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

Osman Khan

Friday 12th 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?

We do not have enough data, as people do not usually go to the doctor unless they have a life threatening situation. The dominant type of error is estimation error, as we do not have enough data.

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

Silver defines extrapolation as the future following the trends of the past. This conflicts with ours, as we defined this to be overfitting.

(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.

- William Petty predicted that global population would be .7 billion in 2012.
- The Ehrlich couple predicted that predicted population to grow so much that million plus deaths would occur in the 1970s.

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

self-fulfilling prophecy: $g(x) = f(x)$

self-canceling prediction: $g(x) \neq f(x)$

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

SIR is underfit, as it uses very few proxies.

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

Only the cat itself provides perfect information. Every other thing (a model) will either underfit or overfit.

(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 the context of our class, learning from data, using cross validating, k-folding, inner & outer looping, are all examples of feedback mechanisms, where we use the information provided to us to improve our model. Otherwise, all we have to rely on is the null model.

(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 uses really good proxies. He keeps an eye and an ear for relevant information that could give him a clue about how the game might play out.

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

Most of the conditions it is conducted in does not repeat well.

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

I think he was more loyal than the king. Statistical inference is, after all, just inference. The data can provide support for a certain point of view, but it is supposed to give insight. Plus, he did not get lung cancer although he was a lifelong smoker.

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

I believe the world is moving in the direction of Bayesianism. People do have priors that are constantly updated as they get more data. Plus, Bayesianism has the ability to absorb complexities, which Frequentism doesn't, which can only hope for gathering as much data as possible.

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

Kasparov did moves that the Deep Blue was not programmed well for, which was why it was exposed.

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

Fisher was able to make such bold and daring moves as he did not have the experience of having played games where such raw-aggressive moves weren't always successful.

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

The usefulness of a result to our search query. Out of the possible metric, this was the most relevant.

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

We call Google's theories models and their testing, learning from 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?

The edge from taking this class is that we can develop our own models for markets that are not saturated. Thereby, we can be the big fish in a small pond, rather than being a small fish in a big, saturated lake.

(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	Hide & Go Seek	Rock-Paper-Scissors
High Skill	Ultimate Frisbee	Duelling

(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.

No-limit hold em is not as restrictive as limit hold em. The option at making more expansive bets, including all in, makes it difficult to build a successful computer program. In addition, the majority of fish in no limit are those who lack basic knowledge of poker.

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

I strongly aggree with Silver's description of what makes people successful. There are certain priors (family one is born in, where you live, luck), but those prior do get updated over time. Sure, some priors are a lot of dominant than others, but it is a Bayesian process at the end of the day.

(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. There is no such thing as a good model. Models can only approach it. It is a fallacy to be happy, if you lose although you made the right moves according to a "good model". The goal should be to improve predictions.

(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?

Due to variance in the performance of a model, a mutual fund did unreplicatively well in a year. The performance was not due to the predictive power but rather luck.

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

No, in order to validate, the investor would have needed to have left some money in the stock (5-10%).

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

Yes, they are, when the graph increases exponentially.

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

It is possible that there will be some dips in the price of the stocks. Despite that, the long-term investor should keep on holding on to the stock.

(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" is "follow the crowd, especially when you don’t know any better". It works well, because, at least in the short-term, if it was a terrible thing to do, so many people would not have been doing it.

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

We don’t know how long it will keep on expanding for. Plus, if we borrow money to short it, then we don’t have absolute control over when it is demanded to be returned.

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

They underfit. Plus, as they only work in the short term, we get into trouble over the long term.

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 expand our candidate set $\mathcal{H}$, so that it could better capture the complexities of the phenomenon. The mathematical theory that justified this soln was the Weierstrauss Polynomial Approximation thm. This turned out to be a good solution with caveats, a la the need to avoid runge’s phenomenon, and other overfitting issues.

(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 change in the coefficient of the slope of x
- b_2 is the change in the slope of the slope of x

(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?

The interval is really small, and changes in b_1 & b_2 can yield significantly different results. If we can have more information about the range of $\mathcal{Y}$, it would be helpful. The smaller the interval of $\mathcal{Y}$, the higher the trust.

(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 $0.1 b_2$. Prove this approximation from first principles.

$\hat{y} = b_0 + b_1 x_1 + b_2 \ln(x_2)$
$\Delta x_2 = x_{2f} - x_{2o}$
$\Delta \hat{y} = (b_0 + b_1 x_1 + b_2 \ln(x_{2f}) - (b_0 + b_1 x_1 + b_2 \ln(x_{2o}))$
$= b_2(\ln(x_{2f}) - \ln(x_{2o}))$
$= b_2(\ln(\frac{x_{2f}}{x_{20}})) \approx b_1(\frac{x_{2f}}{x_{2o}} - 1)$ if x_2 increases by 10%, $\frac{x_{2f}}{x_{2o]}} = 1.1 \to$ the response increases by $0.1 b_2$.$\blacksquare$

(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 works when x_{2o} is close to x_2 (i.e. in a neighborhood).

(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.

If x_1 changes by, say, t, then the response chances by e^{tb_1}. If x_2 changes by s, then the response changes by $x_2 e^{b_2}$.

(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.

$\ln(\hat{\mathbf{y}}) = b_0 + b_1 x_1 + b_2 \ln(x_2)$
$\hat{\mathbf{y}} = e^{b_0 + b_1 x_1 + b_2 \ln(x_2)}$
$= e^{b_0} e^{b_1 x_1} e^{b_2 \ln(x_2)}$
$= e^{b_0} e^{b_1 x_1} e^{\ln(x_2)^{b_2}}$
$= e^{b_0} e^{b_1 x_1} x_2^{b_2}$, where we let $m_i = e^{b_i}$ to get
$= m_0 m_1^{x_1} x_2^{b_2}$
To interpret m_1, we see what happens when x_1 increases by one:
$= m_0 m_1^{x_1 + 1} x_2^{b_2}$
$= m_1 (m_0 m_1^{x_1} x_2^{b_2})$,
which shows us that proportional change in x_1 yields proportional change in the response.

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 defined to be the use of $g(\vec{x_*})$ to predict when $\vec{x_*} \notin$ Range [X]. It is a net negative as the model is not optimally built to predict outside the range.

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

Yes, especially when the models are based on polynomials where small changes in inputs can cause very variable effects at the border (Runge's phenomenon). On the other hand, log-based models are a little more stable when the extrapolation is to the right of the range (note, there is no negative range for logs).

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

The edges of the range of polynomial regression models suffer from Runge's phenomenon, creating unstable curves with high variance. Once we go beyond the range, the unstability can be even worse.

Problem 4

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

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

We have a list of models, which could include linear models, polynomials, log-models, first-order interactions (among others). We also have numerous algorithms. Say this gives us a choice of n models, such as:
$g_1(\vec{x}), g_2(\vec{x}), \cdots g_{(}\vec{x})$
How do we select one? This is not a question about select a correct one, as all models are approximate. The metric we will minimize, in order to pick a model, is oos error.

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

Using two splits of data, into training and test, we would learn from training and predict on test set using all models. We would get in-sample metrics.

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

The metrics we get would be dishonest.

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

Note: error metrics evaluated on $\mathbb{D}_{train} \cup \mathbb{D}_{select}$ would be referred to as in-sample, as opposed to out-of-sample which includes $\mathbb{D}_{train}$.

We would:

- fit all models on the training set to obtain $\hat{y}$'s
- predict all models on the select set to obtain in-sample errors
- we will select the model with the least in-sample error
- predict on the test set to get a conservative estimate of the out-of-sample error.
- build the final model on all of the data using the selected model.

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

It gives us oos error metrics, which are honest. The choice is stable with lower variation.

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

We apply k-innerfolds on select sets to obtain the model with the lower in-sample errors. Then, the k-outerfolds are applied on the test sets to obtain out-of-sample metrics. The model with the best out-of-sample performance is then used to build g_{final}.

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

Assuming we have one algorithm with a hyperparameter. We need to have a range of the hyperparameter. Then, we can divide the range into a grid (the grain of which depends on our computational power). Each value of the grid would be counted as a model and we can apply nested resampling on three splits of the data to find hyperparameter values.

(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 $p \gg n$.

The expansive set will be a power set of:
$poly(p_{raw}, p_{raw}), \sum_{i=1}^{p_{raw}} \ln p_{raw}, \sum_{tr \in sin,cos,tan,csc,sec,cot}^{p_{raw}} tr(p_{raw})$
As this is a power set, it will be a huge number, and it would be difficult to be a dataset in which the n is not «p.

(i) [easy] Describe the methodology from class that can create a linear model on a subset of the transformed features (from the previous problem) that will not overfit.

Lasso! It provides a subset of the features, names $\vec{b}_{lasso}$, that are relevant and do not overfit. Lasso doubles up as a feature/variable selector.

Problem 5

These are some questions related to the CART algorithms.

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

(a) Consider all possible orthogonal-to-axis splits. For each split, there are two daughter nodes. Assign $\hat{y} = \bar{y}$ of the response in the nodes. Calculate SSE in each node: $\sum_{i \in node}(y_i - \bar{y}_{node})^2$.

(b) Locate the best split by minimizing the following objective function:
$SSE_{weight} = \frac{n_L SSE_L + n_R SSE_R}{n_L + n_R}$
& create the split.

(c) Repeat steps 1-2 recursively for each algorithm node until daughter node has $\leq N_o$, the node size (hyperparameter), i.e. the number of observations in 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 one dimension, where where a is the start of the range, b is the end of range, and c is the size of the bin:
$\mathcal{H} = \{w_1 \mathbb{1}_{x \in (a,a+c)} + w_2 \mathbb{1}_{x \in (a+c,a+2c)} + \cdots + w_d \mathbb{1}_{x \in (b-c,b)}\}$

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

We could look at the second moment/sample variance, mode, or median.

(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.

(a) identify an inner node whose daughter nodes are both leaves. This will go to the bottom-most level, and work from right to left, and go upwards after finishing a level.

(b) delete both daughter nodes

(c) the $\hat{y}$ of this new leaf becomes the average of responses in the observations that were in the deleted daughter nodes.

(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.

A linear relationship would have the vanilla OLS beating regression trees in oos predictive accuracy.

(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).

The same as Regree Tree, except instead of SSE_{weight} we use:
$G_{weighted} = \frac{n_L G_L + n + RG + R}{n_L n_R}$,
and instead of $\bar{y}$ we use mode of y.

(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.

We can look at the second moment/sample variance, or median.