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

Mohammed Zohair Hasan

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?

They are hard to predict because of variability and evolution of the virus and the dominant error is due to overfitting.

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

Silver uses the term extrapolation to describe the idea of using existing data to make predictions on new data. He uses the term "linear extrapolation".This terminalogy is conflicting because we cannot assume linearity.

(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.

I can think of Robert Goddard suggesting rockets could one day reach the moon, The New York Times criticized the idea, stating rockets could not work in space.

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

"self-fulfilling prophecy": A prediction that influences actions to make the predicted event occur. "self-canceling prediction:": A prediction that influences actions to prevent the predicted event from occurring.

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

Silver does not directly say whether the SIR model is underfit or overfit. Since this model is a more generalized model designed for general insights rather than precise predictions, we can say that it can be either or neither.

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

I interpret Norbert Weiner's statement "The best model of a cat is a cat" as no matter how sophisticated our models can be, the best and most accurate representation of something is itself. No matter how many factors or predictors we add to our model, it will never beat the representation of the phenomenon we are modeling.

(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?

In this context, a "feedback mechanism" is a system that uses outputs or results as inputs to improve future actions.

(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.

Bob Voulgaris uses statistical analysis and predictive modeling which gives him an edge in sports betting

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

Scientific studies are not reproducible due to specific conditions in which the studies took place before. Factors such as small sample size or experimental conditions are hard to not only fully list out, but also replicate on the stuff that is left out.

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

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

The world is moving towards Bayesianism which allows us to update out beliefs with new and upcoming evidence.

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

Kasparov defeated Deep Blue by using its flaw which made it overfitting its algorithm to expected scenarios and underfitting its response.

(m) [easy] Why was Fischer able to make such bold and daring moves?

Fischer basically built models in his head, calculating potential outcomes and predicting better than his opponents.

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

(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, we can address "theories" as models and "testing" as validating those models such as using some sort of error calculation metric that we discussed in class. One example of validating may be RMSE.

(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?

I can't sell this rigorous class short, the class has given me a deep understanding of data science, more specifically, the ability to interpret and create models as well as understanding whats happening under the hood, mathematically.

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

	Low Luck	High Luck
Low Skill	Slot machine	Lottery
High Skill	getting good test grade	Landing interview

(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.

There is a psychological element where people try to bluff which thew algorithm probably cannot account for.

(s) [easy] Do you agree with Silver's description of what makes people successful (pp326-327)? Explain.

I agree with Silver, high skill and hard work can make for a very lethal combination. I can think of an example such as someone at a job who is not only skilled at what

he does but also grinds out his job very hard, to me this seems like a recipe to get promoted very quickly.

(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

Well, one basic necessity for a model is maintenance. All models get outdated somehow, we need humans to feed new data and adjust the model accordingly.

(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?

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

The manic momentum did not consistently predict future trends accurately. The nature of the stock markets unpredictability itself makes it extremely hard to predict such trends.

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

No it is not noticeable while we are in them, and this makes sense. We can also refer back to the covid-19 years where the market seemed like it was flying high with there being an abundance of tech jobs and tech companies were thriving. But, we only realized in hindsight that this was just a delusion due to many people not working and living off unemployment and paying these big tech companies subscriptions such as Netflix.

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

A long-term investor using Shiller's model might be very cautious when investing in times of high valuations, they might assume that the returns will return back to the mean.

(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?

The heuristic that Silver quotes on p358 is that "the simplest explanation is usually correct." This quote works well because it helps someone living by this quote to drop the more unnecessary complexities. It is rather effective because it encourages focus on the most significant variable and avoids overfitting models with complexities that might just hide or reduce the main variable.

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

Considering the volatility of the market, one reason to prevent me from executing on it would be the timing. Predicting the exact time a bubble will burst is extremely hard to time.

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

While we discussed its simplicity being a good thing, very obviously this is also a huge negative factor to heuristics. When we are not considering the specific circumstances in which the heuristic does not hold true, it can cause us to not predict those exceptions properly. So if we refer back to the Silver quote on p358, we cannot use the simplest explanation to describe how a rocket might work in a physics class and although just saying the broad explanation might be acceptable in multiple settings, a physics class is an exceptions that cannot afford a simple explanation.

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 problem is that linear models cannot properly fit nonlinear relationships between the predictors and the outcome. The mathematical theory that justified this is polynomial regression. Polynomial regression can significantly improve model fit and predictive accuracy when there are nonlinear relationships between the variables

(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.

In the model $\hat{y} = b_0 + b_1 x + b_2 x^2$, the coefficients have the following interpretations:

- b_1 (Coefficient of x): This coefficient represents the linear component of the change in the dependent variable ($\hat{y}$) for a one-unit change in x.

- b_2 (Coefficient of x^2): This coefficient represents the quadratic component of the model, shows the rate of change in the slope of the relationship between x and $\hat{y}$ for a one-unit change in x^2.

(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?

Although we do have some confidence in b_1 and b_2, we can't expect to trust them because of the narrow range of x.

(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.

In the model $\hat{y} = b_0 + b_1 x_1 + b_2 \ln(x_2)$, to show how a 10% increase affects $\hat{y}$, we refer it as $x_2' = x_2 \times 1.1$. The logarithmic change being

$$\Delta(\ln(x_2)) = \ln(x_2') - \ln(x_2) = \ln(1.1x_2) - \ln(x_2)$$

and then we can simplify this as

$$\Delta(\ln(x_2)) = \ln(1.1)$$

This is how we can show the change in $\hat{y}$ caused by the increase in x_2

$$\Delta\hat{y} = b_2\Delta(\ln(x_2)) = b_2 \ln(1.1)$$

Using the approximation $\ln(1 + x) \approx x$ for x close to 0, we get:

$$\ln(1.1) \approx 0.1$$

This proves that if x_2 increases by 10%, $\hat{y}$ is predicted to increase by approximately $0.1b_2$.

(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 $\ln(1+x) \approx x$ for x close to 0 is best when the proportional change in x_2 is small. Mainly when it is close to 0

(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 the model $\ln(\hat{y}) = b_0 + b_1 x_1 + b_2 \ln(x_2)$, we interpret b_1 as the estimated percent change in $\hat{y}$ for a one-unit increase in x_1, holding x_2 constant. The approximate interpretation for b_2 is the $\hat{y}$ with respect to x_2. Meaning, b_2 is the estimated percent change in $\hat{y}$ for a 1% change in x_2.

(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.

Problem 3

These are some questions related to extrapolation.

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

Extrapolation refers to this idea that involves extending our model beyond the known input space to try and predict outputs for values that it had not encountered during training. This is a net-negative during prediction because we are relying on the output of values solely on the outputs for other non related values without any data. This introduces a higher risk of error and uncertainty in the range that we extrapolated.

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

Yes, one clear example would linear and non-linear models. A linear model may extrapolate simply by extending the straight line of best fit further. Meanwhile, non-linear models won't extrapolate in such a linear way, it can extrapolate with more complexity.

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

Polynomial regression models are more complex than just linear models. These models will give complex curves which fit the dataset rather accurately. Extending past the dataset will give very unpredictable and volatile lines.

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 with "model selection" is the idea that we need to pick the most optimal or "best" model implying that there exists this "best" model that will perform well on unseen data. "This involves balancing complexity (to fit the data well) and simplicity (to generalize well), commonly referred to as the bias-variance tradeoff." Online source says this and I find this a great way to state it.

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

When we use two splits we imply that one split is used for training known as the training set and the other is used for validation which is also named the validation set. Then we would train multiple models with different hyperparameters or different types of models on the training set to find the best one.

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

The main limitation is that the two splits will encourage optimistic estimates of the model's performance. This is due to overfitting the validation set.

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

With three splits you can now have a training, validating, and a third set known as the "test set" which implies that after training and validating the model, you can now actually test it to see how will the model will perform with unseen data to get an idea of how it might work with more unseen and new data.

(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 g_{final} 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.

We will use the inner fold of a nested resampling procedure to train models with each set of the hyperparameters. We will then evaluate then evaluate the model's performance on the validation set. We pick the best hyperparameter.

(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 $p \gg n$.

We can use polynomial transformations on p so that p $>$ n given $x_1, \ldots, x_{p_{raw}}$.

(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.

We start by polynomial transformations to generate the expanded set of features, then we do feature selection. We then use filter methods to pick features that are heavily correlated 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.

Steps: 1. We start at the root node 2. For each predictor we have to determine possible splits 3. Calculate MSE for each split 4. Choose the feature and corresponding split point 5. Once the criteria is met we terminate the node.

(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.

For regression trees $\mathcal{H}$ is every combination of features in the dataset D. There can be infinite ways to combine a trees predictors. This is more prevalent when n grows bigger.

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

Instead of average of the responses, we can do the mode of the responses. This method assigns the most frequent response value among the data points in the terminal node as the prediction value.

(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.

- Start with full tree - such that each observation in dataset D ends up in its own leaf.
- Setting pruning criteria - this will help determine the benefit in pruning
- Prune a node - for each node that we will prune, we can remove the daughter leaves and turn the internal node into a leaf
- We are left with a tree that is more simple to interpret

(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.

We can consider a purely linear relationship between the predictor X and the response Y. The vanilla OLS would model this relationship accurately as all conditions are fully met. On the other hand, in a regression tree, the relationship probably won't be captured well due to the nature of regression trees being piecewise. This is a trick question probably because the ability of OLS is overlooked when the conditions are met for it. Meaning, its ability to work on linear relationships is unrivaled.

(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).

- Start at the root
- for each feature, find best split

- select split with lowest misclassification error
- declare the criteria
- rebuild the tree with the best hyperparameter than we found

(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 is Entropy. We can use Entropy to compare "quality" of splits within inner nodes of a classification tree. Entropy excells in datasets where there is a lot of classes.