MATH 342W / 642 / RM 742 Spring 2025 HW #4

Sergio E. Garcia Tapia

Thursday 10th April, 2025

Problem 1

These are questions about 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. We also won't cover chapters 12-13 and the conclusion on the homework.

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

Flu fatalities are estimated by considering R_0, the basic reproduction number. However, to estimate reliably, the disease must sweep through a community. The fatality rate also cannot be measured accurately early on. The result is a need to extrapolate from few data points, which often leads to poor predictions. In this setting, the most dominant type of error is ignorance error δ, since n is small.

(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 "making the assumption that the current trend will continue indefinitely into the future". This sounds like extrapolation is when we believe a model is stationary, meaning that the relationship between the features and the target do not change over time. In class, we defined extrapolation as predicting outside the range of the design matrix X, which was the defined to be a rectangle delimited by the maximum and minimum feature values in each dimension. The time component or idea of stationarity did not play a role in this definition.

(c) [easy] Give a couple examples of extraordinary prediction failures (by very famous people who were considered heavy-hitting experts of their time) that were due to reckless extrapolations.

- In 1682, Sir William Petty predicted a population of 700 million in 2012, but it was instead 7 billion around 2011.
- In 1968, Paul Ehrlich and Anne Ehrlich incorrectly predicted that hundreds of millions of people would die from starvation.

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

In the case of a self-fulfilling prophecy, it is as if a model initially makes a particular prediction $\hat{y}$. Then the $\hat{y}$ becomes a feature or strongly influences a feature, leading to other predictions being close to $\hat{y}$.

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

Underfit. One reason is that the SIR model assumes that the interactions between the population is random, which often does not hold. It also assumes that different subjects are equally likely to be susceptible. However, certain groups may be more susceptible than others, perhaps due to religion, or occupation, among other things. The model assumes that subjects are equally likely to be vaccinated, but that may not hold due to differing beliefs about the risk of contracting the disease, or due to other inherent beliefs held by a population. In short, the SIR model expects certain "homogeneous" conditions to hold, which in general do not. Because it fails to take into account many asymmetries and other complicated interactions, it underfits.

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

He meant that in order for a model to be completely accurate, it must be know everything about a phenomenon. Conversely, no model can be completely accurate because it lacks *something*, and that alone can lead to different predictions. Perhaps the model for a cat won't be a cat itself, even the creature it describes closely resembles a 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?

Feedback mechanisms refers to validation about the predictions that we make. In this class, we split our data set $\mathbb{D}$ into $\mathbb{D}_{\text{test}}$ and $\mathbb{D}_{\text{train}}$. By validating the prediction function $g = \mathcal{A}(\mathbb{D}_{\text{test}}, \mathcal{H})$ from our model $(\mathcal{A}, \mathcal{H})$, we learn whether our predictions are effective. Then we use this information to inform our decision about how to tune our model. This is the "learning from data" approach at play.

(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 obtains many different data points (large n) and many features p. He analyzes the trends in the data carefully, and is careful not to overfit when building a model that he uses to make predictions and place bets.

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

A lot of science is not reproducible because it is based on predictions made by fitting noise. Given we are in the era of Big Data, there is more noise and the same amount of objective truth. Thus scientists are likelier to overfit the data that they gather. Silver suggests that it's a consequence of applying the frequentist approach to probability, which encourages reducing error by collecting more data. However it does not encourage telling the noise from the data.

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

Fisher was biased, and his statistical philosophy conflicted with the practice that was used to arrive at the hypothesis about smoking and cancer. He placed more emphasis on the methods than the interpretation of the results was too victim of getting caught in the noise.

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

Bayesianism, with Silver claiming that some researchers have begun arguing against it in undergraduate study.

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

Kasparov made a move that does not occur often in a master competition. Therefore, the number of data points n and the number of features p were both low in regards to the move (how players have responded to it in the class, how effective it was, the win rate of the player who responded with that move, etc). If the computer were to build a predictive model to know how to counter that move, the model would suffer from underfitting because of the small n and p.

(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?

When Google returns search results, they are measuring “usefulness" or “relevance" of the results. Google measures this in order to improve their search algorithms, thereby refining getting closer to providing you with the information that you are looking for.

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

Google's theories are akin to creating a model, which consists of getting some data $\mathbb{D}$, an algorithm $\mathcal{A}$, and a set of candidate functions $\mathcal{H}$. Testing those theories corresponds to out-of-sample validation with $g = \mathcal{A}(\mathbb{D}, \mathcal{H})$ on $\mathbb{D}_{\text{test}}$.

(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 student that succeeds in this class is well-equipped to assess their predictions and models. We are aware of the dangers of overfitting or even underfitting, the importance of using honest metrics and viewing results statistically without letting the model be the final say.

(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	Uno	Bingo
High skill	Billiards	

(r) [easy] [EC] Why do you think Billing's algorithms (and other algorithms like his) are not very good at no-limit hold em? I can think of a couple reasons why this would be.

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

Yes. If we were to think of "hard work", "natural talent", "opportunities", "environment" as predictors of "success", then yes, I think this is a good model for success. It stands to reason that if you work harder, your chance of being successful increases. Similarly, a person who has more life opportunities takes more "shots", and after enough of them, some of them ought to land in the basket.

(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. Part of what makes a "good" model is the process that led to its creation. Silver mentions that we can never be certain about the reason for an inaccurate prediction. However, being disciplined in our model construction and prediction practices, where we gather a lot of data, from different places, and test our predictions honestly, our models can become better. We mut remember that models do not have the final say, and they need to be imbued with meaning and refined each time.

(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.

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

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

(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?

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

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

Problem 2

These are some questions related to probability estimation modeling. Let X denote the design matrix with n rows and $p+1$ columns (with the first column being $\mathbf{1}_n$ and the other columns being linearly independent predictors) and $\boldsymbol{y}$ is the binary response vector of size n and use this notation throughout your responses.

(a) [easy] What is g_0 if you are modeling probability estimates?

(b) [easy] What is $\mathcal{H}_{pr}$ for the probability estimation algorithm that employs the linear model in the covariates with logistic link function?

(c) [easy] What is $\mathcal{H}_{pr}$ for the probability estimation algorithm that employs the linear model in the covariates with cloglog link function?

(d) [easy] Let $\mathcal{A} : \boldsymbol{b} = \arg\max_{\boldsymbol{w} \in \mathbb{R}^{p+1}} \{\ldots\}$. Derive the expression that replaces the $\ldots$ which will be a function of $X, \boldsymbol{y}, \boldsymbol{w}, n$. Note: this algorithm fits a "logistic regression".

(e) [easy] Why is logistic regression an example of a "generalized linear model" (glm)?

(f) [easy] Consider $\boldsymbol{x}_*$ to be a new unit. For its prediction, the probability estimate that $y_* = 1$ is 37%, what is the log odds of $y_* = 1$?

(g) [easy] If, $\boldsymbol{x}_* \boldsymbol{b} = 3.1415$ where $\boldsymbol{b}$ is the result of the logistic regression fit, what is the probability estimate that $y_* = 1$?

(h) [harder] If, $\boldsymbol{x}_* \boldsymbol{b} = 3.1415$ where $\boldsymbol{b}$ is the result of the probit regression fit, what is the probability estimate that $y_* = 1$?

(i) [easy] In probability estimation modeling, what is the formula for the Brier Score performance metric? Prove the Brier score is always non-positive.

(j) [easy] In probability estimation modeling, what is the formula for the Log Scoring Rule performance metric? Prove the Log Scoring Rule is always non-positive.

(k) [difficult] Generalize linear probability estimation to the case where $\mathcal{Y} = \{C_1, C_2, C_3\}$, i.e. a nominal variable with $L = 3$ levels.

Assume the logistic link function. Write down the objective function that is argmax'd over the parameters (you define what these parameters are — that is part of the question). Once you get the answer you can see how this easily goes to $L > 3$, an arbirtrary[1] number of response levels.

[1] Note: The algorithm for general L is known as all of the following: "multinomial logistic regression", "polytomous LR", "multiclass LR", "softmax regression", "multinomial logit" (mlogit), the "maximum entropy" (MaxEnt) classifier, and the "conditional maximum entropy model". You can inflate your resume with lots of redundant jazzy terms by doing this one question!

For the next two questions, let $n_1 := \sum \mathbb{1}_{y_i=1}$ and $n_0 := \sum \mathbb{1}_{y_i=0}$ so that $n = n_0 + n_1$. Then assume $n_1 \neq n_0$. This is equivalent to letting $n_1 = cn$ and $n_0 = (1-c)n$ and assuming $c \in [0,1] / \left\{\frac{1}{2}\right\}$.

(l) [harder] [MA] Prove the Brier score is always higher for g_0 vs the model where you set $\hat{p}_i = \frac{1}{2}$ for all i. Hint: $(1-c)c < \frac{1}{2}$ for $c \in [0,1] / \left\{\frac{1}{2}\right\}$.

(m) [difficult] [MA] Prove the Log Scoring Rule is always higher for g_0 vs the model where you set $\hat{p}_i = \frac{1}{2}$ for all i.

Problem 3

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?

(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.

(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?

(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.

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

(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.

(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 4

These are some questions related to extrapolation.

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

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

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

Problem 5

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

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

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

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

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

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

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

(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 $p \gg n$.

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

Problem 6

These are some questions related to the CART algorithms.

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

(b) [difficult] Describe $\mathcal{H}$ for regression trees. This is very difficult but doable. If you can't get it in mathematical form, describe it as best as you can in English.

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

(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.

(e) [difficult] Provide an example of an $f(\boldsymbol{x})$ relationship with medium noise δ where vanilla OLS would beat regression trees in oos predictive accuracy. Hint: this is a trick question.

(f) [easy] Write down the step-by-step $\mathcal{A}$ for classification trees. This should be short because you can reference the steps you wrote for the regression trees in (a).

(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.