MATH 390.4 / 650.2 Spring 2018 Homework #3t

Professor Adam Kapelner

Due 11:59PM Friday, March 23, 2018 under the door of KY604

(this document last updated Thursday 15th March, 2018 at 6:10pm)

Instructions and Philosophy

The path to success in this class is to do many problems. Unlike other courses, exclusively doing reading(s) will not help. Coming to lecture is akin to watching workout videos; thinking about and solving problems on your own is the actual "working out." Feel free to "work out" with others; **I want you to work on this in groups.**

Reading is still *required*. For this homework set, read about all the concepts introduced in class online e.g. multivariate least squares linear modeling, orthogonal projections, QR decomposition, etc. This is your responsibility to supplement in-class with your own readings. Also, read ch 2-4 in Silver.

The problems below are color coded: green problems are considered *easy* and marked "[easy]"; yellow problems are considered *intermediate* and marked "[harder]", red problems are considered *difficult* and marked "[difficult]" and purple problems are extra credit. The *easy* problems are intended to be "giveaways" if you went to class. Do as much as you can of the others; I expect you to at least attempt the *difficult* problems.

This homework is worth 100 points but the point distribution will not be determined until after the due date. See syllabus for the policy on late homework.

Up to 10 points are given as a bonus if the homework is typed using LaTeX. Links to instaling LaTeX and program for compiling LaTeX is found on the syllabus. You are encouraged to use `overleaf.com`. If you are handing in homework this way, read the comments in the code; there are two lines to comment out and you should replace my name with yours and write your section. The easiest way to use overleaf is to copy the raw text from hwxx.tex *and* preamble.tex into two new overleaf tex files with the same name. If you are asked to make drawings, you can take a picture of your handwritten drawing and insert them as figures or leave space using the "\vspace" command and draw them in after printing or attach them stapled.

The document is available with spaces for you to write your answers. If not using LaTeX, print this document and write in your answers. I do not accept homeworks which are *not* on this printout. Keep this first page printed for your records.

NAME: ____

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.

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

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

(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.

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"?

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

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

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

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

(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?

(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.).

(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.

(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?

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

(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?

(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?

(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^*, δ, ϵ, 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.).

(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^*, δ, ϵ, 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.).

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

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

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.

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.

(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.

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

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

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

(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.

(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.

(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.

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?

(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?

(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.

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

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

(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.

(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.

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

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