MATH 342W / 650.4 Spring 2024 Homework #3

Professor Adam Kapelner

Due 11:59PM March 17

(this document last updated 11:49pm on Saturday 23rd March, 2024)

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 and we will discuss in class. Do your best.

The problem that arises when attempting to make a model to predict the weather is the level of abstraction we created when approximating out Z's, weather patterns and other aspects are extremely complex which inherently would create alot of error in any model. In the book Lewis Fry Richardson proposes the idea of interpreting the atmosphere as a grid and solve equations for each segment. This would not only require immense computational power but small errors would propagate very quickly through the rest of the model still maintaining the issue at hand and in some sense, making it worse.

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

Weather forcast, particularly those broadcasted in television often exhibit a "wet bias", meaning they predict higher percentages of rain than what actually occurs. This bias is partially due to economic incentives. For instance, weather channels admit to exaggerating the percentages in order to cover themselves in the even of an unforeseen "sprinkle", as people are more likely to remember and react negatively to unanticipated rain than to a missed forecast of dry weather. For more honest and accurate weather predictions, the National Weather Service is recommended.

(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.

The problem with earthquake predictions, as explained in Chapter 5 of Nate Silver's book, is that earthquake prediction suffers from a fundamental lack of understanding about when and where an earthquake will occur., this lack of understanding leads to massive ignorance errors in our models as our proximal causes arent well understood making approximations nearly useless. While seismic activity can be monitored and historical data can provide statistical probabilities of occurrence (like the Gutenberg-Richter law for earthquake frequency and magnitude), these methods do not yield precise, time-specific predictions.

(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?

In Nate Silver's book, the "nonsense predictor" he describes in the whimsical explanation of overfitting is the color of the lock. He gives an example where a criminal is

tasked with figuring out a method for picking locks and comes up with specific combinations for locks based on their colors (red, black, and blue). This approach is nonsensical because it overly fits the data from a limited sample (the three locks) without establishing a general theory of lock-picking that would apply to locks of any color or type. This illustrates the concept of overfitting, where a model or theory is so tightly fit to the specific data points that it fails to predict new, unseen instances accurately.

(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?

He meant that given enough parameters, a model can be made to fit any dataset extremely closely, regardless of whether the fit is meaningful or merely coincidental. For example, with enough parameters, a model can precisely match the peculiarities of a specific dataset (like fitting the shape of an elephant), but this precise fit might not generalize well to new, unseen data.

(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.

The problem with unemployment predictions revolves around the significant uncertainty and complexity of macroeconomics. Unlike weather or earthquakes, economic indicators like unemployment are influenced by a wide array of factors, which are difficult to predict accurately. Economic forecasts often fail due to biases and politial pressures which are proximal causes we will never be able to predict. These forecasts typically present a singular outlook without an adequate explanation to the uncertainties, leading to a disconnect between public perception and the actual reliability of these predictions.

(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 colsp $[\boldsymbol{X}]$ where $\boldsymbol{X}$ is a $n \times (p+1)$ matrix with all columns linearly independent from each other. What is rank $[\boldsymbol{H}]$?

In this case the Rank of X is the p+1. Since the columns in X are linearly independent.

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

$X\ (XtX)^{-}1Xt$

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

The conceptual meaning behind this simplification is that H is the orthogonal projecttion matrix onto the column space of X. This means when we project any vector onto the column space of X using H, we obtain a vector that lies entirely within the column space of X. H maps any vector onto its closest approximation within the column space of X.

(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 colsp $[X]$ is the same exact projection matrix that projects onto colsp $[X']$.

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

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

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

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

The colspace[X]

(i) [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?

In linear regression the degrees of freedom is calculated as the number of observations (n) minus the number of linearly independent parameters.

(j) [easy] 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?

(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 colsp $[\boldsymbol{Q}] =$ colsp $[\boldsymbol{X}]$ 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 in the SST remaining the same.

(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.

Overfitting is the concept of making your model too accurate to the dataset used for training ignoring the true underlying association of the data itself. You completely lose the meaning of the result by only working in the require data.

(s) [easy] Why are "in-sample" error metrics (e.g. R^2, SSE, s_e) dishonest? Note: I'm leaving out RMSE as RMSE attempts to be honest by increasing as p increases due to the denominator. I've chosen to use standard error of the residuals as the error metric of choice going forward.

The reason why error metrics like R^2 are dishonest is because they only measure how well the midel fits the data it was trained on, not how well it generalizes to new, unsee data.

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

The way we can provide honest testing metrics is by creating subsets of the data, a set for training (the majority) and the set for testing (the complement). With this logic we are able to generate "future/unseen" data for out model. Then you calculate the R^2 based on the SSE "out of sample" and the SST "out of sample, tested on the "future" data.

(u) [easy] The procedure in (t) 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?

K-fold cross-validation reduces variation in honest error metrics by systematically partitioning the dataset into k subsets, using each in turn as a testing set while training the model on the remaining data. This method ensures that all data points contribute

equally to the performance metric, mitigating the risks of overfitting and providing a balanced assessment of the model's capabilities. By averaging the error metrics across multiple training-testing splits, k-fold cross-validation yields a more stable and reliable estimate of the model's predictive performance.

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?

The constant K controls loosely speaking the proportion of data allocated to the training versus the testing set. The trade-off is that the larger K provides more data for training the model, improving the ability to generalize. However this leaves less data for training which could make validation less reliable due to an increase in variance in the models performance estimate. On the other hand, a smaller K results in a larger test-set which could provide more reliable estimates for out out of sample metrics, but reducing the training set would cause underfitting.

(b) [harder] Assume you are doing one train-test split where you build the model on the training set and validate on the test set. If n was very large so that there would be trivial misspecification error even when using $K = 2$, would there be any benefit at all to increasing K if your objective was to estimate generalization error? Explain.

In the context of a large sample size n, utilizing a 50/50 train-test split (where $K = 2$) suffices to minimize the error due to model misspecification, as both training and testing sets are adequately large. Incrementing K, as in k-fold cross-validation, might not substantially improve the generalization error's estimation because the ample test data already provide a dependable error estimate. The advantage of increasing K lessens with a larger n, and the extra computational burden may not justify the slight improvement in estimating the generalization error. Nonetheless, a higher K could be beneficial for examining the model's robustness and identifying variations in its performance across different data segments.

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

K-fold cross-validation tries to solve the problem of estimating a model's generalization error accurately, combatting issues like overfitting and inefficient data use. It ensures all data points are used for both training and testing across multiple iterations, reducing the risk of bias that can occur with a single train-test split.

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