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

Benjamin Minkin

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?

This is because the prediction has an affect on how people act. If the prediction is that many people will die, people will in turn be extra cautious.

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

(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.

Blockbuster denying the purchase of Netflix comes to mind. They did not think digital streaming would take off like it has. In hindsight this moves seems silly. Another example was IBM predicting that the internet was just a fad.

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

self-fulfilling prophecy $y = \hat{\boldsymbol{y}}$ This causes a tautology.
self-canceling prediction $y \neq \hat{\boldsymbol{y}}$ This will never be true.

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

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

Sometimes there are substitutes that can work better than creating complex models.

(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?

This can be a model that trains off the same data that it tests on. The feedback loop will make the model think it can predict with a perfect accuracy when in reality it will have poor out of sample performance.

(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.

He was able to separate the bias of following the crowd from his independent thought about the likelihood of success of the lakers.

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

I believe that a lot of science is wrong. This is because of reasons including but not limited to innate human biases, Simpsons paradox and the FWER.

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

He enjoyed smoking and was biased.

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

The world is moving more towards using Bayesian methods.

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

Kasparov tried to create complicated positions that forced deep blue into following heuristics. He thought that because he was a better player than the the programmers, his heuristics would reign supreme.

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

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

It is predicting what you were searching for.

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

We called them model predictions. These can be tested with cross validation.

(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?

A lot of fields are zero sum games. This means that having a strong background can help an aspiring data scientist outperform the competition.

(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.

Poker is a game that has other components aside from math. In some aspects it is more important to play the player instead of playing the game. His model did not have the versatility to play the player so I do not think it would make a good poker player.

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

Yes. This is a type of survivor's bias.

(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

No. I do not believe it is ever smart to trust a model completely. It is, by definition, never a perfect representation of reality.

(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?

This is due to variation. Some models will outperform simply due to luck. We called this estimation error.

(v) [easy] Did the Manic Momentum model validate? Explain.
No. It lost money over time.

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

Possibly. This is hard to know for certain as people will always be saying that we are in a bubble. Then when one does occur, they claim to have seen it all along. This form of survivorship bias makes this type of prediction hard to assess.

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

In the long run, the only way to beat the market is to have information that hasn't be Incorporated into stock prices. This means that one should invest passively.

(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?

"follow the crowd, especially when we don't know better." This works because groups tend to make better decisions than individuals. This is similar to how averages tend to make better guesses than a random sample of one.

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

Some bubbles can last a long time before popping. Even after being identified, the bet against the bubble may become insolvent long before the burst.

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

Heuristics are just guidelines towards a goal. They are themselves just a means to an end. One should not get this twisted and chase heuristics instead of the goal he is trying to accomplish.

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 we were trying to solve was the issue of mispecification error in linear models. This method increases the candidate space $\mathcal{H}$ to include polynomials which should decrease mispecification error. The justification may be a theorem that supposes all continuous functions can be approximated with polynomials. This turned out to be a decent 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.

For interpretation, a one unit increase in x is expected to result in a b_1 increase in y. Similarly, a one unit increase in x^2 is expected to result in a b_2 increase in 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?

No. Based on the X, I believe using this model for interpretation can be risky. There needs to be a good a-priori reason to trust this 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.

Yes. This is what I explained above.

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

This approximation works in most cases however it may not work when x is negative.

(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.

A one unit increase in x1 is predicted to increase yhat by ln(b1) units. Similarly, an increase in one unit of x2 is predicted to increase yhat by one unit.

(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.

To cancel out the ln, put all terms as the power of e.

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 the process of attempting to predict on data from processes that were not used to train the model. Extrapolation is harder than interpolation partially because of overfitting and underfitting. It is not easy to ignore the noise and focus on the true relationships in the data. There also may be differing causal drivers between the training data and the extrapolated data.

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

Yes. For example, a linear model predicts on x values infinitely, even when those values are physically impossible. Those x values will likely have equally improbable $\hat{\boldsymbol{y}}$ values. This is different than the SVM model where new data change change the support vectors.

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

This is because of the higher dimensional parameters. These can amplify small errors especially in extrapolation.

Problem 4

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

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

This problem describes the balancing act between underfitting and overfitting. If the model focuses too much on the training data it may overfit and do poorly in extrapolation. Conversely, if the model doesn't focus enough on the training data, it will also do poorly in extrapolation.

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

I would train the model on one split of the data and test on the remaining split. Then, I would reverse the roles of the two splits for the next model.

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

This method will reduce the amount of data being used to train the model at any given time. This will decrease the effectiveness of the model.

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

Using three splits, I would train on two splits and test on the remaining data. I would repeat this process three times allowing each split to be used as the testing data.

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

This method reduces variance as it allows all of the data to be used as both testing data and training data.

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

The g_{final} will result from averaging the three models created from the data. This will create a 'final' model that has less variance than any one of the other models.

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

I would try many different values for the hyper parameters and keep the one that works the best.

(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 $p \gg n$.

This is the set of every single group of linear combinations of the raw features. These include exponents, sin waves, etc...

(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.

Test the functionality of the algorithm before and after adding each new feature to see whether it positively or negatively affected the algorithm. This is done step by step for each feature in a stepwise fashion.

Problem 5

These are some questions related to the CART algorithms.

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

Step 1. make a random tree
Step 2. at every split point find the best split of the features
step 3. take a random subset of the p features

(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.

This is the function space. It is a set of all "flow charts" that direct towards the classification output.

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

Another possible output can be the mode of the responses.

(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.

Rules:
1. If the parent and children node share the same key-value, combine into a single leaf.
2. Work from the leaves towards the root and combine in a fashion that minimizes the inaccuracy of the tree.

(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.

If within a regression pixel, there is a high variance between the highest and lowest values, the average of that pixel will be inaccurate.

(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).

Step 1. make a random tree
Step 2. at every split point find the best split of the features +
step 3. take a random subset of the p features

The only difference will be that you can't take the mean of classifications and instead will have to return the mode.

(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.

There may be a metric that describes the sections that can classify with perfect accuracy. For example, let accuracyMetric = number of sections that classify perfectly/total.