MATH 342W / 642 / RM 742 Spring 2025 HW #3

Professor Adam Kapelner

Due noon March 19

(this document last updated 11:28am on Thursday 6th March, 2025)

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*. 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 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, chapters 3-6. 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.

(a) [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.

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

(c) [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.

(d) [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?

(e) [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?

(f) [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.

(g) [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 related to the concept of orthogonal projection, QR decomposition and its relationship with least squares linear modeling.

(a) [easy] Let $\boldsymbol{H}$ be the orthogonal projection onto $\text{colsp}\,[\boldsymbol{X}]$ where $\boldsymbol{X}$ is a $n \times (p+1)$ matrix with all columns linearly independent from each other. What is $\text{rank}\,[\boldsymbol{H}]$?

(b) [easy] Simplify $\boldsymbol{HX}$ by substituting for $\boldsymbol{H}$.

(c) [harder] What does your answer from the previous question mean conceptually?

(d) [difficult] Let $\boldsymbol{X}'$ be the matrix of $\boldsymbol{X}$ whose columns are in reverse order meaning that $\boldsymbol{X} = [\mathbf{1}_n \vdots \boldsymbol{x}_{\cdot 1} \vdots \ldots \vdots \boldsymbol{x}_{\cdot p}]$ and $\boldsymbol{X}' = [\boldsymbol{x}_{\cdot p} \vdots \ldots \vdots \boldsymbol{x}_{\cdot 1} \vdots \mathbf{1}_n]$. Show that the projection matrix that projects onto $\text{colsp}\,[X]$ is the same exact projection matrix that projects onto $\text{colsp}\,[X']$.

(e) [easy] Generalize the previous problem by proving that orthogonal projection matrices that project onto any specific subspace are *unique*.

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

(g) [difficult] [MA] Prove the converse of the previous question: that if a square matrix is an orthogonal projection matrix, then it must be both symmetric and idempotent.

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

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

(j) [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?

(k) [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.

(l) [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 $\in \mathbb{R}^{n\times(p+1)}$ and $n = p+1$ in this case to ensure the inverse is defined. Hint: this is purely a linear algebra exercise and it's a one-liner.

(m) [easy] Prove that the least squares projection $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T = \boldsymbol{Q}\boldsymbol{Q}^\top$. Justify each step.

(n) [difficult] [MA] This problem is independent of the others. Let H be an orthogonal projection matrix. 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.

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

(p) [easy] Explain why adding a new column to $\boldsymbol{X}$ results in no change to SST.

(q) [harder] Prove that adding a new column to $\boldsymbol{X}$ results in SSR increasing.

(r) [harder] What is overfitting? Use what you learned in this problem to frame your answer.

(s) [easy] Why are the "in-sample" error metrics R^2 and SSE dishonest? Note: I'm leaving out MSE and RMSE as they attempt to be honest by increasing as p increases due to the denominator.

(t) [easy] How can we provide honest error metrics for R^2, SSE? It may help to draw a picture of the procedure.

(u) [easy] The procedure in the previous question produces highly variable honest error metrics. Can you change the procedure slightly to reduce the variation in the honest error metrics? What is this procedure called and how is it done?

Problem 3

These are some questions related to validation.

(a) [easy] Assume you are doing one train-test split where you build the model on the training set and validate on the test set. What does the constant K control? And what is its tradeoff?

(b) [easy] What problem does K-fold CV try to solve?

(c) [difficult] [MA] Theoretically, how does K-fold CV solve this problem? The Internet is your friend.

Problem 4

These are some questions related to probability estimation modeling. Let X denote the design matrix with n rows and $p+1$ columns (with the first column being $\mathbf{1}_n$ and the other columns being linearly independent predictors) and $\boldsymbol{y}$ is the binary response vector of size n and use this notation throughout your responses.

(a) [easy] What is g_0 if you are modeling probability estimates?

(b) [easy] What is $\mathcal{H}_{pr}$ for the probability estimation algorithm that employs the linear model in the covariates with logistic link function?

(c) [easy] What is $\mathcal{H}_{pr}$ for the probability estimation algorithm that employs the linear model in the covariates with cloglog link function?

(d) [easy] Let $\mathcal{A} : \boldsymbol{b} = \arg\max_{\boldsymbol{w} \in \mathbb{R}^{p+1}} \{\ldots\}$. Derive the expression that replaces the ... which will be a function of $X, \boldsymbol{y}, \boldsymbol{w}, n$. Note: this algorithm fits a "logistic regression".

(e) [easy] Why is logistic regression an example of a "generalized linear model" (glm)?

(f) [easy] Consider $\boldsymbol{x}_*$ to be a new unit. For its prediction, the probability estimate that $y_* = 1$ is 37%, what is the log odds of $y_* = 1$?

(g) [easy] If, $\boldsymbol{x}_* \boldsymbol{b} = 3.1415$ where $\boldsymbol{b}$ is the result of the logistic regression fit, what is the probability estimate that $y_* = 1$?

(h) [harder] If, $\boldsymbol{x}_* \boldsymbol{b} = 3.1415$ where $\boldsymbol{b}$ is the result of the probit regression fit, what is the probability estimate that $y_* = 1$?

(i) [easy] In probability estimation modeling, what is the formula for the Brier Score performance metric? Prove the Brier score is always non-positive.

(j) [easy] In probability estimation modeling, what is the formula for the Log Scoring Rule performance metric? Prove the Log Scoring Rule is always non-positive.

(k) [difficult] Generalize linear probability estimation to the case where $\mathcal{Y} = \{C_1, C_2, C_3\}$, i.e. a nominal variable with $L = 3$ levels.

Assume the logistic link function. Write down the objective function that is argmax'd over the parameters (you define what these parameters are — that is part of the question). Once you get the answer you can see how this easily goes to $L > 3$, an arbirtrary[1] number of response levels.

[1]Note: The algorithm for general L is known as all of the following: "multinomial logistic regression", "polytomous LR", "multiclass LR", "softmax regression", "multinomial logit" (mlogit), the "maximum entropy" (MaxEnt) classifier, and the "conditional maximum entropy model". You can inflate your resume with lots of redundant jazzy terms by doing this one question!

For the next two questions, let $n_1 := \sum \mathbb{1}_{y_i=1}$ and $n_0 := \sum \mathbb{1}_{y_i=0}$ so that $n = n_0 + n_1$. Then assume $n_1 \neq n_0$. This is equivalent to letting $n_1 = cn$ and $n_0 = (1-c)n$ and assuming $c \in [0,1]/\left\{\frac{1}{2}\right\}$.

(l) [harder] [MA] Prove the Brier score is always higher for g_0 vs the model where you set $\hat{p}_i = \frac{1}{2}$ for all i. Hint: $(1-c)c < \frac{1}{2}$ for $c \in [0,1]/\left\{\frac{1}{2}\right\}$.

(m) [difficult] [MA] Prove the Log Scoring Rule is always higher for g_0 vs the model where you set $\hat{p}_i = \frac{1}{2}$ for all i.