MATH 390.4 / 650.2 Spring 2018 Homework #5t

Darshan Patel

Friday 18th May, 2018

Problem 1

These are questions about the Finlay's introduction to his book.

(a) [easy] Finlay introduces predictive analytics by using the case study of what supervised learning problem? Explain.

Finlay introduces predictive analytics by using the case study of credit scoring, which we also looked at in the beginning of the semester. Predictive analytics was used to determine if a customer could take out a loan, credit card, mortgage, etc. This was initially done by human underwriters who just made an expert opinion on each application. Now, with predictive models, we are better able to determine credit scoring using FICO scores to determine whether a customer's application should go through or not.

(b) [difficult] What does a credit score of 700 mean? Use figure 1.2 on page 5 when answering this question.

A credit score of 700 means an individual has a 1024 odds of default. This means that if you have 1025 borrowers that score 700, the expectation is that 1024 will repay what they borrowed and 1 will not.

(c) [difficult] How much more likely is someone to default if that have 9 or more credit cards than someone with 4-8 credit cards?

If someone has 9 or more credit cards than someone with 4 to 8 credit cards, their credit score will fall by 18 points. Assuming all other variables the same, the one with the credit score of 700 will default at the odds of $1024 : 1$ while the one with credit score of $700 - 18 = 682$ will default at the odds of $\approx 800 : 1$. This means that if you have 801 borrowers that score 682, the expectation is that only 800 will repay what they borrowed. The change is

$$\frac{\frac{1}{801} - \frac{1}{1025}}{\frac{1}{1025}} \approx 28\%$$

(d) [easy] Summarize Finlay's conception of "big data".

Big data has four features: volume, variety, volatility and multi-sourced. Big data deals with any database that is too large to be comfortably managed on an average PC/laptop/server. It contains many different types of structured and unstructured data. Big data also deals with data that is not stable, such as heart rate and address. Some big data sources are generated from one source whereas some are from many different sources which introduces additional issues around data quality, privacy and security.

Problem 2

This question is about probability estimation. We limit our discussion to estimating the probability that a single event occurs.

(a) [easy] What is the difference between the regression framework and the probability estimation framework?

The regression framework gives a model a full picture of what an outcome will be. By using the probability estimation framework, it gives the model probabilities of how likely one or more feature will affect a predicted phenomenon.

(b) [easy] Is probability estimation more similar to regression or classification and why?

Probability estimation is more similar to classification because it tries to predict whether something is a 1 or 0, or success or failure.

(c) [difficult] Why was it necessary to think of the response Y as a random variable and why in particular the Bernoulli random variable?

It is necessary to think of the response Y as a Bernoulli random variable because a Bernoulli random variable only has 2 outcomes, 0 and 1. The response Y also has two outcomes only.

(d) [difficult] If we use the Bernoulli r.v. for Y, are there any error terms (i.e. δ, ϵ, e) anymore? Yes/[✓ no].

(e) [easy] What is the difference between f in the regression framework and f_{pr} in the probabilistic classification framework?

The difference between f in the regression framework and f_{pr} in the probabilistic classification framework is that f is a polynomial and $f(\vec{x}) \in \mathbb{R}$ whereas $f_{pr} \in (0, 1)$ since it goes into the Bernoulli distribution as a parameter. f_{pr} is the probability $\mathbb{P}(Y|\vec{x}) = 1$.

(f) [difficult] Is there a t_{pr}? If so, what does it look like?

There is a t_{pr} and it looks like g_{pr} because we want f_{pr} to be close to t_{pr}.

(g) [easy] Write out the likelihood as a function of f_{pr}, the $\boldsymbol{x}_i$'s and the y_i's.

$$\mathbb{P}\left(Y_1, \ldots, Y_n\right) = \prod_{i=1}^{n} f_{pr}(\boldsymbol{x}_i)^{y_i}(1 - f_{pr}(\boldsymbol{x}_i))^{1-y_i}$$

(h) [difficult] What assumption did you have to make and what would happen if you didn't make this assumption?

We assumed that there is no dependence structure between $Y_1, \ldots, Y_n$. If we don't make this assumption, then $\mathbb{P}\left(Y_1, \ldots, Y_n\right)$ is unknown and probably will be messy to get in a canonical form.

(i) [easy] Is f_{pr} knowable? Yes/[✓ no].

Problem 3

This question continues the discussion of probability estimation for one event via the logistic regression approach.

(a) [harder] As before, if we are to get anywhere at all, we need to approximate the true function f_{pr} with a function in a hypothesis set, $\mathcal{H}_{pr}$. Let us examine the range of all elements in $\mathcal{H}_{pr}$. What values can these functions return and why?

These functions can only return values in $(0, 1)$ because we want to define a probability for attaining a 1.

(b) [difficult] We would also feel warm and fuzzy inside if the elements of $\mathcal{H}_{pr}$ contained the term $\boldsymbol{w} \cdot \boldsymbol{x}$. What is the main reason we would like our prediction functions to contain this linear component?

Our prediction functions should contain this linear component so if we have a high $\boldsymbol{w} \cdot \boldsymbol{x}$, we can return a high probability using the function in $\mathcal{H}_{pr}$.

(c) [easy] The problem is $\boldsymbol{w} \cdot \boldsymbol{x} \in \mathbb{R}$ but in (a) there is a special range of allowable functions. We need a way to transform $\boldsymbol{w} \cdot \boldsymbol{x}$ into the range from (a). What is this function called?

link/activation function.

(d) [easy] Give some examples of such functions.

- Logistic function: $\Phi(u) = \frac{e^u}{1+e^u}$
- Complementary Log-Log function: $\Phi(u) = 1 - e^{-e^u}$
- Hyperbolic Tangent function: $\Phi(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}}$

(e) [easy] We will choose the logistic function. Write the likelihood again from 2(g) but replace f_{pr} with the element from $\mathcal{H}_{pr}$ that uses the logistic function.

$$\mathbb{P}\left(Y_1, \ldots, Y_n\right) = \prod_{i=1}^{n} (\frac{e^{\boldsymbol{w} \cdot \boldsymbol{x}}}{1 + e^{\boldsymbol{w} \cdot \boldsymbol{x}}})^{y_i}(1 - \frac{e^{\boldsymbol{w} \cdot \boldsymbol{x}}}{1 + e^{\boldsymbol{w} \cdot \boldsymbol{x}}})^{1-y_i}$$

(f) [difficult] Simplify your answer from (e) so that you arrive at:

$$\sum_{i=1}^{n} \ln\left(1 + e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i}\right)$$

$$\begin{aligned}
\boldsymbol{b} &= \underset{\boldsymbol{w}}{\arg\max}\ \left\{\mathbb{P}\left(Y_i, \ldots, Y_n\right)\right\} \\
&= \underset{\boldsymbol{w}}{\arg\max}\ \left\{\prod_{i=1}^{n} (\frac{e^{\boldsymbol{w}\cdot\boldsymbol{x}_i}}{1+e^{\boldsymbol{w}\cdot\boldsymbol{x}_i}})^{y_i}(1 - \frac{e^{\boldsymbol{w}\cdot\boldsymbol{x}_i}}{1+e^{\boldsymbol{w}\cdot\boldsymbol{x}_i}})^{1-y_i}\right\} \\
&= \underset{\boldsymbol{w}}{\arg\max}\ \left\{\prod_{i=1}^{n} (\frac{1}{1+e^{-\boldsymbol{w}\cdot\boldsymbol{x}_i}})^{y_i}(\frac{1}{1+e^{\boldsymbol{w}\cdot\boldsymbol{x}_i}})^{1-y_i}\right\} \\
&= \underset{\boldsymbol{w}}{\arg\max}\ \left\{\prod_{i=1}^{n} \begin{cases} (1+e^{-\vec{w}\cdot\vec{x}_i})^{-1} & \text{if } y_i = 1 \\ (1+e^{\vec{w}\cdot\vec{x}_i})^{-1} & \text{if } y_i = 0 \end{cases}\right\} \\
&= \underset{\boldsymbol{w}}{\arg\max}\ \left\{\prod_{i=1}^{n} (1+e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i})^{-1}\right\} \\
&= \underset{\boldsymbol{w}}{\arg\max}\ \left\{\sum_{i=1}^{n} \ln(1+e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i})\right\}
\end{aligned}$$

(g) [E.C.] We will now maximize this likelihood w.r.t to $\boldsymbol{w}$ to find $\boldsymbol{b}$, the best fitting solution which will be used within g_{pr} i.e.

$$\boldsymbol{b} = \underset{\boldsymbol{w}\in\mathbb{R}^{p+1}}{\arg\max}\ \left\{\sum_{i=1}^{n} \ln\left(1 + e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i}\right)\right\}$$

to do so, we should find the derivative and set it equal to zero i.e.

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{w}}\left[\sum_{i=1}^{n} \ln\left(1 + e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i}\right)\right] \overset{\text{set}}{=} 0$$

Try to find the derivative and solve. Get as far as you can. Do so on a separate page

See page 12.

(h) [easy] If you attempted the last problem, you found that there is no closed form solution. What type of methods are used to approximate $\boldsymbol{b}$? Note: once you use such methods and arrive at a $\boldsymbol{b}$, that is called "running a logistic regression".

To approximate $\boldsymbol{b}$, use numerical methods such as gradient descent (or my favorite, stochastic gradient descent).

(i) [easy] In class we used the notation $\hat{p} = g_{pr}$. Why?

We want to find the probability of attaining a 1 in our model using $\boldsymbol{x}$ and so we let that equal g_{pr}.

(j) [easy] Write down $\hat{p}$ as a function of $\boldsymbol{b}$ and $\boldsymbol{x}$.

$$\hat{p} = (1 + e^{-\boldsymbol{b} \cdot \boldsymbol{x}})^{-1}$$

(k) [harder] What is the interpretation of the linear component $\boldsymbol{b} \cdot \boldsymbol{x}$? What does it mean for $\hat{p}$? No need to give the full, careful interpretation.

The interpretation of the linear component is $\ln(\frac{\hat{p}}{1-\hat{p}})$. This is the log-odds of Y given $\boldsymbol{x}$. The more positive and larger the log odds is, the higher the probability is.

(l) [difficult] How does one go about *validating* a logistic regression model? What is the fundamental problem with doing so that you didn't have to face with regression or classification? Discuss.

To validate a logistic regression model, validate against f_{pr}; but this is not possible and so we have to validate $\hat{p}$ versus $\boldsymbol{y}$. This is different from validation in regression and classification because we are not directly comparing our model to f_{pr}, but rather just looking at probabilities. $\boldsymbol{y}$ is in $\mathbb{R}$ whereas $\hat{p}$ is in $(0, 1)$. We have to turn a $\hat{p}$ into a $\hat{\boldsymbol{y}}$ using probability estimation to do classification.

Problem 4

This question is about probabilistic classification i.e. using probability estimation to classify. We limit our discussion to binary classification.

(a) [easy] How do you use a probability estimation model to classify. Provide the formula which provides $\hat{y}(\hat{p})$ i.e. the estimate of whether the event of interest occurs as a function of the probability estimate of the event occurring. Use the "default" rule.

$$\hat{y}_i = \mathbb{1}_{\hat{p}_i \geq 0.5}$$

(b) [easy] In the formula from (a), there is an option to be made, write the formula again below with this option denoted p_{th}.

$$\hat{y}_i = \mathbb{1}_{\hat{p}_i \geq p_{th}}$$

(c) [harder] What happens when p_{th} is low and what happens when p_{th} is high? What is the tradeoff being made?

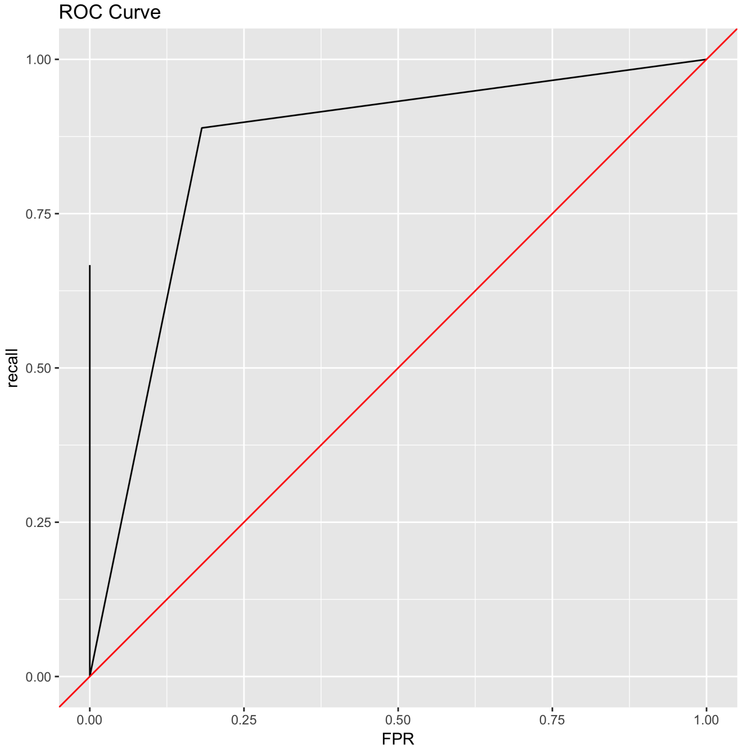
When p_{th} is low, there are more false positives and less false negatives because it is easier to get a 1 than a 0. On the other hand, when p_{th} is high, there are less false positives and more false negatives because it is harder to get a 1 than a 0. The tradeoff being made is between how many false positive and false negatives you have.

(d) [difficult] Below is the first 20 rows of in-sample prediction results from a logistic regression whose reponse is $> 50K$ (the positive class) or $\leq 50K$ (the negative class). You have the $\hat{p}_i$'s and the y_i's. Create a performance table that includes the four numbers in the confusion table as well as FPR and recall. Leave some room for one additional column we will compute later in the question. The rows in the table should be indexed by $p_{th} \in \{0, 0.2, \ldots, 0.8, 1\}$ which you should use as the first column. Hint: you may want to sort by $\hat{p}$ and convert y to binary before you begin.

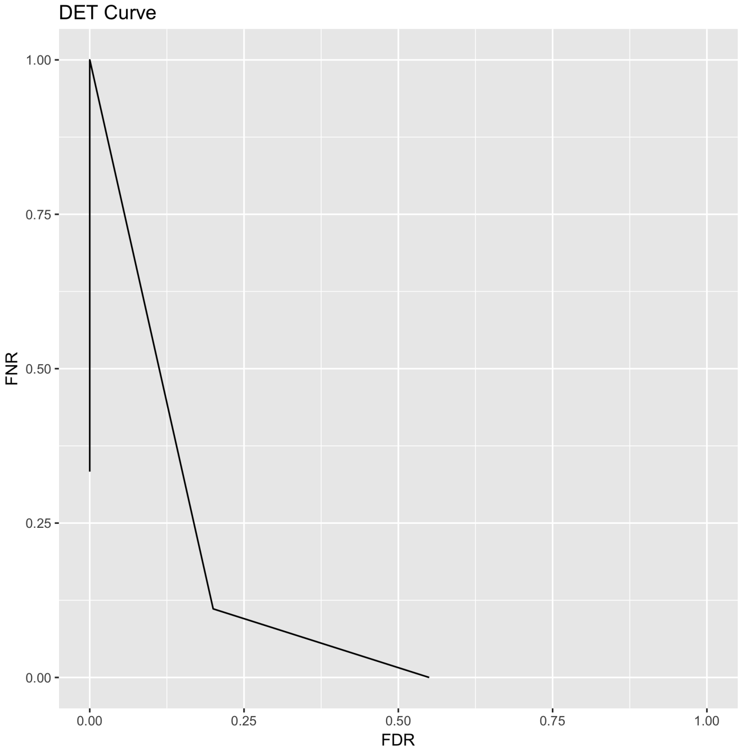
$\hat{p}$	y	y	$\mathbb{1}_{\hat{p}\geq 0.0}$	$\mathbb{1}_{\hat{p}\geq 0.2}$	$\mathbb{1}_{\hat{p}\geq 0.4}$	$\mathbb{1}_{\hat{p}\geq 0.6}$	$\mathbb{1}_{\hat{p}\geq 0.8}$	$\mathbb{1}_{\hat{p}\geq 1.0}$
0.00	<=50K	0	1	0	0	0	0	0
0.00	<=50K	0	1	0	0	0	0	0
0.01	<=50K	0	1	0	0	0	0	0
0.01	<=50K	0	1	0	0	0	0	0
0.01	<=50K	0	1	0	0	0	0	0
0.02	<=50K	0	1	0	0	0	0	0
0.07	<=50K	0	1	0	0	0	0	0
0.07	<=50K	0	1	0	0	0	0	0
0.07	>50K	1	1	0	0	0	0	0
0.08	<=50K	0	1	0	0	0	0	0
0.32	<=50K	0	1	1	0	0	0	0
0.35	>50K	1	1	1	0	0	0	0
0.35	>50K	1	1	1	0	0	0	0
0.38	<=50K	0	1	1	0	0	0	0
0.49	>50K	1	1	1	1	0	0	0
0.59	>50K	1	1	1	1	0	0	0
0.69	>50K	1	1	1	1	1	0	0
0.73	>50K	1	1	1	1	1	0	0
0.76	>50K	1	1	1	1	1	0	0
0.91	>50K	1	1	1	1	1	1	0

p_{th}	TP	TN	FP	FN	FPR	Recall	FDR	FNR
0.0	9	0	11	0	$\nicefrac{11}{11}$	$\nicefrac{9}{9}$	$\nicefrac{11}{20}$	$\nicefrac{0}{9}$
0.2	8	9	2	1	$\nicefrac{2}{11}$	$\nicefrac{8}{9}$	$\nicefrac{2}{10}$	$\nicefrac{1}{9}$
0.4	6	11	0	3	$\nicefrac{0}{11}$	$\nicefrac{6}{9}$	$\nicefrac{0}{6}$	$\nicefrac{3}{9}$
0.6	4	11	0	5	$\nicefrac{0}{11}$	$\nicefrac{4}{9}$	$\nicefrac{0}{4}$	$\nicefrac{5}{9}$
0.8	1	11	0	8	$\nicefrac{0}{11}$	$\nicefrac{1}{9}$	$\nicefrac{0}{1}$	$\nicefrac{8}{9}$
1.0	0	11	0	9	$\nicefrac{0}{11}$	$\nicefrac{0}{9}$	$\nicefrac{0}{0}$	$\nicefrac{9}{9}$

(e) [harder] Using the performance table from (d), trace out an approximate ROC curve.



(f) [harder] Using the performance table from (d), trace out an approximate DET curve.



(g) [easy] Consider the $c_{FP} = \$5$ and $c_{FN} = \$1,000$. Explain how you would find the probabilistic classifier model that minimizes cost among the p_{th} values you considered in your performance table in (d) but do not do any computations.

To minimize costs, you want to minimize FN because since the cost of getting a false negative is $\$1,000$. To do this, keep p_{th} small. Create probabilistic classifier models using small p_{th} on $\mathcal{D}_{\text{train}}$ and then test it on $\mathcal{D}_{\text{test}}$. Determine which p_{th} creates the biggest average reward and use that p_{th}.

Problem 5

These are questions related to bias-variance decomposition, bagging and random forests.

(a) [easy] List the assumptions for the bias-variance decomposition.

Assume that $\mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] = f(\boldsymbol{x})$, or $\mathbb{E}[\Delta \mid \boldsymbol{X} = \boldsymbol{x}] = 0$ where Δ is the random variable for error. Also assume that the variance does not depend on $\boldsymbol{x}$, meaning $\mathbb{V}\text{ar}[\Delta \mid \boldsymbol{X} = \boldsymbol{x}] = \mathbb{V}\text{ar}[\Delta] = \sigma^2$ and so $\mathbb{E}[\Delta^2] = 0$.

(b) [harder] Why is $f(\boldsymbol{x})$ called the "conditional expectation function"?

$f(\boldsymbol{x})$ is called the conditional expectation function because it is the expected value of Y given $\boldsymbol{X} = \boldsymbol{x}$, or $\mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}]$.

(c) [easy] Provide an expression for the bias-variance decomposition formula for the average MSE over the distribution $\mathbb{P}(\boldsymbol{X})$ for $y = g + (f - g) + \delta$. You should have three terms in the expression. Make sure you explain conceptually each term in English.

$$\text{MSE} = \sigma^2 + \text{E}_{\boldsymbol{X}}[\mathbb{V}\text{ar}[g(\boldsymbol{x})]] + \text{E}_{\boldsymbol{X}}[\text{Bias}[g(\boldsymbol{x})]^2]$$

where σ^2 is the variance of Y, the middle term explains how much g varies around its mean function $\mathbb{E}[g(\boldsymbol{x})]$ and the last term explains how far away the average g is from f.

(d) [E.C.] Rederive the bias-variance decomposition formula for the average MSE over the distribution $\mathbb{P}(\boldsymbol{X})$ for $y = g + (h^* - g) + (f - h^*) + \delta$. You should group the final expression into *four* terms where two will be the same as the expression found in (c), one will be similar to a term found in (c) and one will be new. Make sure you explain conceptually each term in English. Do so on an additional page.

See page 13.

(e) [harder] Assume a $\mathbb{D}$ where n is large and p is small and you fit a linear model g to all features. Your in-sample R^2 is low. In the expression from (c), indicate term(s) which are likely large, which term(s) are likely small and explain why.

If n is large and p is small and in-sample R^2 is low, then the variance term $\text{E}_{\boldsymbol{X}}[\mathbb{V}\text{ar}[g(\boldsymbol{x})]]$ is small due to the large sample size. The σ^2 term is fixed all throughout because it is always there. The bias term $\text{E}_{\boldsymbol{X}}[\text{Bias}[g(\boldsymbol{x})]^2]$ will also become small since the average g will not be much far away from f since the sample size is huge.

(f) [harder] Assume a $\mathbb{D}$ where n is large and p is small and you fit a tree model g to all features. Your in-sample R^2 is low. In the expression from (c), indicate term(s) are likely large, which term(s) are likely small and explain why.

If a tree model g is fit to all models where n is large and p is small and R^2 is low, then the bias term will be small since the bias term for each tree will be small because trees are not biased. The variance term will also be small since n is large. The $\sigma * 2$ term will be small because it is the best the tree model can do.

(g) [easy] Provide an expression for the bias-variance decomposition formula for the average MSE over the distribution $\mathbb{P}(\boldsymbol{X})$ for $y = g + (f - g) + \delta$ where g now represents the average taken over constituent models $g_1, g_2, \ldots, g_T$. (This is known as "model averaging" or "ensemble learning"). You can assume that $\rho := \text{Corr}[g_{t_1}, g_{t_2}]$ is the same for all $t_1 \neq t_2$.

$$\text{MSE} = \sigma^2 + \left(\rho \mathbb{V}\text{ar}[g_t] + \frac{1-\rho}{T}\mathbb{V}\text{ar}[g_t]\right) + \text{Bias}[g_t]^2$$

(h) [easy] If $T \to \infty$, rewrite the bias-variance decomposition you found in (k).

$$\text{MSE} = \sigma^2 + \text{Bias}[g_t]^2$$

(i) [easy] If $g_1, g_2, \ldots, g_T$ are built with the same data $\mathbb{D}$ and $\mathcal{A}$ is not random, then $g_1 = g_2 = \ldots = g_T$. What would ρ be in this case?

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

(j) [easy] Even though each of the constituent models $g_1, g_2, \ldots, g_T$ are built with the same data $\mathbb{D}$, what idea can you use to induce $\rho < 1$? This idea is called "bagging" which is a whimsical portmanteau of the words "bootstrap aggregation".

To induce $\rho < 1$, aggregate trees $g_i = \mathcal{A}(\mathcal{H}, \mathbb{D}_{(i)})$ where $\mathcal{A}$ and $\mathcal{H}$ are regression trees and $\mathbb{D}_{(i)}$ is the original $\mathbb{D}$. Then find the average/aggregate tree by averaging all the trees. When many trees are averaged together, the variance of the average g will show that $\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} = \frac{\sigma_{ij}}{\sigma^2}$ because $\sigma_i = \sigma_j$.

(k) [easy] Explain how examining predictions averaged on the out of bag (oob) data for each $g_1, g_2, \ldots, g_T$ can constitute model validation for the bagged model.

Examining predictions averaged on the out of bag data can constitute model validation because each tree can validate itself by predicting on $\mathbb{D}_{\text{test},t}$. When each tree does this and then averaged out, all observations will be validated.

(l) [easy] Explain how the Random Forests® algorithm differs from the CART (classification and regression trees) algorithm.

The Random Forests® algorithm is different from the CART algorithm in the manner that the choices for which features to use is randomized. In the Random Forests® algorithm, not all features of $\boldsymbol{Y}$ are used whereas in CART algorithm, all of $\boldsymbol{Y}$ is used.

(m) [easy] Explain why the MSE for the Random Forests® algorithm expected to be better than for the bag of CART models.

The MSE for the Random Forests® algorithm is expected to be better than the bag of CART models because each better features that explain $\boldsymbol{Y}$ will be used to create the model and so f will be closer to h^*.

(n) [easy] List the three major advantages of Random Forests® for supervised learning / machine learning.

The three major advantages of Random Forests® are reduction in overfitting, lower variance and high accuracy.

Problem 6

These are questions related to correlation, causation and the interpretation of coefficients in linear models / logistic regression.

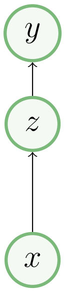
(a) [easy] You are provided with the responses measured from a phenomenon of interest $y_1, \ldots, y_n$ and associated measurements $x_1, \ldots, x_n$ where n is large. The sample correlation is estimated to be $r = 0.74$. Is $\boldsymbol{x}$ "correlated" with $\boldsymbol{y}$?

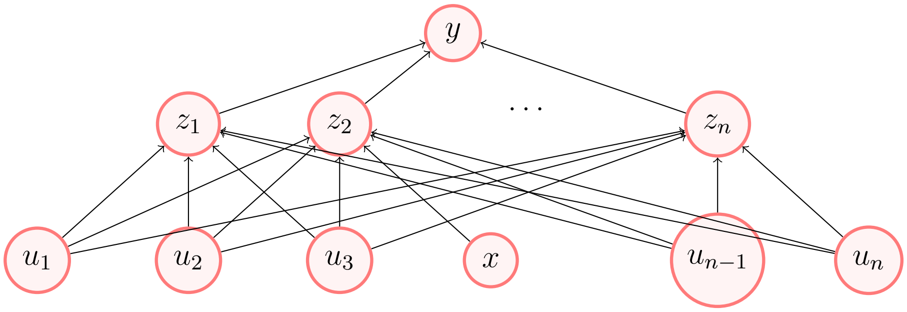
$\boldsymbol{x}$ is correlated with $\boldsymbol{y}$.

(b) [harder] Consider the case in (a), would $\boldsymbol{x}$ be a "causal" factor for $\boldsymbol{y}$? Explain.

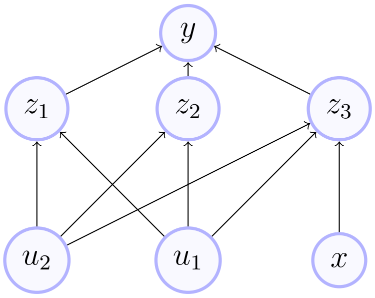
There is not enough information to claim that $\boldsymbol{x}$ could be a "causal" factor for $\boldsymbol{y}$. More testing will be needed.

(c) [harder] Consider the case in (a) and create two plausible causal models using the graphical depiction style used in class (nodes representing variables and lines represent causal contribution where node A below node B means node A is measured before node B). Your model has to include x and y but is not limited to only those variables.





(d) [harder] Consider the case in (a) but now n is small. Create a third plausible causal model (in addition to the two you created in the last problem) using the same graphical depiction style. Your model has to include x and y but is not limited to only those variables.



(e) [easy] Explain briefly how you would prove beyond a reasonable doubt that $\boldsymbol{x}$ is not only correlated with $\boldsymbol{y}$ but that $\boldsymbol{x}$ is a causal factor of $\boldsymbol{y}$.

To prove that $\boldsymbol{x}$ is a causal factor of $\boldsymbol{y}$, perform randomized controlled experiments.

(f) [easy] Consider $\boldsymbol{x}$ is college GPA and $\boldsymbol{y}$ is career average income. Is $\boldsymbol{x}$ correlated with $\boldsymbol{y}$?

If $\boldsymbol{x}$ is college GPA and $\boldsymbol{y}$ is career average income, then $\boldsymbol{x}$ is correlated with $\boldsymbol{y}$.

(g) [harder] Consider $\boldsymbol{x}$ is college GPA and $\boldsymbol{y}$ is career average income. Is $\boldsymbol{x}$ a causal factor of $\boldsymbol{y}$?

If $\boldsymbol{x}$ is college GPA and $\boldsymbol{y}$ is career average income, then $\boldsymbol{x}$ is not a causal factor of $\boldsymbol{y}$.

(h) [harder] Consider $\boldsymbol{x}$ is college GPA and $\boldsymbol{y}$ is career average income. Can you think of a $\boldsymbol{z}$ which is a lurking variable? Explain the variable and why you believe it fits the description of a lurking variable.

If $\boldsymbol{x}$ is college GPA and $\boldsymbol{y}$ is career average income, then z, a lurking variable, can be the location of career. Different locations can have different career average income. Furthermore, college GPA will come into play when determining whether employers in the location will consider the student for the workforce.

(i) [harder] If you fit a linear model for $\boldsymbol{y}$, $g = b_0 + b_x x + b_z z$, what would the b_x value be close to? Why?

If a linear model for $\boldsymbol{y}$ is fitted, then the b_x value would be close to 0. This is because when z is held constant, the only causal factor in $\boldsymbol{y}$, manipulating x won't affect y and so b_x would be close to 0.

(j) [E.C.] Create a causal model using the same graphical depiction style that justifies the four linear regression assumptions. Do so on a different page.

(k) [harder] When running a regression of `price` on all variables in the `diamonds` dataset, the coefficient for `carat` is about \$6,500. Interpret this value as best as you can.

When comparing two observations, A and B, sampled in the same way as the data in `diamonds`, where A has a `carat` value one unit higher than the `carat` value of B, and all other features are exactly the same, then A is predicted to have a response $\texttt{price}_A$ that differs by \$6,500, on average, from $\texttt{price}_B$ assuming the linear model optimal for least squares.

(l) [harder] When running a logistic regression of class `malignant` on all variables in the `biopsy` dataset, the coefficient for `V1` (which measures clump thickness) is about 0.54. Interpret this value as best as you can.

When comparing two observations, A and B, sampled in the same way as the data in `biopsy`, where A has a `V1` value one unit higher than the `V1` value of B, and all other features are exactly the same, then A is predicted to have a response $\texttt{malignant}_A$ that differs by 0.54, on average, from $\texttt{malignant}_B$ assuming the logistic model optimal for least squares.

Problem 7

Extra Credit Problems

(a) [E.C.] We will now maximize this likelihood w.r.t to $\boldsymbol{w}$ to find $\boldsymbol{b}$, the best fitting solution which will be used within g_{pr} i.e.

$$\boldsymbol{b} = \underset{\boldsymbol{w}\in\mathbb{R}^{p+1}}{\arg\max}\left\{\sum_{i=1}^{n}\ln\left(1+e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i}\right)\right\}$$

to do so, we should find the derivative and set it equal to zero i.e.

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{w}}\left[\sum_{i=1}^{n}\ln\left(1+e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i}\right)\right] \overset{\text{set}}{=} 0$$

$$\begin{aligned}
\frac{d}{d\boldsymbol{w}}\Big(\sum_{i=1}^{n}\ln(1+e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i})\Big) &= \sum_{i=1}^{n}\frac{d}{d\boldsymbol{w}}\ln(1+e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i})\\
&= \sum_{i=1}^{n}\frac{1}{1+e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i}}\cdot e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i}\cdot(1-2y_i)\boldsymbol{x}_i\\
&= \sum_{i=1}^{n}\frac{(1-2y_i)\boldsymbol{x}_i e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i}}{1+e^{(1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i}}\\
&= \sum_{i=1}^{n}(1-2y_i)\boldsymbol{x}_i\Phi((1-2y_i)\boldsymbol{w}\cdot\boldsymbol{x}_i)
\end{aligned}$$

where $\Phi(u) = \frac{e^u}{1+e^u}$, the logistic function. Note that there is no closed form for this when set equal to 0.

(b) [E.C.] Rederive the bias-variance decomposition formula for the average MSE over the distribution $\mathbb{P}(\boldsymbol{X})$ for $y = g + (h^* - g) + (f - h^*) + \delta$. You should group the final expression into *four* terms where two will be the same as the expression found in (c), one will be similar to a term found in (c) and one will be new. Make sure you explain conceptually each term in English. Do so on an additional page.

Here, $Y \mid \boldsymbol{X} = \boldsymbol{x} = g(\boldsymbol{x}) + (h^*(\boldsymbol{x}) - g(\boldsymbol{x})) + (f(\boldsymbol{x}) - h^*(\boldsymbol{x})) + \Delta$. Then

$$\mathbb{E}\left[(Y - g(\boldsymbol{x})|\boldsymbol{X})^2\right] = \mathbb{E}\left[(h^*(\boldsymbol{x}) - g(\boldsymbol{x})) + (f(\boldsymbol{x} - h^*(\boldsymbol{x})) + \Delta)^2\right]$$

Now

$$\begin{aligned} MSE &= \mathbb{E}\left[\underbrace{(h^* - g)^2}_{V} + (\overset{B}{\underbrace{f - h^*}})^2 + \Delta^2 + 2(\underbrace{h^* - g}_{V})\Delta + 2(\underbrace{f - h^*}_{B})\Delta + 2(\underbrace{h^* - g}_{V})(\underbrace{f - h^*}_{B})\right] \\ &= \mathbb{E}\left[V^2\right] + \mathbb{E}\left[B^2\right] + \mathbb{E}\left[\Delta^2\right] + 2\mathbb{E}\left[V\right]\Delta + 2\mathbb{E}\left[B\Delta\right] + 2\mathbb{E}\left[VB\right] \\ &= \mathbb{E}\left[V^2\right] + B^2 + \sigma^2 + 2B\mathbb{E}\left[V\right] \end{aligned}$$

The σ^2 term is the variance of Y. The B^2 term is the bias term, explaining how far f is from the true phenomenon h^*. The $\mathbb{E}\left[V^2\right]$ term explains how much g varies from the its mean function $\mathbb{E}\left[(h^* - g)^2\right]$.