MATH 390.4 / 650.2 Spring 2020 Homework #3

Professor Adam Kapelner

Due noon Wedndesday, March 11, 2020 under the door of KY604

(this document last updated 2:31pm on Friday 28th February, 2020)

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 Chapters 3-6 of Silver's book. You should be googling and reading about all the concepts introduced in class online. This is your responsibility to supplement in-class with *your own* readings.

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 7 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 *including this first page* and write in your answers. I do not accept homeworks which are *not* on this printout.

NAME: _____

Problem 1

These are questions about Silver's book, chapters ... For all parts in this question, answer using notation from class (i.e. $t, f, g, h^*, \delta, \epsilon, e, t, z_1, \ldots, z_t, \mathbb{D}, \mathcal{H}, \mathcal{A}, \mathcal{X}, \mathcal{Y}, X, y, n, p, x_{\cdot 1}, \ldots, x_{\cdot p}$, $x_{1\cdot}, \ldots, x_{n\cdot}$, etc. and also we now have $f_{pr}, h^*_{pr}, g_{pr}, p_{th}$, etc from probabilistic classification as well as different types of validation schemes).

(a) [easy] What algorithm that we studied in class is PECOTA most similar to?

(b) [easy] Is baseball performance as a function of age a linear model? Discuss.

(c) [harder] How can baseball scouts do better than a prediction system like PECOTA?

(d) [harder] Why hasn't anyone (at the time of the writing of Silver's book) taken advantage of Pitch f/x data to predict future success?

(e) [difficult] Chapter 4 is all about predicting weather. Broadly speaking, what is the problem with weather predictions? Make sure you use the framework and notation from class. This is not an easy question and we will discuss in class. Do your best.

(f) [easy] Why does the weatherman lie about the chance of rain? And where should you go if you want honest forecasts?

(g) [difficult] Chapter 5 is all about predicting earthquakes. Broadly speaking, what is the problem with earthquake predictions? It is *not* the same as the problem of predicting weather. Read page 162 a few times. Make sure you use the framework and notation from class.

(h) [easy] Silver has quite a whimsical explanation of overfitting on page 163 but it is really educational! What is the nonsense predictor in the model he describes?

(i) [easy] John von Neumann was credited with saying that "with four parameters I can fit an elephant and with five I can make him wiggle his trunk". What did he mean by that and what is the message to you, the budding data scientist?

(j) [difficult] Chapter 6 is all about predicting unemployment, an index of macroeconomic performance of a country. Broadly speaking, what is the problem with unemployment predictions? It is *not* the same as the problem of predicting weather or earthquakes. Make sure you use the framework and notation from class.

(k) [E.C.] Many times in this chapter Silver says something on the order of "you need to have theories about how things function in order to make good predictions." Do you agree? Discuss.

Problem 2

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.

(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 $g([1\ \bar{x}_1\ \bar{x}_2\ \ldots\ \bar{x}_p]) = \bar{y}$ in OLS.

(i) [harder] Prove that $\bar{e} = 0$ in OLS.

(j) [difficult] If you model $\boldsymbol{y}$ with one categorical nominal variable that has levels A, B, C, prove that the OLS estimates look like $\bar{y}_A$ if $x = A$, $\bar{y}_B$ if $x = B$ and $\bar{y}_C$ if $x = C$. You can choose to use an intercept or not. Likely without is easier.

Problem 3

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

(a) [difficult] [MA] Prove that if a square matrix is both symmetric and idempotent then it must be an orthogonal projection matrix.

(b) [easy] Prove that I_n is an orthogonal projection matrix $\forall n$.

(c) [easy] What subspace does I_n project onto?

(d) [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?

(e) [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 in OLS?

(f) [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 $\mathbf{0}_n$ than the first $\boldsymbol{e}$? That is, wouldn't this yield a better model on iteration #2? Yes/no and explain.

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

(h) [harder] Prove that the least squares projection $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T = \boldsymbol{Q}\boldsymbol{Q}^\top$.

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

(j) [easy] Prove that adding a new column to $\boldsymbol{X}$ results in SST remaining the same.

(k) [difficult] [MA] Prove that $\text{rank}\,[\boldsymbol{H}] = \text{tr}\,[\boldsymbol{H}]$. Hint: you will need to use facts about eigenvalues and the eigendecomposition of projection matrices that we learned in class.

Problem 4

All of these are extra credit. This is for students who want to get a taste of a first year linear model theory class at the graduate level. The prereq to do these problems is Math 368/621. Only attempt these if you have time!

(a) [E.C.] In OLS, $\mathcal{H} = \{\boldsymbol{x}\boldsymbol{w} \ : \ \boldsymbol{w} \in \mathbb{R}^{p+1}\}$. Thus, there is a best function $h^*(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta}$ and $y = h^*(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta} + \mathcal{E}$. Imagine that for all n observations in $\mathbb{D}$, the $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\mathcal{E}}$ where $\boldsymbol{\mathcal{E}} \sim \mathcal{N}_n\left(\mathbf{0}_n, \sigma^2\boldsymbol{I}_n\right)$. Assume σ^2 is known. Show that $\boldsymbol{Y} \sim \mathcal{N}_n\left(X\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n\right)$.

(b) [E.C.] Let $\boldsymbol{B} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, i.e. the r.v. that represents the OLS estimator of which $\boldsymbol{b}$ is one realization which changes based on the realizations of the error-vector r.v. $\boldsymbol{\mathcal{E}}$. Find the distribution of $\boldsymbol{B}$ and once this is done, its expectation and variance-covariance matrix. Do the entries in $\boldsymbol{B}$ have dependence?

(c) [E.C.] Find the distribution of $\hat{\boldsymbol{Y}}$, the vector r.v. of predictions.

(d) [E.C.] Find the distribution of $\boldsymbol{E}$, the vector r.v. of residuals.

(e) [E.C.] Find the distribution of SST.

(f) [E.C.] Find the distribution of SSE.

(g) [E.C.] Find the distribution of SSR.

(h) [E.C.] Find the distribution of R^2.

(i) [E.C.] Let $\boldsymbol{U} \sim \mathcal{N}_n(\mathbf{0}_n, \boldsymbol{I}_n)$ independent of $\boldsymbol{V} \sim \mathcal{N}_n(\mathbf{0}_n, \boldsymbol{I}_n)$. Let θ be the r.v. model of the angle between $\boldsymbol{U}$ and $\boldsymbol{V}$. How is θ distributed?