorous statistical analysis to establish causality. He thought thtat the observations between the link of smokeng

to lung cancer were not controlled experiments and he believed that correlation did not equate to causation. Or he could also just been very bias since he was a smoker himself. People believed that he had ties to the tobacco industry and was a heavy opposition due to the pursuit of personal motive.

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

The world is not moving exclusively towards either Fisher's frequentism or Bayesianism but is rather embracing both, depending on the context and needs of specific applications. Bayesian methods are gaining popularity due to their flexibility in incorporating prior knowledge, comprehensive uncertainty estimation, and advancements in computational tools that make them applicable to complex models. However, frequentist approaches remain foundational in many fields, especially in regulatory contexts and large-sample scenarios. Overall, there's a growing trend towards a pluralistic approach in statistics, where both Bayesian and frequentist methodologies are used complementarily to address diverse scientific and analytical challenges.

(l) [easy] How did Kasparov defeat Deep Blue? Can you put this into the context of over and underfiting?

Garry Kasparov and IBM's Deep Blue chess computer in 1996 resulted in a 4-2 win for Garry Kasparov. Deep blue is a perfect example of both overfit and underfit. It was extremely overfit and seemed to posses overly complex algorithms to detect all possbile positions of well-known plays that are common to the game. It probably will be undefeated in situations where the match is structured or the opponent is utilizing popular strategies that Deep blue has seen before but extremely underfit to unorthodox plays that a grandmaster's heuristic can utilize. It just did not know how to react or generalize to things it has never seen before.

(m) [easy] Why was Fischer able to make such bold and daring moves?
Bobby Fischer is known for his bold and daring moves in chess. He was able to execute such strategies due to a combination of exceptional talent, innovative thinking, deep understanding of the game, and to finish it off strong psychological acumen. He knows what moves an opponent can play a few moves ahead and has the ability to respond to opposing advances while setting up his own advance.

(n) [easy] What metric y is Google predicting when it returns search results to you? Why did they choose this metric?
When Google returns search results to you, it is predicting and optimizing a metric closely related to relevance and user satisfaction. These metrics are typically assessed through various direct and indirect signals that might include click-through rates (CTR), time spent on a page, bounce rates, and user engagement levels. The ultimate goal of Google's search algorithm is to deliver the most relevant and useful information for each user query.

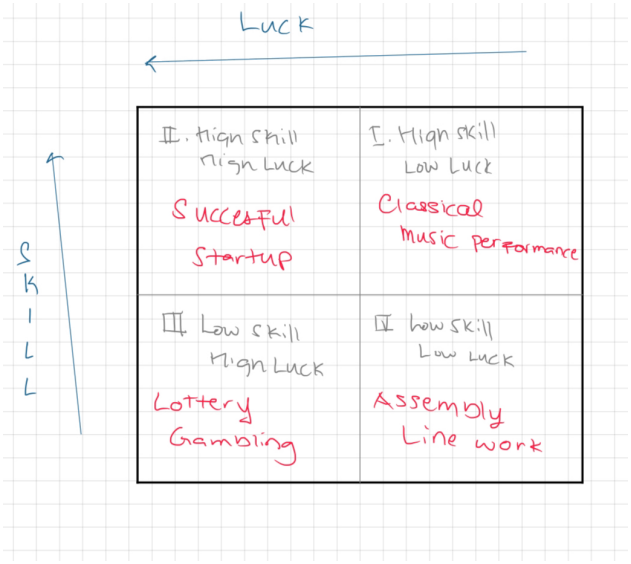
(o) [easy] What do we call Google's "theories" in this class? And what do we call "testing" of those theories?

In the context of this class *theories* could be the A or the algorithm that a model could utilize to learn from the data. There are infinite possibilities from different models to their individual set of hyper parameters. Given that we have to test the multitude of models that we develop to ensure that we come up with the best one. Since theories are algorithms, we could then test our theories by validating the model that the algorithm using validation techniques. This ensures that we truly have the best possible model that appropriately learned the patters of the data.

(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?

In the realm of data science, while push-button tools that automatically fit models can be highly efficient and user friendly, they often lack flexibility and may not provide the deepest insights to the data's underlying patterns and nuances. For an aspiring data scientiest, gaining a competitive edge over those who primarily rely on these tools involves several key aspects which are :
- Deeper understanding of the fundamentals : Understand the intricacies of models and know which one to use for specific situations - Data Manipulation and Preparation Skills : Data scientists with strong fundamentals would know every step of the pipeline which could in tern provide better models by utilizing the right data set, features, model, validation technique, etc. - Customization and Optimization opportunities - Ethical judgement and bias recognition - Can communicate findings effectively

(q) [easy] Create your own 2×2 luck-skill matrix (Fig. 10-10) with your own examples (not the ones used in the book).



2X2 Luck - Skill Matrix

(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.

Billing's algorithms, like many others, may struggle in no limit hold'em due to the game's complexity and need for nuanced decision-making. No-limit hold'em involves not only strategic betting but also bluffing, hand reading, and adjusting to opponent's play styles, which are challenging for algorithms to accurately model. Additionally, the immense number of possible game states in no-limit hold'em makes it difficult for algorithms to explore and evaluate all options effectively, leading to sub optimal decisions.

(s) [easy] Do you agree with Silver's description of what makes people successful (pp326-327)? Explain.
I agree with Silver's take on the role that luck plays into the reasons for success. It is true that people need to be skilled enough as well as disciplined enough to keep working and improving but the opportunity to rise is the biggest factor. The best programming students could exist everywhere but if they don't apply for jobs they may never be discovered. The converse is true a lot of unskilled developers are out there due to opportunities such as referrals that put them in front of more deserving people. CS Students can study and code all day receiving the highest grades and creating the coolest projects but if they dont take the time to network and allow themselves to become a desirable candidate / applicant, they will forever live in the shadows to wonder why they have no jobs.

(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
This is a classical case of Human vs Automation. Like all things there are trade offs. In this specific domain for predictive enterprise, Humans could be removed from the entire mathematical computation but for analysis I think humans are still required. Yes a model can also analyze and interpret the data using Large language models like CHATGPT but If everything is fully automated that can assume stationarity. Things may change and the model may be inaccurate with time. Second the model itself maybe good but bias, a human interpretting the results could act as a checks and balance between the model the same way that a model may remove a human's bias. Models are also just machines that spit out numbers, they have no ethics. To remove a human from automation may prevent the generalization of factors that the model did not consider. Overall I think humans can be removed in some aspects but humans are still necessary in analysing and interpretting the models results. It is a human's job to validate the model and its results because at the end of the day models are just tools.

(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?
To explain using the context of our class the reason that a mutual fund that performs well but is unable to replicate its success could be a victim of high bias that year. It may just be that at that specific year it really did well making people believe in that

mutual fund. Another thing is maybe the analysts assumed stationarity that every year forward it would perform well which is never usually the case.

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

The Manic Momentum model, discussed in Nate Silver's book "The Signal and the Noise," relates to the prediction of financial markets, particularly focusing on stock prices and their movements. The concept of "Manic Momentum" refers to the idea that stock prices can exhibit momentum that is, trends in stock prices can persist in one direction for a period of time before possibly reversing. It was validated through several ways such as **Market efficiency**, **Volatility**, **Empirical Evidence,** etc. We can conclude that the Manic momentum does capture some aspects of real-world market dynamics, especially thing that persist in the short term.

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

Noticing stock market bubbles while we are in them is relatively hard because during a bubble, market sentiment is often overwhelmingly positive, which can cloud judgment. Investors may become overly optimistic about the continued growth and ignore sings of overvaluation. The rapid increase could also encourage other investors to ride the wave, this is called herd mentality. It also isn't that simple to label and identify a bubble since there is no precise or universally accepted definition of what a bubble is, which makes it hard to identify one until after it has burst.

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

For long-term investors, Shiller's model implies that when stock valuations are high relative to historical earnings, future returns are likely to be lower. Therefore, adjusting asset allocations based on valuation levels and taking a more conservative stance when valuations are high and might improve long-term investment outcomes.

(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 Nate Silver's book, he discusses the heuristic of assuming continuity over change in uncertain situations. Specifically, he mentions, "When something is new or hard to understand, bet that the trend will continue rather than that it will reverse." This heuristic simplifies decision-making in complex situations by suggesting that existing trends are likely to persist. In many areas, once a trend is established, it often continues due to underlying forces. For example, in economics, markets that are rising or falling tend to continue to do so. Systems, whether they are natural, economic, or social, typically have inherent resistance to change. It requires a lot more forces to change which is not usually common.

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

Even with a reliable model for predicting financial bubbles, practical challenges like the difficulty of timing the market accurately, psychological pressures, market impact of large trades, regulatory constraints, financial limitations, counterparty risks, and potential feedback loops can all hinder the successful execution of trading strategies based on such predictions. These factors make it complex to act on theoretical models in the dynamic and often unpredictable real-world financial markets.

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

Heuristics simplify decision-making but can lead to problems such as overgeneralization, embedded cognitive biases, resistance to new information, misjudgment of probabilities, and an illusion of validity. They can also result in stereotyping and neglecting the complexity of situations, potentially leading to discriminatory practices and oversimplified solutions to complex problems. These limitations highlight the need for careful management and critical evaluation of heuristic-based decisions.

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?

The issue that occurs was when the data is overly complex that our candidate set $\mathcal{H}$ was unable to find any set of functions that could fully grasp the complexity of $\mathcal{D}$. For example if there is an upward curve trend within the data and our $\mathcal{H}$ only consisted of linear functions we would incur a lot of mispecification error. The solution we introduced was resorting to higher term polynomials. Since the computer just takes input we can trick it by utilizing higher level polynomials that are still in the form of linear functions. It is also convenient mathematically because OLS will still be a viable option despite making the equation more complex.

(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.

b_1 is the linear coefficient of the model. It represents the linear effect of the predictor variable x on the response variable $\hat{\boldsymbol{y}}$. b_2 is the coefficient of the quadratic term x^2. It quantifies the curvature effect of x on $\hat{\boldsymbol{y}}$.

(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?

Given the narrow range of x, estimates b_1 and b_2 should be treated with caution. They are likely to be very sensitive to the specific dataset and might not generalize well outside this narrow window. If the model's purpose is to make predictions within a similar narrow range or to understand the relationship in this specific interval, the estimates might be useful. However, for broader applications or predictions over a more extensive range of x, these coefficients may not provide reliable or trustworthy insights. Additional data covering a broader range of x values would be beneficial to test the robustness and applicability of the model.

(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.

The explanation provided shows how a percentage change in x_2 affects the response variable $\hat{\boldsymbol{y}}$ in the model $\hat{\boldsymbol{y}} = b_0 + b_1 x_1 + b_2 \ln(x_2)$. Here's a summary of the key steps:

(a) **Model Derivation**: The model suggests that changes in x_2 through the natural logarithm of x_2.

(b) **Rate of Change Calculation:** The derivative of $\hat{\boldsymbol{y}}$ with respect to x_2 is b_2/x_2. This derivative indicates how much $\hat{\boldsymbol{y}}$ changes for a small unit change in x_2

(c) **Effect of Percentage Change**: When x_2 increases by a percentage p%, x_2 changes to x_2(1+p/100). Using the properties of logarithms, the change in $\hat{\boldsymbol{y}}(\hat{\boldsymbol{y}})$ can be estimated by $b_2$ln(1+p/100).

(d) **Approximation for Small** p: For small percentage changes p, the logarithm ln(1+p/100) can be approximated by $p/100$ using the Taylor expansion. This results in ($\hat{\boldsymbol{y}} \approx b_2$(p/100)

(e) **Interpretation** : Thus, if x_2 increases by 10%, the increase in $\hat{\boldsymbol{y}}$ is approximately $0.1b_2$. This calculation shows how the model predicts a proportional change in the response based on a percentage change in x_2

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

The approximation $\ln\left(1 + \frac{p}{100}\right) \approx \frac{p}{100}$ is very useful for quick calculations and is often sufficiently accurate for practical purposes when dealing with small changes. However, for larger changes, for systems where precision is paramount, or where the response is highly non-linear, more robust methods of analysis should be considered. The decision to use such an approximation should be guided by the context and the acceptable level of error in the specific application.

(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.

In summary, b_1 quantifies the multiplicative effect on $\hat{y}$ for a one-unit change in x_1 and can be interpreted straightforwardly as an exponential effect. The coefficient b_2, on the other hand, represents the elasticity of $\hat{y}$ with respect to x_2, indicating how much $\hat{y}$ changes in percentage terms in response to a 1% change in x_2. These interpretations are essential for understanding the impact of variables in models where logarithmic

transformations are employed, providing insights into the proportional and relative effects of changes in predictor variables.

(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.

Step 1: Original Model :
$\ln(\hat{y}) = b_0 + b_1 x_1 + b_2 \ln(x_2)$

Step 2: Exponentiate Both Sides
$\hat{y} = e^{b_0 + b_1 x_1 + b_2 \ln(x_2)}$

Step 3: Simplify
Utilize identity $e^{ln(a)} = a$:
$\hat{\boldsymbol{y}} = e^{b_0} \cdot e^{b_1 x_1} \cdot x_2^{b_2}$

Step 4: Define Constants
Define $m_0 = e^{b_0}$ and $m_1 = e^{b_1}$
$\hat{\boldsymbol{y}} = m_0 \cdot m_1^{x_1} \cdot x_2^{b_2}$

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 when you predict future events based on trends from existing data that go beyond the range you originally studied. For example, if you have only seen the sales data for the first six months of the year, and you try to guess what will happen at the end of the year, you're extrapolating. This method can be risky and often produces inaccurate results because it assumes that current trends will continue unchanged. For instance, if you noticed more ice cream sales in the early summer months and extrapolated that this would continue, you might predict extremely high sales for December, which is unlikely due to the colder weather reducing demand for ice cream.

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

Yes, different models extrapolate differently based on how they are built and what they assume about the data. For instance, a simple model might assume that sales grow by a fixed amount each month, so if you tell it that sales have been increasing from January to June, it might just continue that trend into the future. More complex models might consider other factors like seasonal changes or economic conditions, which could lead to different predictions for the future. Each model has its own way of handling data outside of its original range, leading to varying results when extrapolating.

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

Polynomial regression models suffer terribly from extrapolation because they follow curves that can change dramatically outside the range of data they were trained on. For example, if a polynomial model is used to fit a set of points that rise slowly, it might curve sharply upward or downward when trying to predict values beyond those points. This makes the model very unreliable for predicting outside its original data set, as it might give extreme values that don't make sense with real-world behavior.

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 each hyperparameter, each algorithm, the amount of data you use, etc, will all produce different models. There is a multitude of models you can come up with. The problem is how do we know if we obtained the best model possible that generalizes well on unseen data. We would need a model that is closest to f.

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

Splitting the dataset into two implies that the majority off it will be used to train and the remained will be used for testing which we would give you a rough estimate on how well the model will do.

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

The test set could only be used once. If you utilize it multiple times your model can over fit. The dilemma is how do i make adjustments to my model while being able to have honest metrics.

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

Three splits introduce a new split called validation set. This is an intermediarry step that we is part of the training set. We could train the model and proxy-test it using the validation set which we can use again and again, with the possibility of overfitting to the validation set increases if you are not careful, before testing on the actual test set. This allows more flexibility during 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?

Using inner and outer folds means that the inner fold is used for tuning the hyperparameter/ selecting the best model and the outer fold is used to show how well the model selected from the inner fold will perform on unseen data. As the name implies, this is cross-validating and will help in improving the model selection procedure.

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

When using nested resampling on three splits, the inner fold determines the best model/ hyperparameter. The selected model/hyperparameters from each inner fold are retrained on the entire outer fold training set. The gfinal is the final model that is ready to be used on unseen data

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

Model selection for models that have hyperparameters in their algorithms could be done iteratively. For example x is a hyperparameter. We could have give x a range for example 1-100. Iteratively train, test, collect error metrics. Iterative checking could be done by a foor loop. Once we search all possible outcomes we can just select the combination of hyperparemeters that produced the lowest errors.

(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 $p \gg n$.

To produce the most expansive set we could Generate all polynomial combinations of the features up to a certain degree. For example, with two features x_1 and x_2, and up to degree 2, you'd create more complex versions of the raw features. The next thing we can do beyond simple polynomial terms is to include interaction between terms for different features.

(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.

First we generate the expanded set of features using the transformations discussed such as polynomial expansion and interaction of terms. Then feature selection. We utilize filter methods to select features that have a strong relationship with the target variable.

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) Start at the root : Begin with all the training data at the root node

(c) Feature selection : For each feature, attempt to find the best split that divides the data into two groups. Select the split with the lowest MSE.

(d) Actually split the data

(e) Create a leaf node : Establish a stopping criteria. Assign model prediction to leaf nodes.

(f) Once you identify the best rebuild the entire tree with the most optimal hyperparameters (Height) and best features.

(g) [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.

H for regression trees is every possible combination of features in D as indicators functions. There could be infinite ways to combine a trees predictors especially as n grows larger.

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

In regression trees, choosing a suitable leaf assignment rule is crucial for the model's performance. Besides using the average of the responses in the node, another effective rule for assigning values to leaf nodes is the median of the responses in the node. This allows us to be more robust since median is less sensitive to outliers compared to the mean.

(i) [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.

Steps:

Start Fully Grown : Begin with a tree where each data point in your dataset forms a leaf. Each leaf's prediction is just the target value of that single data point.

Check for Pruning Candidates: Look through all the internal nodes of the tree. Identify any internal node where both children are leaves (these nodes are candidates for pruning).

Prune a Node: For each candidate node, consider the two daughter leaves. Remove these leaves and turn the internal node into a new leaf. The prediction for this new leaf becomes the average of the target values from the two removed leaves.

Iterate: Repeat the process: After each pruning, re-evaluate the tree to find new pruning candidates. Continue pruning until no further internal nodes have two leaf children, or until you reach a predetermined stopping criterion (like a desired number of leaves).

Result: The end result is a simpler tree that is less likely to overfit the data since it generalizes by averaging responses over more observations per leaf.

(j) [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.

If the actual underlying relationship between the predictors and the response is linear, OLS, which directly models this linearity, would likely have better predictive accuracy. This is because it efficiently estimates the parameters with methods that minimize the sum of squared errors, directly targeting the core of the linear relationship. OTHER THAN THAT REGRESSION TREES IS KING!

(k) [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).

a. Start at the root : Begin with all the training data at the root node b. Feature selection : For each feature, attempt to find the best split that divides the data into two groups. Select the split with the lowest **missclassification error**. c. Actually split the data d. Create a leaf node : Establish a stopping criteria. Assign model prediction to leaf nodes. e. Once you identify the best rebuild the entire tree with the most optimal hyper parameters (Height) and best features.

(l) [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.

Entropy. This measure is a fundamental concept from information theory and is particularly effective in the context of decision trees for classification, often used in the construction of C4.5 trees, a successor of the basic ID3 algorithm.