MATH 390.4 / 650.2 Spring 2018 Homework #3t

Chaim Eisenbach

Thursday 22$^{\text{nd}}$ 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.

For a fox there are multiple $x's$ that are closer to our $z's$. Also, they reduce our errors by having a larger data set and a more nuanced algorithm. As opposed to hedgehogs whose $x's$ may not be as close to our $z's$. Also Hedgehogs stick with one $\mathcal{A}$ while Fox's can adjust their $\mathcal{A}$ depending on new information to get a better g.

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

He wanted definite answers to issues. This can be seen in a lot of his policy moves. Such as dropping the atom bombs, the containment plan, the Korean war, union strikes etc. He wasn't interested in a less definitive more nuanced opinion.
Many people thing this way. It is easier to think of something in concrete terms than in more nuanced ones. BIG BOLD PREDICTIONS

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

They have more opportunities to permute and manipulate the information they have to confirm their biases.

(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.

Returning a probability is better than returning a definitive answer. We want a range of possible outcomes, which would be a more honest expression of the uncertainty found in the real world. It would be foolish to pin things down to an exact number.

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"?

Y: output space.

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

There are business requirements and constraints that need to be taken into account. Sacrifice a small amount of predictive accuracy to ensure that business requirements are met. In a real business environment the bottom line comes before perfection.

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

Weighted scores, they used a decision tree which is a typed of classification model.

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

A regression, profits are being measured.

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

cultural norms. Organizations have a set way of doing things. Which is why Finlay prefers "embedded analytics" over an "analytics culture".

(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?

The model to predict these didn't take into account how rare foreclosure actually is. They ignored the culture. They couldn't differentiate between a good model and good business.

(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.).

With human judgment it is easier to get closer to f, to develop a better algorithm. To give greater weight to more important $x's$.

(h) [difficult] In Chapter 4 Finlay 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.

Even though according to Silver it is important to look at many different causes for an event to make better and more accurate predictions. Finlay is telling us that once you surpass a certain number of variables you no longer have information useful enough to improve your algorithm or whatever you are trying to do. There is a limited number of really important factors in determining the cause of something.

(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?

Pre-sifted data save the time to find your own good data among all the useless data. Predictive analytics is good business. An investment in good data can produce a more profitable business model. Having better data leads to a more accurate model.

(j) [easy] Table 4 lists "data types" about what type of observations?
Behavior of interest. To predict burglary.

(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?

Data about primary behavior. It predicts a certain pattern about ones behavior.

(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 fruitful for predictive ability?

It will provide only incremental benefit to the model. If we know one spouses age we already have a pretty good picture of the range of the other spouses age. "If new data is highly correlated with existing data then it won't add much to the power of your predictions".

(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.).

Eventually reality veers away from our model. So our $x's$ are further away from our $z's$ which are different from the initial $z's$. Also, our g strays away from f, maybe f leaves our $\mathcal{H}$ entirely.

(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.).

We need a $\mathbb{D}$ that encompasses enough X and Y to become closer to t. Essentially enough to get to a $\mathcal{H}$ that encompasses an h^* that is close enough to an f that is close enough to t.

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

According to Finlay data is only important if it's good data. Storage has become so cheap that there is a lot of useless data that needs to be sifted through.

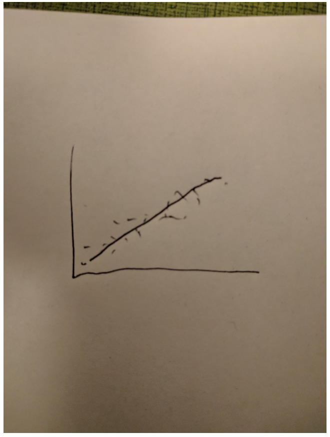
(p) [easy] What is Finlay's solution to "overfitting" (p84)?

It will be less likely to occur if there are large samples being used.

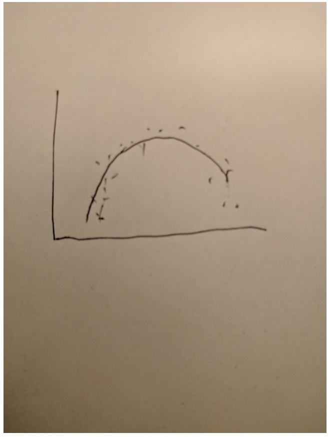
Problem 3

These are questions about association and correlation.

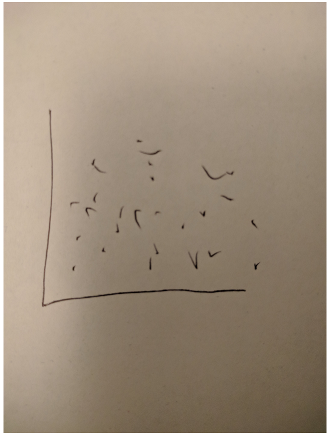
(a) [easy] Give an example of two variables that are both correlated and associated by drawing a plot.



(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.

No. correlation ∈ association.

Problem 4

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

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

Looking at: $\boldsymbol{c}^\top A\boldsymbol{c} = c_1(c_{11} + c_2a_{21}....c_na_{n1}) + c_2(ca_{21} + ... + ca_{2n}) + c_n.....$

$\Rightarrow \sum_{i=1}^{n} c_i \sum_{j=1}^{n} c_ja_{ij}$

$\frac{\partial}{\partial \boldsymbol{c}_i}\left[\boldsymbol{c}^\top A\boldsymbol{c}\right] = \sum_{j=1}^{n} c_ja_{ij} + c_ja_{ji}$

(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.

$$\frac{\partial}{\partial \boldsymbol{w}}\left[\vec{y}^\top\vec{y} - 2\vec{w}^\top x^\top\vec{y} + \vec{w}^\top(x^\top x)\vec{w}\right] = 0_{p+1}^\top \Rightarrow (x^\top x)^{-1}x^\top x\vec{w} = (x^\top x)^{-1}x^\top\vec{y}$$

$\vec{b} = (x^\top x)^{-1}x^\top\vec{y}$

(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 $b_1 = r\frac{s_y}{s_x}$.

$$(X^\top X)^{-1}X^\top\vec{y} = \tfrac{1}{n\sum x_i^2 - n^2\bar{x}^2}\underbrace{\begin{pmatrix}\sum x_i^2 & -n\bar{x}\\ -n\bar{x} & n\end{pmatrix}}_{(X^\top X)^{-1}} * \underbrace{\begin{pmatrix}1 & . & . & .1 & \\ x_1 & . & . & . & x_n\end{pmatrix}}_{X^\top}\underbrace{\begin{pmatrix}y_1\\ .\\ .\\ .\\ y_n\end{pmatrix}}_{\vec{y}}$$

$$= \tfrac{1}{n\sum x_i^2 - n^2\bar{x}^2}\underbrace{\begin{pmatrix}\sum x_i^2 & -n\bar{x}\\ -n\bar{x} & n\end{pmatrix}}_{(X^\top X)^{-1}}\underbrace{\begin{pmatrix}n\bar{y}\\ \sum y_i x_i\end{pmatrix}}_{X^\top\vec{y}} = \tfrac{1}{n\sum x_i^2 - n^2\bar{x}^2}\begin{pmatrix}n\bar{y}\sum x_i^2 & -n\sum y_i x_i\\ -n^2\bar{x}\bar{y} & +n\sum y_i x_i\end{pmatrix}$$

$$b_0 = \tfrac{\bar{y}(\sum x_i^2 - n\bar{x}^2) - \bar{x}(\sum y_i x_i - n\bar{x}\bar{y})}{\sum x_i^2 - n\bar{x}^2}$$

$$b_1 = \tfrac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

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

Eliminate the linearly dependent vectors.

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

Rank-Nullity Theorem: rank $[X] = p + 1 - \dim(null(x)) = p + 1 - \dim(null(X^\top X)) =$ rank $\left[X^\top X\right]$

(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.

$\sum c_i(y_i - x_i w)^2$
$(y - xw)^\top c(y - xw)$
$(y^\top cy - y^\top cxw - w^\top x^\top cy + w^\top x^\top cxw)$
taking the derivative we get $2(-x^\top cy + x^\top cxb) = 0$
$b = (x^\top cx)^{-1}x^\top cy$

(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.

$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2} \Rightarrow \frac{(\sum(x-\bar{x})(y-\bar{y}))^2}{(\sum(x-\bar{x}))^2(\sum(y-\bar{y}))^2} \Rightarrow \frac{Cov(x,y)^2}{S_x^2 S_y^2} = r^2$

(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.

$\bar{y} = \frac{1}{n}\sum b_0 + b_1 x_{i1} + \ldots + b_n x_i n$
$\bar{y} = b_0 + b_1\bar{x}_n + \ldots + b_n\bar{x}_n$

So the line goes through the averages of x and y.

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?

$p+1$ the degrees of freedom because it is the dimension of the column space of X. It's the independent pieces of data.

(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?

$\mathcal{X} = \text{colsp}[x]$
$\text{Proj}_{\mathcal{X}}[\vec{\boldsymbol{y}}] = X\vec{w}$
$X^\top(X\vec{w} - \vec{y}) = 0 \Leftarrow$ Because perpendicular.
$X^\top X\vec{w} = X^\top\vec{y}$
$\vec{w} = (X^\top X)^{-1}X^\top\vec{y}$
$X\vec{w} = X(X^\top X^{-1})X^\top\vec{y}$
$\text{Proj}_{\mathcal{X}}[\boldsymbol{y}] = X(X^\top X)^{-1}X^\top\vec{y}$

(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.

No, because the projection is the minimum error.

(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.

$$Q^\top Q = \begin{pmatrix} q_1^\top \\ . \\ . \\ . \\ q_n^\top \end{pmatrix} \begin{pmatrix} q_1 & . & . & . & q_n \end{pmatrix} = I_{p+1}$$

$Q^\top Q = I$
So $Q^\top = Q^{-1}$

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

$X \in \mathbb{R}^{n \times n}$
$X = QR$, $Q^\top Q = I_n$, $R \in \mathbb{R}^{n \times n}$
$(X^\top X)^{-1} X^\top = (R^\top Q^\top Q R)^{-1} R^\top Q^\top = R^{-1} Q^\top$
$H = X(X^\top X)^{-1} X^\top = X R^{-1} Q^\top = Q Q^\top$

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

$$\text{Proj}_Q\left[\vec{a}\right] = Q(Q^\top Q)^{-1} Q^\top \vec{a} = \sum_{j=1}^{k} \frac{\vec{q_j}\vec{q_j}^\top}{||\vec{q_j}||^2} = \sum_{j=1}^{k} \vec{q_j}\vec{q_j}^\top \vec{a} = Q Q^\top \vec{a}$$

(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.

As SSR $\uparrow \Rightarrow$ SSE $\downarrow$ because SST $=$ SSR $+$ SSE
Adding a new piece, means rank is now $p + 1 + 1$
$\text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum y_i^2 - n\bar{y^2} = \sum_{j=1}^{p+1} \left|\left|\text{Proj}_{qj}\left[\vec{y}\right]\right|\right|^2 + \left|\left|\text{Proj}_{q_{new}}\left[\vec{y}\right]\right|\right|^2$
$\Rightarrow SSR_{new} \geq SSR \Rightarrow SSE_{new} \leq SSE \Rightarrow R^2_{new} \geq R^2$

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

R^2 can be made to look very good by adding more parameters without our model actually being better.

(i) [E.C.] Prove that rank $[H] =$ tr $[H]$.
tr $[H] =$ tr $\left[X(X^\top X)^{-1} X^{-1}\right] =$ tr $\left[X^\top X (X^\top X)^{-1}\right] =$ tr $[I_{p+1}] = p + 1$