MATH 390.4 / 650.2 Spring 2018 Homework #3t

Joseph Peltroche

Thursday 22nd March, 2018

Problem 1

These are questions about Silver's book, chapter 2.

(a) [harder] If one's goal is to fit a model for a phenomenon y, what is the difference between the approaches of the hedgehog and the fox? 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.). Connecting this to the modeling framework should really make you think about what Tetlock's observation means for political and historical phenomena.

Solution. Foxes consider the uncertainty of a situation to have a certain outcome, and realize their predictions are an appoximation to reality. They see their model as

$$y = \hat{y} + e = g(\vec{x}) + e$$

aware of the inevitable error present when forecasting, incorporating it into their model and thereby becoming stronger forecasters. The hedgehogs on the other hand, are overly confident in their predictions and treat it as if it were the reality

$$y = \hat{y} = g(\vec{x})$$

ignoring the error, becoming more vulnerable to mistakes as a result and thereby being the weaker forecaster. □

(b) [easy] Why did Harry Truman like hedgehogs? Are there a lot of people that think this way?

Solution. Harry Truman liked hedgehogs over foxes because he was fed up with what he intrepreted as a lack of conviction; the foxes in his adminstration couldn't give him a qualified answer because they were explicit with the uncertainities and the noise they were able to recognize in their data. Foxes, although capable of making better predictions, come off as less confident than hedgehogs, who are brimming with confidence yet are susceptible to produce predictions that are incorrect. Stubborn and close-minded, the hedgehogs can give a confidently direct answer, even if it is incorrect

or bound to be incorrect, as long as they believe in it. They do not consider the uncertainities and noise to the same degree as the foxes, and win many people in doing so, including Harry Truman who wanted a confident forecaster who could give him an explicit answer. There are a lot of people who think in this "type A" cultural mindset which is the reason by big, bold hedgehogs like Dick Morris are so popular in television and other forms of media. □

(c) [difficult] Why is it that the more education one acquires, the less accurate one's predictions become?

Solution. The more education one requires, the stronger possibility there is too introduce more noise into your data, and making it harder to distinguish the signal from the noise. The likelihood of getting caught up with unnecessary information increases, and there is a danger if the forecaster is a hedgehog. Contrary to foxes which are immune to this effect, hedgehogs use the additional information to the advantage of their bias. The more facts hedgehogs have, the more opportunities they have to manipulate them in ways that confirm their beliefs and biases, making them more prone to weaker predictions. □

(d) [easy] Why are probabilistic classifiers (i.e. algorithms that output functions that return probabilities) better than vanilla classifiers (i.e. algorithms that only return the class label)? We will move in this direction in class soon.

Solution. Rather than producing a prediction that is a class label, a probabilisitic classifier brings the uncertainties of our prediction and model to our attention. A vanilla classifier neglects the probablisitic nature of a prediction; instead the algorithm produces a clear answer but without any uncertainty. Despite probabilities (output of the probablisitic classifier) being intrepreted as lacking clarity, it correctly claims the imperfections of theories attempting to mimick reality and can help us interpret what our model is really producing and to improve our models towards a better approximation. □

Problem 2

These are questions about Finlay's book, chapter 2-4. We will hold off on chapter 1 until we cover probability estimation after midterm 2.

(a) [easy] What term did we use in class for "behavioral (outome) data"?

Solution. the behavioral data in Finlay's book is the y term we use in our training data; in class it has taken the name the output, the response, endpoint, or the dependent variable. □

(b) [easy] Write about some reasons why data scientists implement models that are subpar in predictive performance (p27).

Solution. Data scientist will sacrifice a small amount of predictive performance in order the meet the demands of business requirements and constraints that have to be taken into account. Of course data scientists wouldn't want to give up any predictability, but these requirements are unavoidable. This may force the model developr to force certain variables to feature in the model or ensure that certain ones are excluded. An example would be creating a model for an industry simple enough so non-experts could understand and implement. □

(c) [easy] In the first wine example, what is the outcome metric and what kind of supervised learning was employed?

Solution. The outcome metric in the first wine example was the number of new customers and the supervised learning employed was a "Decision Tree" model (a classification model) using previously obtained training data. □

(d) [easy] In the second wine example, what is the outcome metric and kind of supervised learning was employed?

Solution. The outcome metric in the second wine example was the net profit itself and the kind of supervised learning employed was the gross profit model (a regression model). □

(e) [easy] In the third chapter, why is it that some organizations cannot use predictive modeling to improve their business?

Solution. Some organizations cannot use predictive modeling in an effort to improve their business because it has not been fully accepted throughout the organization. If there is a clash between the culture of the organization and the acceptance of predictve modeling then the disagreement can cause to predictive model to never be used in the first place. This includes senior managers to front line staff that must be comfortable with the idea of automated data driven decision making. □

(f) [easy] In the bankruptcy case, what is the problem with merely using g to obtain a $\hat{y}$ without any other information from the model?

Solution. The problem was that the model had failed to correctly intrepret cases where something clearly didn't stack up and pass the information on to the investigaor to use as leverage when investigating a case. This was contrary to the review team, which were able to assess the vallue and "workability" of cases, i.e. cases with a reasonable chance of finding the fraudster and recovering money from them. □

(g) [easy] Chapter 3 talks about using the model with human judgment. Under what circumstances is this beneficial? 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.).

Solution. Human involvement depends on the volume and value of the decision (the dimensions and importance of y; the n), e.g. life or death decision regarding thousands of patients contracting a certain disease. In this example it would be better to introduce a expert opinion in cojunction with this model. Really this means the model would be used a prioritization tool for the human experts to judge, i.e. our g would develop a heirachy for every $\hat{y}_i$ and the list produced would be passed to experts in the field who can review each case. □

(h) [difficult] In Chapter 4 Finaly makes an interesting observation based on his experience in data science. He says most predictive models have $p \leq 30$. Why do you think this is? Discuss.

Solution. Finaly says most predictive models have $p \leq 30$ based on his experience in data science. I believe this is due to the complexity of the model and how it can worsen the forecast and the model's capability. Having a $p < 30$ isn't impossible, but can hurt your prediction-making capabilities because the forecaster runs the risk of introducing excessive information that can cause the model to be more complex than it needs to be. It runs of the risk of introducing correlated data types that were not necessary in the first place and runs the risk of introducing information that simply isn't revelant to the specific behavioral outcome. In the era of Big Data, this is a big concern and incorporating too much data (excessive predictors) with little revelance can often hurt more than help forecasting. □

(i) [easy] He says there is "almost always other data that could be acquired ... [which] doesn't always come for free". The "data" he is talking about here specifically means "more predictors" i.e. increasing p. In what cases would someone be willing to pay for this data?

Solution. People would be willing to increase the number of predictors if the return value for the predictors is significant enough to purchase. Out of the predictors that can be incorporated into a model for forecasting, the primary and secondary behaviors and geo-demographic data are some of the most important preditctors to obtain and worth paying for. □

(j) [easy] Table 4 lists "data types" about what type of observations?

Solution. The different data types are categorizing predictors into notable groups of predictive capability. Each of the different data types are observing not merely the behavior they want to predict but any behavior correlated and associated to it, along with information that are harder to access but could have a some associaton to the primary behavior sucha as sentiments and network data types. In Table 4, burglary was used an example to predict, but this data classification can be applied to all sorts of prediction problems. □

(k) [easy] What type of data does he find in his experience to be the most important to predictive modeling? Why do you think this is so?

Solution. Finlay finds the primary behavior to be the most important by a considerable margin. I believe it is because any historical evidence of what behavior the model is attempting to predict can mandate whether it can actually happen again in the future. The likelihood of something occuring in the future shoots up if it has happened before in the past; if someone does something once then there is a greater chance that same person will do it again. This is the idea of supervised learning. □

(l) [easy] If $x_{\cdot 17}$ was age and $x_{\cdot 18}$ is age of spouse, what is the most likely reason why adding $x_{\cdot 18}$ to $\mathbb{D}$ not be friutful for predictive ability?

Solution. Adding x_{18} to $\mathbb{D}$ will not really be fruitful and will be like adding excessive information to the training data because having x_{17} will be a strong indicator as to what the age of the spouse is, usually about 90% that the age of the spouse is relatively close to your own age. Hence it will be like adding strongly correlated information and will not contribute any additional predictive ability. □

(m) [difficult] What is the lifespan of a predictive model? Why does it not last forever? 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.).

Solution. The lifespan of a predictive model varies, depending on the relationships($g(\vec{x}_i)$) found between the predictor data $\vec{x}_1$ and the outcome data y_i. Over time, the relationship $g(\vec{x}_i)$ deteriorates. For some domain applications, it can take years for a model to deteriorate to the point that is needs replacing, i.e. $g(\vec{x}_i) \neq y_i$, while for markets that are constantly changing in real time, models are rebuilt on a daily or more frequent basis. □

(n) [difficult] What does "large enough to representative of the full population" (p80) mean? 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.).

Solution. By this quote Finaly means a way of optimizing $\mathbb{D}$ and obtaining a small enough sample size n with a small enough $p+1$ predictors such that it captures the patterns of behavior $f(\vec{x})$ while being small enough to actually analyze $\mathbb{D}$ and practical enough to use for our model. □

(o) [easy] Is there a hype about "big data" i.e. including millions of observations instead of a few thousand? Discuss Finlay's opinion.

Solution. According to Finaly there is a hype about big data; every few years a myth circulates the predictive analytics community about the benefits produced from using big data in your model. In his opinion, yes bigger samples mean better, but there is a pay-off that he stresses. Sometimes the difference in accuracy between different orders of magnitudes of sample sizes will not be large enough even pursue. Sometimes the increasing a sample size by orders of magnitude would produce only about a fraction of a percent better accuracy than using the smaller sample, yet the obtaining the larger sample will require investing heavily on technology to be able to do full population modeling. □

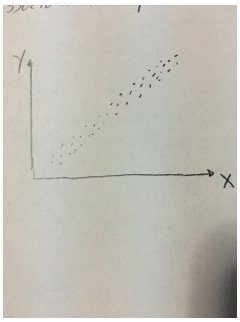
(p) [easy] What is Finlay's solution to "overfitting" (p84)?

Solution. Finlay's solution to overfitting is to use larger samples. There is a lesser likelihood overfitting would occur with a larger sample over a smaller sample. Finaly discusses how accuracy between two models is larger than it should be if a model is overfitted due to a smaller sample vs a model constructed using a large sample where over-fitting has not occured. □

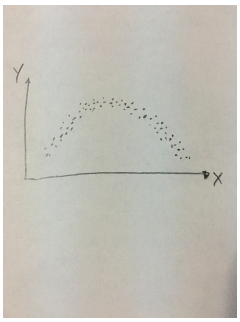
Problem 3

These are questions about association and correlation.

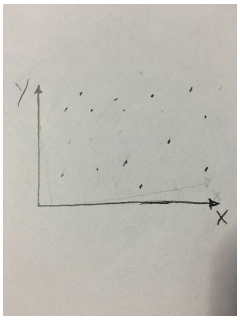
(a) [easy] Give an example of two variables that are both correlated and associated by drawing a plot.



(b) [easy] Give an example of two variables that are not correlated but are associated by drawing a plot.



(c) [easy] Give an example of two variables that are not correlated nor associated by drawing a plot.



(d) [easy] Can two variables be correlated but not associated? Explain.

Solution. Since correlation is within the greater concept of dependence, that is, correlation$\in$ association, then to be correlated means to be associated. Correlation is a form of association, and so one can not be correlated without being associated, hence it is not possible to be correlated yet independent (not associated) between the variables in consideration). □

Problem 4

These are questions about multivariate linear model fitting using the least squares algorithm.

(a) [difficult] Derive $\dfrac{\partial}{\partial \boldsymbol{c}}\left[\boldsymbol{c}^\top A\boldsymbol{c}\right]$ where $\boldsymbol{c} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$ but *not* symmetric. Get as far as you can.

Solution.

$$\frac{\partial}{\partial \vec{c}}[\underbrace{\vec{c}^T A\vec{c}}_{\text{quadratic form}}] = \frac{\partial}{\partial \vec{c}}[\vec{c}^T(A\vec{c})]$$

$$A\vec{c} = \begin{bmatrix} a_{11}c_1 + a_{12}c_2 + \ldots + a_{1n}c_n \\ \vdots \\ a_{n1}c_1 + a_{n2}c_2 + \ldots + a_{nn}c_n \end{bmatrix} \in \mathbb{R}^{n \times 1}$$

$$\vec{c}^T(A\vec{c}) = c_1(a_{11}c_1 + a_{12}c_2 + \ldots + a_{1n}c_n) + c_2(a_{21}c_1 + a_{22}c_2 + \ldots + a_{2n}c_n) + \ldots$$
$$\ldots + c_n(a_{n1}c_1 + a_{n2}c_2 + \ldots + a_{nn}c_n)$$

So, just one partial derivative is:

$$\frac{\partial}{\partial c_1}[\vec{c}^T(A\vec{c})] = (2c_1a_{11} + c_2a_{12} + c_3a_{13} + \ldots + c_na_{1n}) + c_2a_{21} + c_3a_{31} + \ldots + c_na_{n1}$$

$$= 2c_1a_{11} + c_2(a_{12} + a_{21}) + c_3(a_{13} + a_{31}) + \ldots + c_n(a_{1n} + a_{n1})$$

A similar outcome results for all proceeding partial derivaitves

$$\frac{\partial}{\partial c_n}[\vec{c}^T(A\vec{c})] = c_1a_{1n} + c_2a_{2n} + c_3a_{3n} + \ldots + c_1a_{n1} + c_2a_{n2} + \ldots + 2c_na_{nn}$$

$$= c_1(a_{1n} + a_{n1}) + c_2(a_{2n} + a_{n2}) + \ldots + 2c_na_{nn}$$

$$\frac{\partial}{\partial \vec{c}}[\vec{c}^T(A\vec{c})] = \begin{bmatrix} 2c_1a_{11} + c_2(a_{12} + a_{21}) + c_3(a_{13} + a_{31}) + \ldots + c_n(a_{1n} + a_{n1}) \\ c_1(a_{21} + a_{12}) + 2c_2a_{22} + c_3(a_{32} + a_{23}) + \ldots + c_n(a_{2n} + a_{n2}) \\ \vdots \\ c_1(a_{n1} + a_{1n}) + c_2(a_{n2} + a_{2n}) + c_3(a_{n3} + a_{3n}) + \ldots + 2c_na_{nn} \end{bmatrix}$$

□

(b) [easy] Given matrix $X \in \mathbb{R}^{n \times (p+1)}$, full rank and first column consisting of the $\mathbf{1}_n$ vector, rederive the least squares solution $\boldsymbol{b}$ (the vector of coefficients in the linear model shipped in the prediction function g). No need to rederive the facts about vector derivatives.

Solution.

$$SSE = \sum_i^n (y_i - \hat{y}_i)^2 = (\vec{y} - \vec{\hat{y}})^T(\vec{y} - \vec{\hat{y}}) \qquad (\vec{v}^T\vec{v} = ||\vec{v}||^2)$$

$$= (\vec{y}^T - \vec{\hat{y}}^T)(\vec{y} - \vec{\hat{y}})$$

$$= \vec{y}^T\vec{y} - \vec{y}^T\vec{\hat{y}} - \vec{\hat{y}}^T\vec{y} + \vec{\hat{y}}^T\vec{\hat{y}}$$

$$
\begin{aligned}
&= \vec{y}^T\vec{y} - 2\hat{\vec{y}}^T\vec{y} + \hat{\vec{y}}^T\hat{\vec{y}} && (a^Tb = b^Ta)\\
&= \vec{y}^T\vec{y} - 2(X\vec{w})^T\vec{y} + (X\vec{w})^T(X\vec{w})\\
&= \vec{y}^T\vec{y} - 2\vec{w}^TX^T\vec{y} + \vec{w}^TX^TX\vec{w} && ((ab)^T = b^Ta^T)
\end{aligned}
$$

to minimize the function, we take partials with respect to the vector $\vec{w}$ and set the first derivative to 0_{p+1}:

$$
\begin{aligned}
\frac{\partial}{\partial\vec{w}}[SSE] &= \frac{\partial}{\partial\vec{w}}[\vec{y}^T\vec{y} - 2\vec{w}^TX^T\vec{y} + \vec{w}^TX^TX\vec{w}]\\
&= \vec{0}_{p+1} - 2X^T\vec{y} + 2X^TX\vec{w} = \vec{0}_{p+1}\\
&\Rightarrow (X^TX)^{-1}X^TX\vec{w} = (X^TX)^{-1}X^T\vec{y}
\end{aligned}
$$

$$
\boxed{\vec{b} = (X^TX)^{-1}X^T\vec{y}}
$$

□

(c) [harder] Consider the case where $p = 1$. Show that the solution for $\boldsymbol{b}$ you just derived is the same solution that we proved for simple regression in Lecture 8. That is, the first element of $\boldsymbol{b}$ is the same as $b_0 = \bar{y} - r\frac{s_y}{s_x}\bar{x}$ and the second element of $\boldsymbol{b}$ is $b_1 = r\frac{s_y}{s_x}$.

Solution. If $p = 1$ then $X \in \mathbb{R}^{n\times 2}$ $(p + 1 = 2)$. Calculating X^TX

$$
X^TX = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n {x_i}^2 \end{bmatrix}_{2\times 2}
$$

However, to compute $\vec{b}$ we need the inverse of the matrix. Fortunately, this matrix is a 2 by 2 matrix which makes it easier to compute the determinant:

$$
\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc}\begin{bmatrix} a & b \\ c & d \end{bmatrix}
$$

$$
(X^TX)^{-1} = \frac{1}{n\sum_{i=1}^n {x_i}^2 - n^2\bar{x}^2}\begin{bmatrix} \sum_{i=1}^n {x_i}^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix}
$$

Continuing on :

$$
(X^TX)^{-1}X^T\vec{y} = \frac{1}{n\sum_{i=1}^n {x_i}^2 - n^2\bar{x}^2}\begin{bmatrix} \sum_{i=1}^n {x_i}^2 & -n\bar{x} \\ -n\bar{x} & n \end{bmatrix}\begin{bmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_n \end{bmatrix}\vec{y}
$$

$$
= \frac{1}{n\sum_{i=1}^n {x_i}^2 - n^2\bar{x}^2}\begin{bmatrix} \sum_{i=1}^n {x_i}^2 - nx_1\bar{x} & \dots & \sum_{i=1}^n {x_i}^2 - nx_n\bar{x} \\ -n\bar{x} + x_1n & \dots & -n\bar{x} - nx_n \end{bmatrix}_{2\times n}\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1}
$$

$$= \frac{1}{n\sum_{i=1}^n {x_i}^2 - n^2\bar{x}^2} \begin{bmatrix} y_1(\sum_{i=1}^n {x_i}^2 - nx_1\bar{x}) & \dots & y_n(\sum_{i=1}^n {x_i}^2 - nx_n\bar{x}) \\ y_1(-n\bar{x} + x_1 n) & \dots & y_n(-n\bar{x} - nx_n) \end{bmatrix}$$
$$= \frac{1}{n\sum_{i=1}^n {x_i}^2 - n^2\bar{x}^2} \begin{bmatrix} (\sum_{i=1}^n y_i)(\sum_{i=1}^n {x_i}^2) - n\bar{x}\sum_{i=1}^n x_i y_i \\ n\bar{x}\sum_{i=1}^n y_i + n\sum_{i=1}^n x_i y_i \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

Focusing on just b_1, we get

$$\begin{aligned} b_1 &= \frac{1}{n\sum_{i=1}^n {x_i}^2 - n^2\bar{x}^2} \left(n\bar{x}\sum_{i=1}^n y_i + n\sum_{i=1}^n x_i y_i \right) \\ &= \frac{1}{\cancel{n}\sum_{i=1}^n {x_i}^2 - n\cancel{^2}^{1}\bar{x}^2} \left(\cancel{n}\bar{x}n\bar{y} + \cancel{n}\sum_{i=1}^n x_i y_i \right) \\ &= \frac{1}{\sum_{i=1}^n {x_i}^2 - n\bar{x}^2} \left(\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} \right) \end{aligned}$$

Note I used the fact $\sum_{i=1}^n y_i = n\bar{y}$. The expression obtained is that derived in the ordinary least squares algorithm which represents $r\frac{S_y}{S_x}$. Hence $b_1 = r\frac{S_y}{S_x}$.Focusing on b_0,

$$\begin{aligned} b_0 &= \frac{1}{n\sum_{i=1}^n {x_i}^2 - n^2\bar{x}^2} \left((\sum_{i=1}^n y_i)(\sum_{i=1}^n {x_i}^2) - n\bar{x}\sum_{i=1}^n x_i y_i \right) \\ &= \frac{1}{\cancel{n}\sum_{i=1}^n {x_i}^2 - n\cancel{^2}^{1}\bar{x}^2} \left((\cancel{n}\bar{y})(\sum_{i=1}^n {x_i}^2) - \cancel{n}\bar{x}\sum_{i=1}^n x_i y_i \right) \\ &= \frac{1}{\sum_{i=1}^n {x_i}^2 - n\bar{x}^2} \left(\bar{y}\sum_{i=1}^n {x_i}^2 - \bar{x}\sum_{i=1}^n x_i y_i + n\bar{x}^2\bar{y} - n\bar{x}^2\bar{y} \right) \\ &= \bar{y}\frac{n\sum_{i=1}^n {x_i}^2 - \cancel{n^2\bar{x}^2}^{1}}{\sum_{i=1}^n {x_i}^2 - n\bar{x}^2} - \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n {x_i}^2 - n\bar{x}^2}\bar{x} \\ &= \bar{y} - \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n {x_i}^2 - n\bar{x}^2} \\ &= \bar{y} - b_1\bar{x} \end{aligned}$$

□

(d) [easy] If X is rank deficient, how can you solve for $\boldsymbol{b}$? Explain in English.

Solution. If X is rank deficient, then we can't apply algorithm to solve for b because there wouldn't be an inverse of X to even use (by the linear algebra equivalent statements). We can, however, go solve for b by modifying X. We can delete columns of X until it is full rank, thereby allowing us to use the derived expression for b. One may argue why we should go about deleting information, but if the columns we are deleting are linearly dependent then we are really just deleting excessive information that isn't necessary for our mdoel. □

(e) [difficult] Prove $\text{rank}\,[X] = \text{rank}\left[X^\top X\right]$.

Proof. Let $X \in \mathbb{R}^{n\times(p+1)}$. Hence both X and X^TX have the same number of columns. In order to use the algorithm, X^TX must be invertible, hence it has full rank. On the contrary, assume $\text{rank}\,[X] \neq \text{rank}\left[X^TX\right]$ implying that $\text{rank}\,[X] < \text{rank}\left[X^TX\right] = p+1$. Therefore, there exists a vector $\vec{u} \neq \vec{0}_{p+1} \in \mathbb{R}^{p+1}$ such that

$$X\vec{u} = \vec{0}_n$$

This vector $\vec{u}$ can then be applied to X^TX

$$X^TX\vec{u} = X^T(X\vec{u}) = X^T0_n = 0_{p+1}$$

which means X^TX is not full rank, a contradiction. Thus, both ranks of X and of X^TX must be equal to each other by contradiction. □

(f) [difficult] Given matrix $X \in \mathbb{R}^{n\times(p+1)}$, full rank and first column consisting of the $\mathbf{1}_n$ vector, now consider cost multiples ("weights") $c_1, c_2, \ldots, c_n$ for each mistake e_i. As an example, previously the mistake for the 17th observation was $e_{17} := y_{17} - \hat{y}_{17}$ but now it would be $e_{17} := c_{17}(y_{17} - \hat{y}_{17})$. Derive the weighted least squares solution $\boldsymbol{b}$. No need to rederive the facts about vector derivatives. Hints: (1) show that SSE is a quadratic form with the matrix C in the middle (2) Split this matrix up into two pieces i.e. $C = C^{\frac{1}{2}}C^{\frac{1}{2}}$, distribute and then foil (3) note that a scalar value equals its own transpose and (4) use the vector derivative formulas.

(g) [difficult] If $p = 1$, prove $r^2 = R^2$ i.e. the linear correlation is the same as proportion of sample variance explained in a least squares linear model.

Proof. Considering $p = 1$, the degrees of freedom become restricted such that $S_y^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2$. A similar thing happens to S_x. $R^2 = SSR/SST$ where

$$SSR = \sum_{i=1}^n (\hat{y} - \bar{y})^2 = \sum_{i=1}^n \hat{y}_i^2 - n\bar{y}$$

(derived in class, due to the fact $\sum_{i=1}^n \hat{y}_i = \vec{\hat{y}}^T\vec{1} = (H\vec{y})^T\vec{1} = \vec{y}^TH^T\vec{1} = \vec{y}^T = \sum_{i=1}^n y_i$) and $SST = \sum_{i=1}^n (y_i - \bar{y})^2$. Recall that $\hat{y}_i = b_0 + b_1x_i$ and that the solutions of $\vec{b}$ in the $p = 1$ case are $b_0 = \bar{y} - b_1\bar{x}$ and $b_1 = r\frac{S_y}{S_x}$. Then

$$\begin{aligned} R^2 = \frac{\sum_{i=1}^n \hat{y}_i^2 - n\bar{y}}{\sum_{i=1}^n (y_i - \bar{y})^2} &= \frac{\sum_{i=1}^n (b_0 + b_1x_i)^2 - \bar{y}^2n}{(n-1)S_y^2} \\ &= \frac{\sum_{i=1}^n b_0^2 + b_1x_i^2 + 2b_0b_1x_i - \bar{y}^2n}{(n-1)S_y^2} \\ &= \frac{b_0^2n + b_1^2\sum_{i=1}^n {x_i}^2 + 2b_0b_1\sum_{i=1}^n x_i - \bar{y}^2n}{(n-1)S_y^2} \end{aligned}$$

$$
\begin{aligned}
&= \frac{(\bar{y} - b_1\bar{x})^2 n + b_1^2 \sum_{i=1}^n {x_i}^2 + 2b_1(\bar{y} - b_1\bar{x})\bar{x}n - \bar{y}^2 n}{(n-1)S_y^2} \\
&= \frac{\cancel{\bar{y}^2 n} + \cancel{b_1^2 n\bar{x}^2} - \cancel{2\bar{y}b_1\bar{x}n} + b_1^2 \sum_{i=1}^n {x_i}^2 + \cancel{2b_1\bar{y}\bar{x}n} - \overset{1}{\cancel{2}} b_1^2\bar{x}^2 n - \cancel{\bar{y}^2 n}}{(n-1)S_y^2} \\
&= \frac{b_1^2 \sum_{i=1}^n {x_i}^2 - b_1^2\bar{x}^2 n}{(n-1)S_y^2} \\
&= \frac{b_1^2(\sum_{i=1}^n {x_i}^2 - n\bar{x}^2)}{(n-1)S_y^2} \\
&= \frac{r^2 \frac{\cancel{S_y^2}}{S_x^2} \sum_{i=1}^n (x_i - \bar{x})^2}{(n-1)\cancel{S_y^2}} \\
&= \frac{r^2 \cancel{(n-1)S_x^2}}{\cancel{S_x^2(n-1)}} \\
&= r^2
\end{aligned}
$$

□

(h) [harder] Prove that the point $< 1, \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p, \bar{y} >$ is a point on the least squares linear solution.

Proof. For the least squares case, we know $\sum_{i=1}^n \hat{y}_i = \sum_{i=1}^n y_i$. Yet we know

$$y_i = b_0 + b_1 x_{i1} + \ldots + b_p x_{ip}$$

Therefore, we can simply plug in two the equation for least squares:

$$
\begin{aligned}
\sum_{i=1}^n y_i &= \sum_{i=1}^n b_0 + b_1 x_1 + \ldots + b_p x_p \\
&= \sum_{i=1}^n b_0 + b_1 \sum_{i=1}^n x_{i1} + \ldots + b_p \sum_{i=1}^n x_{p1} \\
&= nb_0 + b_1 n\bar{x}_1 + \ldots + b_p n\bar{x}_n
\end{aligned}
$$

Dividing by n on both sides results in

$$\frac{1}{\cancel{n}}(\cancel{n}b_0 + b_1\cancel{n}\bar{x}_1 + \ldots + b_p\cancel{n}\bar{x}_n) = \frac{1}{n}\sum_{i=1}^n \hat{y}_i = \frac{1}{n}\sum_{i=1}^n y_i = \bar{y}$$

Thus,

$$\Rightarrow b_0 + b_1\bar{x}_1 + \ldots + b_p + \bar{x}_n = \bar{y}$$

implying that the point $< 1, \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p, \bar{y} >$ is a point on the least squares linear solution. □

Problem 5

These are questions related to the concept of orthogonal projection, QR decomposition and its relationship with least squares linear modeling.

(a) [easy] Consider least squares linear regression using a design matrix X with rank $p+1$. What are the degrees of freedom in the resulting model? What does this mean?

Solution. The degrees of freedom in the resulting model are $p+1$ which means that the X training data has $p+1$ linearly independent columns, meaning the columns of predictors cannot be represented as a non-zero multiple of another column and thereby are not repeating information in the X matrix. □

(b) [harder] If you are orthogonally projecting the vector $\boldsymbol{y}$ onto the column space of X which is of rank $p+1$, derive the formula for $\text{Proj}_{\text{colsp}[X]}[\boldsymbol{y}]$. Is this the same as the least squares solution?

Solution. The space of X is $X = [\vec{x}_1|\vec{x}_2|\ldots|\vec{x}_{p+1}]$. Then the project onto the column space is

$$\begin{aligned}\text{Proj}_{\text{colsp}[X]}[\vec{y}] &= \text{Proj}_{\vec{x}_1}[\vec{y}] + \text{Proj}_{\vec{x}_2}[\vec{y}] \ldots + \text{Proj}_{\vec{x}_{p+1}}[\vec{y}] = \sum_{i=1}^{p+1} \text{Proj}_{\vec{x}_j}[\vec{y}] \\ &= w_1\vec{x}_i + w_2\vec{x}_2 + \ldots + w_{p+1}\vec{x}_{p+1} = X\vec{w}\end{aligned}$$

For some vector $\vec{w}$. From this, the residuals $\boldsymbol{e}$ is defined as the $\vec{y}$ minus the projection, and is completly orthogonal to the column space of X. Therefore

$$\boldsymbol{e} = \vec{y} - \text{Proj}_{\text{colsp}[X]}[\vec{y}] = \vec{y} - X\vec{w} \tag{5.b.1}$$

$$\boldsymbol{e} \cdot \vec{x}_1 = 0, \quad \boldsymbol{e} \cdot \vec{x}_2 = 0, \ldots, \boldsymbol{e} \cdot \vec{x}_{p+1} = 0 \tag{5.b.2}$$

Plugging in (5.b.1) into the equations of (5.b.2) produce

$$\vec{x}_1^T(\hat{y} - X\vec{w}) = 0, \quad \vec{x}_2^T(\hat{y} - X\vec{w}) = 0, \ldots, \vec{x}_{p+1}^T(\hat{y} - X\vec{w}) = 0 \tag{5.b.3}$$

But each equation of (5.b.3) is really just the entry of the matrix multiplication $X^T(\vec{y} - X\vec{w}) = \vec{0}$. Thus we can focus on the equation

$$\begin{aligned}&X^T(\vec{y} - X\vec{w}) = \vec{0} \\ &= X^T\hat{y} - X^TX\vec{w} = \vec{0} \\ &\Rightarrow (X^TX)^{-1}(X^T\hat{y} = X^TX\vec{w}) \\ &\Rightarrow \vec{w} = (X^TX)^{-1}X^T\vec{y}\end{aligned}$$

The projection is therefore

$$\text{Proj}_{\text{colsp}[X]}\,[\vec{y}] = X\vec{w} = X(X^TX)^{-1}X^T\vec{y}$$

which is the same as the least squares solution for the linear squares linear modeling. □

(c) [difficult] We saw that the perceptron is an *iterative algorithm*. This means that it goes through multiple iterations in order to converge to a closer and closer $\boldsymbol{w}$. Why not do the same with linear least squares regression? Consider the following. Regress $\boldsymbol{y}$ using $\boldsymbol{X}$ to get $\hat{\boldsymbol{y}}$. This generates residuals $\boldsymbol{e}$ (the leftover piece of $\boldsymbol{y}$ that wasn't explained by the regression's fit, $\hat{\boldsymbol{y}}$). Now try again! Regress $\boldsymbol{e}$ using $\boldsymbol{X}$ and then get new residuals $\boldsymbol{e}_{new}$. Would $\boldsymbol{e}_{new}$ be closer to $\boldsymbol{0}_n$ than the first $\boldsymbol{e}$? That is, wouldn't this yield a better model on iteration #2? Yes/no and explain.

Solution. Regressing $\boldsymbol{y}$ using X in the linear least squares regression would utlize the hat matrix H composed of X, i.e. $H = X(X^T)^-1X^T$. The hat matrix would project y onto the column space of X which would be $\hat{y}$ and the difference between $\boldsymbol{y}$ and $\hat{y}$ would be the residuals $\boldsymbol{e}$. This $\boldsymbol{e}$ vector would be orthogonal to the column space of X, so any second iteration of the hat matrix onto the column space of X would simply produce the same predictions as before. Projecting $\boldsymbol{e}$ onto the column space of X would be zero since the vector is othorgonal to the space it is being projected on. Since this is the least squares regression, these $\vec{b}$ would already be the minimum matrix of coefficients that can exist. Any other answer would not produce a value smaller than a minimum and thus does not warrant any further concern. □

(d) [harder] Prove that $Q^\top = Q^{-1}$ where Q is an orthonormal matrix such that $\text{colsp}\,[Q] = \text{colsp}\,[X]$ and Q and X are both matrices $\in \mathbb{R}^{n\times(p+1)}$. Hint: this is purely a linear algebra exercise.

Proof. If we can prove that $Q^TQ = I_{p+1}$ then that is sufficient to show $Q^T = Q^{-1}$; by definition of the inverse, whatever matrix that multiplies to the matrix of consideration to produce the identity is the inverse. However, note that Q is an orthonormal matrix. Since the columns of Q are of length 1 and orthogonal to each other, the matrix multiplication would simply produce a 1 for every entry that is a result of every column being multiplied by its transpose and a 0 otherwise:

$$\begin{bmatrix} \leftarrow q_1^T \rightarrow \\ \leftarrow q_2^T \rightarrow \\ \vdots \\ \leftarrow q_{p+1}^T \rightarrow \end{bmatrix} \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ q_1 & q_2 & \cdots & q_{p+1} \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

The matrix produced is the identity and therefore $Q^T = Q^{-1}$. □

(e) [harder] Prove that the least squares projection $H = X\left(X^\top X\right)^{-1} X^\top$ is the same as $QQ^\top$.

Proof. Let $X = QR$ for QR where Q is an orthonormal matrix where Q is an orthonormal matrix such that $\text{colsp}\,[Q] = \text{colsp}\,[X]$ and Q and X are both matrices $\in \mathbb{R}^{n\times(p+1)}$. Then

$$\begin{aligned} X(X^TX)^{-1}X^T &= (QR)[(QR)^T(QR)]^{-1}(QR)^T \\ &= (QR)[R^TQ^T(QR)]^{-1}(QR)^T \end{aligned}$$

Note that $Q^TQ = I_{p+1}$ due the orthonormal property of Q (see above problem). Continuing with the proof:

$$\begin{aligned} &= QR(R^TIR)^{-1}R^TQ^T \\ &= QRR^{-1}(R^T)^{-1}Q^T \\ &= QQ^T \end{aligned}$$

□

(f) [harder] Prove that an orthogonal projection onto the $\text{colsp}\,[Q]$ is the same as the sum of the projections onto each column of Q.

(g) [difficult] Trouble in paradise. Prove that the SSE of a multivariate linear least squares model always decreases (equivalently, R^2 always increases) upon the addition of a new independent predictor. Keep in mind this holds true even if this new predictor has no information about the true causal inputs to the phenomenon y.

Proof. The project of y can be defined in the following way for the multivariate linear least squares model:

$$\vec{\hat{y}} = H\vec{y} = QQ^T\vec{y} = \sum_{j=1}^{p+1} \text{Proj}_{q_j}\,[\vec{y}]$$

and the sum of squared regression

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2 = ||\vec{\hat{y}}||^2 = \sum_{j=1}^{p+1} ||\text{Proj}_{q_j}\,[\vec{y}]\,||^2$$

Recall the relation $SST = SSR + SSE$ where SST is the sum of squared total and SSE is the sum of squared errors. Equivalently, $SSE = SST - SSR$. Now let the columnspace of X grow such that there is a new predictor: $X_{new} = [X|\vec{x}_{new}]$. Hence

$$\vec{\hat{y}}_{new} = H_{new}\vec{y} = Q_{new}Q_{new}^T\vec{y} = \sum_{j=1}^{p+1} \text{Proj}_{q_j}\,[\vec{y}] + \text{Proj}_{q_{new}}\,[\vec{y}]$$

and new the sum of squared regression

$$SSR_{new} = ||\vec{\hat{y}}_{new}||^2 = \sum_{j=1}^{p+1} ||\text{Proj}_{q_j}[\vec{y}]||^2 + ||\text{Proj}_{q_{new}}[\vec{y}]||^2 \geq SSR$$

SST will remain the same since it is not dependent on the $x's$. Then $SST - SSR = SSE \geq SST - SSR_{new} = SSE_{new}$. It is very rare to obtain the projection of $\vec{y}$ onto the new $\vec{q}_{new}$ that is 0, meaning that $\vec{y}$ and the column space are orthogonal. Consequently, the inequality $SSE \geq SSE_{new}$ turns to $SSE > SSE_{new}$ a majority of the time. Therefore the addition of new predictors will cause the SSE to decrease and the R^2 will increase. □

(h) [harder] Why is this a bad thing? Explain in English.

Solution. This is bad thing because these added predictors could be any piece of irrelevant information and the SSE would still decrease and the R^2 would increase, giving the impression that the model is becoming more accurate and producing better results. The reality is, however, that there has been no improvement to the predictablity of the model and there has only been an addition of unnecessary information. □

(i) [E.C.] Prove that $\text{rank}[H] = \text{tr}[H]$.