# MATH 342W / 650.4 Spring 2022 Homework #5

Peter Antonaros

Sunday 15th May, 2022

## 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.)

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. It is obviously important in Data Science (that's why Math 341 is a required course in the data science and statistics major).

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

Flu fatalities are hard to predict because of error due to ignorance. It is true that we have a lot of data surrounding these sorts of things, but is the data of high quality and reporting enough metrics that contribute to a fatality? At this point, error due to estimation is very low considering there are many great algorithms out there and error due to misspecification can be minimized. This leads us to ignorance error which is substantial in this prediction task.

Does our data indicate every detail about a person who died...weight, height, age, sex, place of residence, hospital they died at, blood pressure, etc. There are simply too many possible factors each playing an important role in the final outcome.

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

'It involves the assumption that the current trend will continue indefinitely.' His definition of extrapolation conflicts slightly with our definition. He is defining it more from the perspective of time, while we are generalizing to simply the range in which the features are defined on. I would not say that these two definitions conflict, rather our definition is a further generalization of his definition.

(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.

One interesting example, not from any one person, was the idea of flying cars that became popular in the mid 1900s. People saw how in a matter of a couple decades we went from horse drawn wagons to cars being built on assembly lines and assumed this sort of progress would continue. This assumption was based on the fact that their observations were within a certain time interval. They then extended this notion to 'by the year 2000, there will be flying cars'. Essentially these ideas came from extrapolation of past and current trends, of which we know now in 2022 flying cars have not become mainstream.

Another interesting example, following suit with example 1 is the idea of we should have been on Mars by now. People felt that after the space race, that this sort of technological progress would continue at the rapid pace it did and within another decade or two there would humans on mars. As of 2022 this has not yet occurred and its due to people extrapolating, not taking into account other factors that would alter this technological course.

A noteable example of someone famous extrapolating is Paul R. Ehlrich who predicted that millions would die from starvation in the 1970s. This did not occur because the fertility rates which were at record highs when he made this prediction did not continue to stay high. We can say that his model was built on fertility rates $x_1 \in [A, B]$, but by the 1970s, $x_1 \in [A_1, B_1]$ where $A_1 < A, B_1 < B$, leading to inaccurate predictions.

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

Self-fulfilling prophecy: A prediction leads to the very thing it was attempting to predict for. Our model $g$ is now being fed information that was the result of $\hat{y}$ which then further reinforces the prediction it is making.

Self-canceling prediction: A predictions leads to the very thing it was attempting to predict against. Our model $g$ is now being fed information that was the result of $\hat{y}$, but unlike in Self-Fulfilling, since it is trying to predict against this response it then transitions back to the original state.

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

The SIR model of infectious disease is under fit. There is simply too few $p$ account for in the SIR model to be considered optimal or overfit. This isn't to say it provides no value, but knowing the model is under fit can help in future decision making. Practically any model for this sort of scenario would be over fit unless of course we are considering multi-million/billion $p$ models as created by Google as an example.

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

Norbert Weiner said this because no matter how precise mathematics is, biology is a messy field. It is full of complexities both known and hidden, and as of current we have no way to model every last protein, response, pathway etc. This means that no matter how good the mathematical model is, it will never replicate the very thing we are modeling.

In this way, the best model to a cat is simply another cat, which although different, still contains all these complexities. Norbert Weiner is essentially rephrasing the popular quote from George Box. Rather than saying all models are wrong, he is saying that the best model of something is 'the something' itself.

(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 this class, feedback mechanisms would be when the $\hat{y}$ of our model $g$ affect future features $x_{i's}$, which in turn alter the future $\hat{y}_{i's}$. This can either be a positive or negative feedback loop depending on what the model is predicting for / against.

(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 had an edge because he detected that the underlying point totals (team A points + team B points) was frequently going over the predicted totals. Voulgaris obtained an edge in the $\delta$ (error due to ignorance) because he realized that the point guard for team A would be a free agent soon and was doing everything he could to score points. Although this seems obvious he was obtaining a clear view of the entire picture which allowed him to make slightly better predictions, just the edge needed in betting

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

Lately there has been a lot of talk about the reproducibility crisis that science is facing. There can be a variety of reason that this may occur...

Their data set has too few observations $n$. This can lead to results in a publication that will be hard to reproduce despite following same methods the original researchers used and so on. Albeit this is probably not a main reason for this criss, but it does occur quite a bit especially in the social science fields. This is because the data they aim to collect is often harder to obtain as opposed to getting data from say massive economic reports, or banks etc.

Imprecise language specifically when referencing error and statistically tests. There are researchers who say 'our p-test value is $> 0.05$' to indicate that their dealing with something that is statistically significant. The fact that we are saying greater than is imprecise though and can lead to others not being able to reproduce results. Is your p value greater in the sense that is is equal to 0.1 or 0.05000001, there is a big difference.

Another big reason I believe that a lot of science is not reproducible, is because people rarely publish bad results. This means that hundreds of researchers could have investigated the same thing, all yielding the same inconclusive or 'bad' result, and we would never know. If one person does this same research and gets different perhaps more positive results then it will be published as 'original research'. This means we have a sort of bias now, which can lead to people not being able to reproduce the results. We don't know every last thing the one group did differently in comparison to all the previous and so this is a huge problem.

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

I personally believe that Fisher had two primary reasons for believing that smoking does not cause lung cancer. For starters he worked closely with the Tobacco industry which we would consider to be a conflict of monetary interest. The second was that he was a smoker himself, and this in my opinion probably had a bigger effect on his belief. He indicates that there was a correlation, but causation could not be proved, and I think his own smoking lead to him making this decision. Despite his genius, it is difficult for humans to regard something they enjoy as harmful to themselves, and will often times work to dismiss it as he did. There are professors now who make it a point to say that you should not research something that is near and dear to you as it will influence how you see the data. Then again I also would not regard this as a flaw of Fisher because science moves forward based on disagreement. He disagreed with the premise which lead to further study and eventually he was proved wrong. If every scientist agreed then we would have an echo chamber of research rather than meaningful discussion.

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

I would argue that the world is moving more towards Bayesianism since much of what we say now is with consideration to other factors both observed and unobserved. We can see that this is true given the rise of Machine Learning and Deep Learning which has basis in Bayesianism.

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

Simply put Kasparov was able to defeat Deep Blue by switching strategies mid game. This was of benefit to him since once Deep Blue began to notice a pattern in the opponents moves it would play specifically to beat those predicted moves. Essentially we can see this as over fitting to prior moves in order to predict future moves made by Kasparov. Eventually this constant switching of strategies throughout the game and over multiple games lead to a decrease in Deep Blue's accuracy resulting in its loss.

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

Fischer did not adhere to the typical chess heuristics such as 'never give up your queen unless for another queen...'. Although these work 99% of the time, it fails in the 1% of times where that is the best move. Fischer was able to identify that this was the best possible move at the time and defeated his opponent. By allowing himself to be unrestricted by these notions it resulted in a overall better game. The computer an unemotional winning machine also evaluated these moves as the best possible moves, because it had no notion of these human chess heuristics.

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

Google returns pages in an order of most useful to least useful to you according to the search you entered. Essentially they are trying to predict use-fullness/utility given your search. They chose this metric to reduce the amount of useless information. If you find what you're looking for with Google, chances are you will use their browser again for the next search.

For example if I were to search 'Chinese Food', Google has many choices of what to return to me. They can return me top Chinese food stores in China, United States, or my own neighborhood. In order to maximize my search they will order these according to other known information about me. My location would dictate that these food stores in Astoria New York would be on top, thus increasing my satisfaction with the results from the search.

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

Google's theories can be considered to be phenomena that they are interested in exploring and predicting for.

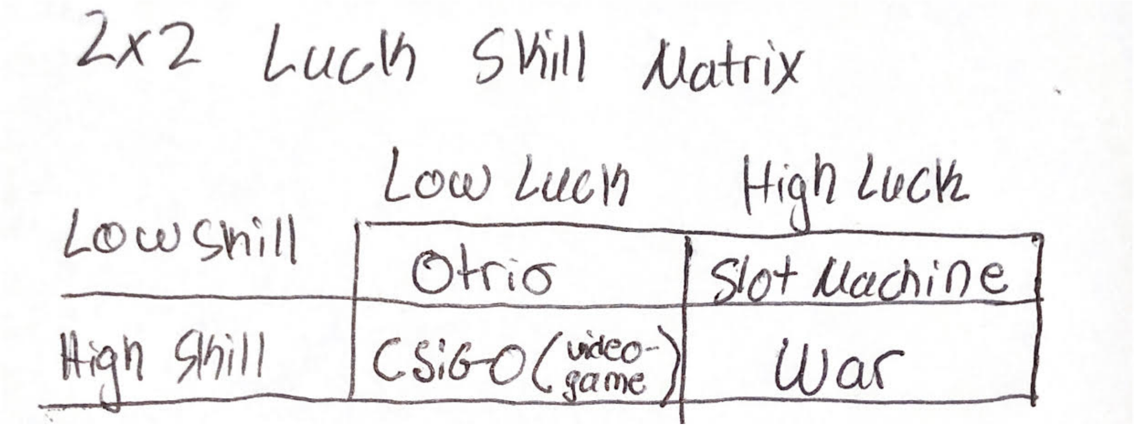
Testing of those theories would be validation of their particular predictions for their phenomena (theory).

(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 ability to go beyond push button tools. This all encompassing statements means that people who have actual in depth knowledge of the field can simply do more than those without this knowledge. A big example, is when something goes wrong and your predictions are not as good as you expected them to be. For someone who has taken this class as an example, we can see which algorithms are more effective for different sorts of data. Rather than simply trying all permutations of algorithms, train sets, hyper parameters etc we can narrow this space significantly. Although in the end you still may be using a push button tool, its knowing what button to press that gives you an advantage to others.

I'd also argue that simply understanding even some of the mathematics that actually goes into these tools makes you more applicable to different fields. You'll better understand what error to use for a particular field, what the prediction actually means and so on. Eventually people who know the inner workings of these process get a certain 'feel' for what they need to do rather than pushing enough buttons to arrive at a respectable model.

(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

| | Low Luck | High Luck |
|---|---|---|
| Low Skill | Otrio | Slot Machine |
| High Skill | CSGO (video-game) | War |

(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.

I think Billing's algorithms (and similar) are not very good at no-limit hold em for the following reasons among others.

1) Too little information to work with. You don't see anybody's cards so you don't have the opportunity to narrow down the possible combination space.

2) A lot of the game is based on intuitive feel based on what the person looks like, their gender, their age, level of sobriety and so on. The algorithm doesn't have access to this information.

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

Silver describes a successful person as getting to the point of being considered 'successful' through a combination of hard work, natural talent, and luck. I would completely agree with this as it is both affirming a person's work while also maintaining that if someone else worked just as hard the outcome might not be the same. He is relatively neutral on this which is fair, since each unique person will have a varied amount of hard work, natural talent, and luck molded together to form their own success. I would agree with this sentiment, and to reiterate, it accounts for the variation in outcome we see despite small changes in input.

(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

I don't know if we should remove Humans entirely from this once a good model has been built, but for the most part yes. Humans often times will insert emotion into decision making which will most likely hurt what the model is 'predicting'. If you've properly created this model and have strong reason to believe it is accurate then you should trust its predictions. Any doubt comes from emotion. For example if our amazing model predicts a bubble and you deny it, you may be thinking with greed and profits in mind rather than I may lose everything in the coming months.

(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?

According to Fama the explanation for a mutual fund performing spectacularly one year, but fails to perform again in the next, is the following...

Consider a model that has been trained on 50 years of data. Its validation occurs on the 51st year and we find that it performs amazing. Fama states that this 51st year would not be important, but that consistency reigns supreme. If this model (mutual fund) can perform well for another 10-15 years then perhaps it is accurate to say that it is beating the market. Fama states that for any short period of time there will be people who perform much better than others.

In the context of models, training and validation splits we see that considering 100 years of data perhaps our validation split should be much larger. Rather than $K = 5$ or $K = 10$ for example we should try something like $K = 2$, to see if our model is consistent over a long market period.

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

I suppose we can consider the maniac momentum model validated since it was tested from the years 1976-1986 and built on data from the 60s into the 70s. Unfortunately despite a positive validation (ignoring transaction costs), its performance would have been terrible in the 2000s (ignoring transaction costs again).

The main reason for this is most models are built on the assumption of stationary conditions. This model was built in a time that was radically different from the 2000s, and so factors like this will lead to massively different market behavior. The mapping between inputs and outputs has shifted which means our model's validation is essentially useless, since those conditions simply don't exist anymore.

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

Fama states that 'if you can't notice you're in a bubble then, it's not a bubble', which implies that bubbles are noticeable when we are in them. In order for a bubble to violate the efficient market hypothesis, it needs to be predictable in real time. (Taken directly from Silver's book). In general, yes it is obvious when a bubble is occurring, but people rarely accept it. They are too deeply entrenched in the momentum of the market and believe this will continue indefinitely. Eventually the bubble bursts and then everyone recognizes it was a bubble with the benefit of hindsight. Generally a rule of thumb to identify a bubble is to compare current growth with long term average growth. If the current rate far exceeds the average then perhaps we are in a bubble.

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

The implication of this model for long term investors is that they should stick to stock with P/E ratios that are lower to ensure higher rates of returns over the long term. In the short term P/E ratio to return is much noisier, but over the long term the model clearly indicates that high P/E rations lead to lower returns or no returns at all.

(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?

Silver quotes the heuristic 'follow the crowd, especially when you don't know any better'. It works well because generally if we as an individual aren't knowledgeable about something we assume that groups of humans as a collective are. Generally it works well for this very reason, the collective knowledge base and decision making on average perform better than the individual (considering all individuals, not just the 'high performing' ones).

(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 good bubble predicting model, the thing that would have greatest influence over you not executing on it would be the movement of the market itself. You may begin to doubt yourself and the model when you continue to see things rise, and not execute based on the models predictions. Especially now in the information age, when our independence is reduced its easy to become misguided by collective statements on the internet such as 'Dogecoin to the moon'. Millions are also saying the same, so you assume there is no way this can go bust despite your model indicating so, and then suddenly it collapses 50%.

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

Heuristics can get you into trouble when they become self reinforcing. In the example in the previous question of following the herd, if the majority of the group don't know what they're doing then mistakes can pile up, and spiral out of control. We follow the mistakes since we are following the group and the process reinforces because then we believe we have to follow more closely to the group in order to prevent these mistakes and so on...

## Problem 2

These are some questions related to probability estimation modeling and asymmetric cost modeling.

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

Logistic Regression is an example of generalized linear model because it uses a link function to transform $b_0 + b_1x_1 + \ldots + b_px_p$. The link function servers the purpose to transform outputs $\in \mathbb{R}$ to $\in [0, 1]$.

We can say that if our algorithm uses the underlying trans-formative function $\Phi(u) = \frac{e^u}{1+e^u}$, then we are logistically regressing against our linear estimators.

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

$\mathcal{H}_{pr} = \{\Phi(\vec{w} \cdot \vec{x})\}$, such that the link function $\Phi(u) = \frac{e^u}{1+e^u}$

(c) [easy] If logistic regression predicts 3.1415 for a new $\boldsymbol{x}_*$, what is the probability estimate that $y = 1$ for this $\boldsymbol{x}_*$?

If $x_* = 3.1415$ then we can say the probability estimate that $y = 1$ for this is given by...

$\frac{1}{1+e^{-x_*}} \approx 95.8\%$

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

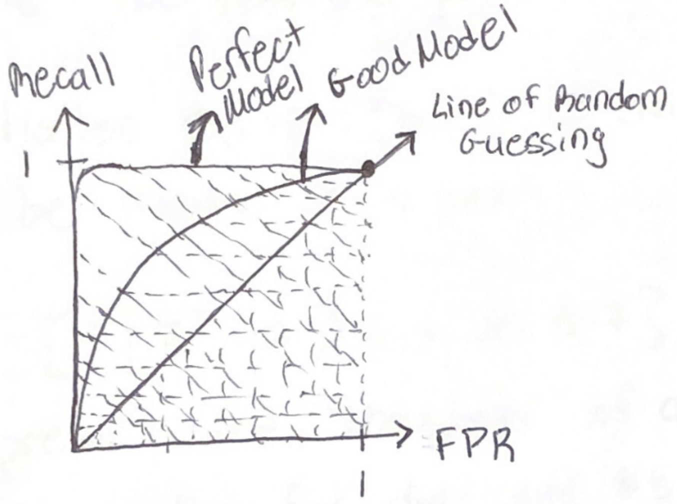
cloglog = Complementary Log Log (Link) Function

$\mathcal{H}_{pr} = \{\Phi(\vec{w} \cdot \vec{x})\}$, such that the link function $\Phi(u) = 1 - e^{-e^{-u}}$

(e) [difficult] Generalize linear probability estimation to the case where $\mathcal{Y} = \{C_1, C_2, C_3\}$. Use the logistic link function like in logistic regression. Write down the objective function that you would numerically maximize. This objective function is one 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 $K > 3$ response categories. The algorithm for general $K$ is known as "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 jazz by doing this one question!

(f) [easy] Graph a canonical ROC and label the axes. In your drawing estimate AUC. Explain very clearly what is measured by the $x$ axis and the $y$ axis.



AUC Estimates

Random Guessing ≈ .5

Good Model ≈ .75

Perfect Model ≈ 1.0

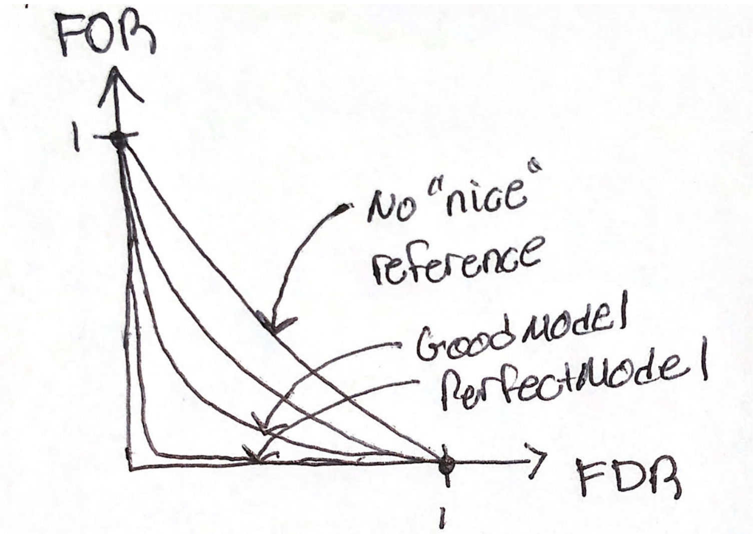
Along the X-Axis we have FPR (False Positive Rate). False Positive Rate is defined as $\frac{FP}{N}$. Here we see that $FP$ is the total occurrences of $\hat{y} = 1$ when the real value $y = 0$ and $N$ is $TN + FP$.

Along the Y-Axis we have Recall. Recall is defined as $\frac{TP}{P}$. Here we see that $TP$ is the total occurrences of $\hat{y} = 1$ when the real value $y = 1$ and $P$ is $TP + FN$.

(g) [easy] Pick one point on your ROC curve from the previous question. Explain a situation why you would employ this model.

Scenario: Cancer Diagnosis leading to treatment Assuming we are using the "good model" curve, we pick a point lower on the FPR and high on Recall. The reason for this choice, is that we don't want people undergoing chemotherapy when they really didn't have cancer as this has a strong chance of ruining their health.

(h) [easy] Graph a canonical DET curve and label the axes. Explain very clearly what is measured by the $x$ axis and the $y$ axis.



Along the X-Axis we have FDR (False Discovery Rate). False Discovery Rate is defined as $\frac{FP}{PP}$. Here we see that $FP$ is the total occurrences of $\hat{y} = 1$ when the real value $y = 0$ and $PP$ is $TP + FP$.

Along the Y-Axis we have FOR (False Omission Rate). False Discovery Rate is defined as $\frac{FN}{PN}$. Here we see that $FN$ is the total occurrences of $\hat{y} = 0$ when the real value $y = 1$ and $PN$ is $TN + FN$.

(i) [easy] Pick one point on your DET curve from the previous question. Explain a situation why you would employ this model.

Scenario: Covid Testing Assuming we are using the "good model" curve, we pick a far along the FDR curve and lower on the FOR curve. The reason for this choice is that we would rather have someone simply stay home and rest even if they don't have covid rather than saying someone doesn't have covid when they do and have them spread it.

(j) [difficult] The line of random guessing on the ROC curve is the diagonal line with slope one extending from the origin. What is the corresponding line of random guessing in the DET curve? This is not easy...

This is a bit tricky for the following reason. On our ROC curve, the diagonal line has a slope of 1 because our false positive rate and true positive rate are inline with each other in the confusion matrix.

On the DET curve, we examine false positive rate and false negative rate which are not inline with each other. My best ‘guess’ as to how the DET curve of random guessing would look like is the following...

## Problem 3

These are some questions related to bias-variance decomposition. Assume the two assumptions from the notes about the random variable model that produces the $\delta$ values, the error due to ignorance.

(a) [easy] Write down (do not derive) the decomposition of MSE for a given $\boldsymbol{x}_*$ where $\mathbb{D}$ is assumed fixed but the response associated with $\boldsymbol{x}_*$ is assumed random.

$$Bias[\vec{x}_*]^2 + \sigma^2$$

(b) [easy] Write down (do not derive) the decomposition of MSE for a given $\boldsymbol{x}_*$ where the responses in $\mathbb{D}$ is random but the $\boldsymbol{X}$ matrix is assumed fixed and the response associated with $\boldsymbol{x}_*$ is assumed random like previously.

$$Bias[\vec{x}_*]^2 + Var[G(\vec{x}_*)] + \sigma^2$$

(c) [easy] Write down (do not derive) the decomposition of MSE for general predictions of a phenomenon where all quantities are considered random.

$$E_{X's}[Bias[\vec{x}_*]^2] + E_{X's}[Var[G(\vec{x}_*)]] + \sigma^2$$

(d) [difficult] Why is it in (a) there is only a “bias” but no “variance” term? Why did the additional source of randomness in (b) spawn the variance term, a new source of error?

Variance stems from how dependent the learned model is on a particular data set. In (a) there is no variance term since $\mathbb{D}$ is assumed to be fixed. This means in the world of possible $D_{i's}$ there is only 1, and so our learned model variance is 0. The error can only come from ignorance and misspecification $\equiv \sigma^2$ and $Bias[\vec{x}_*]$.

(e) [harder] A high bias / low variance algorithm is underfit or overfit?

This algorithm is more likely to be underfit.

(f) [harder] A low bias / high variance algorithm is underfit or overfit?

This algorithm is more likely to be overfit.

(g) [harder] Explain why bagging reduces MSE for "free" regardless of the algorithm employed.

Bagging reduces variance. In our decomposition we can see that variance is an error term and so by lowering this we are lowering our overall MSE. This is done for free since the bagging process uses no extra sources of data, perhaps the only part of the bagging algorithm that costs anything is the increase in computation time. This is minor and well worth the decrease in variance -> decrease in MSE of our model.

(h) [harder] Explain why RF reduces MSE atop bagging $M$ trees and specifically mention the target that it attacks in the MSE decomposition formula and why it's able to reduce that target.

Random Forest is able to reduce MSE atop bagging $M$ trees because for each random forest that is constructed a random sample $p_1$ of the $p$ features are used, such that $p_1 < p$. This means that the likelihood of over fitting has decreased which implies that the overall variance of the model is lowered. Since variance is a component of MSE, we can see how RF targets this in our MSE, thus topping $M$ trees in most cases.

(i) [difficult] When can RF lose to bagging $M$ trees? Hint: setting this critical hyperparameter too low will do the trick.

If we set the random sample of features that RF selects from to be too low, then our model will be underfit. The random forest algorithm will not learn enough about the data which means we are essentially combining a bunch of poorly performing trees to obtain our random forest. This has a strong potentional of losing to bagged $M$ trees as these would have learned with all of our features.

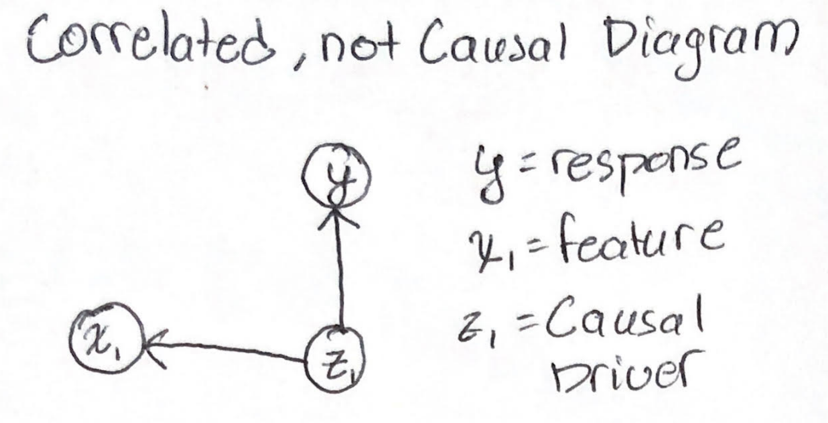
## Problem 4

These are some questions related to correlation-causation and interpretation of OLS coefficients.

(a) [easy] Consider a fitted OLS model for y with features $x_1$, $x_2$, ..., $x_p$. Provide the most correct interpretation of the quantity $b_1$ you can.

When comparing two mutually observed observations (A) and (B) which are sampled in the same way, where (A) has a value one unit higher than that of (B) in $x_i$, but otherwise share the same values of other $x_{i's}$, then (A) on average is predicted to have a response that differs by $b_1$ units in y from the predicted response of (B) under the assumption that the linear model is true.

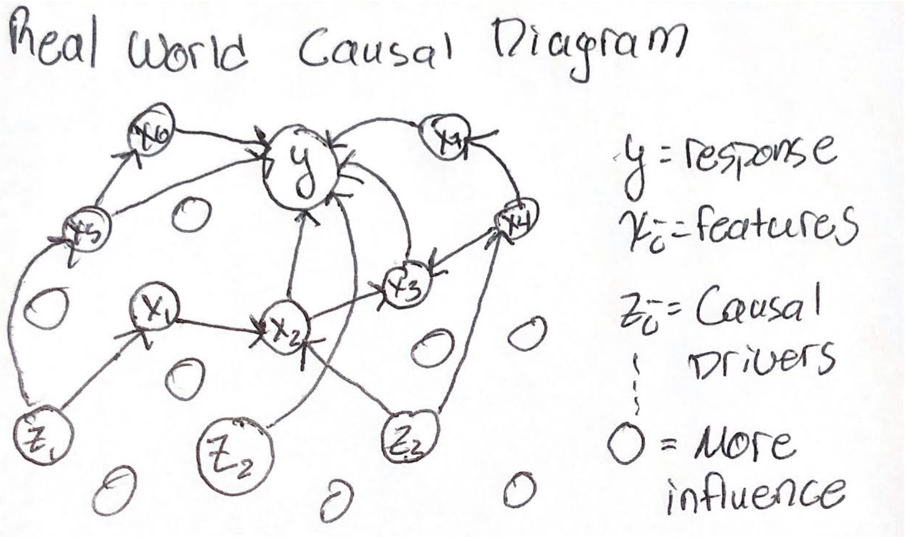
(b) [easy] If $x$ and $y$ are correlated but their relationship isn't causal, draw a diagram below that includes $z$.



(c) [easy] To show that $x$ is causal for $y$, what specifically has to be demonstrated? Answer with a couple of sentences.

In order to show that $x$ is causal for $y$ we must demonstrate that $x$ came before $y$, that their relationship is not by chance, and that nothing else has influence over their relationship.

(d) [harder] If we fit a model for y using $x_1$, $x_2$, $\ldots$, $x_7$, provide an example real-world illustration of the causal diagram for $y$ including the $z_1$, $z_2$, $z_3$.



## Problem 5

These are some questions related to missingness.

(a) [easy] What are the three missing data mechanisms? Provide an example when each occurs (i.e., a real world situation).

The three types of missing data mechanisms are..

MCAR (Missing completely at random): Missing completely at random implies there was total unpredictability in how the data went missing, and something that is totally out of our control. An example of this would be if our computer storage suddenly flipped some bits and certain chunks of information are not either corrupted or totally gone. Another way to phrase this is that the causes of our missing data is unrelated to the data itself.

MAR (Missing at random): Similar to MCAR, in the sense that we may not know why the data is missing, but its causes are related to the overall data. An example would be if in our data set to predict income, we have a missing value for capital gains this may imply that the person has too little money to invest, thus giving us insight to their income despite the field not being present.

NMAR (Not missing at random): NMAR would be data that is missing and unrelated to the observations/factors that the researcher is concerned with. There is not much you can do with this because there are simply unobserved features out of your control affecting the data. An example would be if a person simply refused to take a survey. We cannot get this data, don't know why its missing, just that it is missing and cannot take it into account.

(b) [easy] Why is listwise-deletion a terrible idea to employ in your $\mathbb{D}$ when doing supervised learning?

Listwise-deletion is a terrible idea to employ for supervised learning. This is because supervised learning predicates itself on the basis of having large amounts of data to learn from. Real data is often filled with missing values scattered throughout $\mathbb{D}$. This means if we use listwise deletion we will severely reduce $n$, which can lead to massive problems in our SL algorithms. Not enough data with too many features will lead to over fitting, not enough data with little features will lead to large error in predictions and so on.

In addition, sometimes missing data can actually give us insight into the 'thing' we are predicting. By removing this data with list wise deletion we are immediately rejecting any of the possible benefits the missing data could have provided us.

(c) [easy] Why is it good practice to augment $\mathbb{D}$ to include missingness dummies? In other words, why would this increase oos predictive accuracy?

It is good practice to include missingness dummies because the fact that the data is missing can be related to the response we are predicting for. We can think of this as adding more relevant features to our data set thus lowering our original $\delta$. Although the largest benefit would probably be from having the data to begin with, if this is not possible, then using missing dummies will help to increase our oos predictive accuracy more than excluding the missing data.

We can choose to either have...

$y = f(x_1, x_2, \ldots, x_p) + \delta_1$ or

$y = f(x_1, x_2, \ldots, x_p, m_1, m_2, \ldots, m_p) + \delta_2$, where $m_1 \ldots m_p$ represent missing dummies, and $\delta_2 \leq \delta_1$.

We can see that the missing dummies clearly provide some value to describing something more about our phenomena.

(d) [easy] To impute missing values in $\mathbb{D}$, what is a good default strategy and why?

To impute data a good default strategy is to run Miss Forest treating the current column you want to impute as the response. This will iteratively converge to a values for the missing data, which results in error due to ignorance being lower than our original, which is a positive.

Another strategy would be to treat the missing data as new dummy features. This can sometimes yield more insight to your actual prediction target if there is reason to believe the missing data is linked to the response.

## Problem 6

These are some questions related to lasso, ridge and the elastic net.

(a) [easy] Write down the objective function to be minimized for ridge. Use $\lambda$ as the hyperparameter.

$\boldsymbol{b}_{ridge} := \text{argmin}\{SSE + \lambda||\vec{w}||_2^2\}$

(b) [easy] Write down the objective function to be minimized for lasso. Use $\lambda$ as the hyperparameter.

$\boldsymbol{b}_{lasso} := \text{argmin}\{SSE + \lambda||\vec{w}||_1\}$

(c) [easy] We spoke in class about when ridge and lasso are employed. Based on this discussion, why should we restrict $\lambda > 0$?

Briefly we said that ridge regression should be used for prediction and lasso should be used for feature selection.

(d) [harder] Why is lasso sometimes used a preprocessing step to remove variables that likely are not important in predicting the response?

Lasso is sometimes used as a preprocessing step, since it performs well when there are a small number of significant features and a large number of insignificant ones. We can use this to out advantage by trying to fit the best possible lasso model, which then in turn tells us which features are the most influential on our response by bringing the weights of the 'useless' features to 0. The reason as to why this is done is because a reduction in the number of features will decrease our model complexity, which decreases the variance and brings us closer to the optimal model. This is especially useful when our large number of original features would lead to an over fit model, we can decide which important subset to keep.

(e) [easy] Assume $\boldsymbol{X}$ is orthonormal. One can derive $\boldsymbol{b}_{\text{lasso}}$ in closed form. Copy the answer from the wikipedia page. Compare $\boldsymbol{b}_{\text{lasso}}$ to $\boldsymbol{b}_{\text{OLS}}$.

$\vec{b}_{OLS} = (X^T X)^{-1} X^T \vec{y}$

$\vec{b}_{Lasso} = S_\lambda(\vec{b}_{OLS})$

Here we can see that lasso regression operates on the regular OLS betas, by translating them towards 0, or exactly 0 if they are small enough.

(f) [harder] Write down the objective function to be minimized for the elastic net. Use $\alpha$ and $\lambda$ as the hyperparameters.

$\boldsymbol{b}_{elastic} := \text{argmin}\{SSE + \lambda(\alpha||\vec{w}||_1 + (1-\alpha)||\vec{w}||_2^2\}$

(g) [easy] We spoke in class about the concept of the elastic net. Based on this discussion, why should we restrict $\alpha \in (0, 1)$?

In the elastic model, $\alpha$ acts as a hyper parameter weight as to 'how much ridge' vs 'how much lasso' the model should be using. It makes sense that $\alpha \in [0, 1]$ since it can be seen as either all ridge or all lasso and any value in between. Values outside this range would not make a difference of 'how much' of either sort of model is being used.