MATH 342W / 650.4 Spring 2024 Homework #3

Osman Khan

Sunday 17th March, 2024

Problem 1

These are questions about Silver's book, chapters 4-6. 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.

While we have a good understanding of what happens on the molecular level, we do not have the best information in regards to initial states. As weather systems are dynamic, and subject to chaos theory, slight changes in intial conditions leads to wide differences, which makes weather predictions problematic.

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

The weatherman lies because it is better to predict there is rain than not. If rain occurs, people will be prepared, and if it doesn't, people will enjoy it (with the only hassle being having to carry an umbrella). On the other hand, for the opposite situation, if it rains, people will be drenched, and if it doesn't people would have expected it and would not appreciate the dryness as much.

For honest forecasts, the government's weather department is good.

(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.

We do not know what happens with the crust and tectonic level, at least not exactly. This leads to reliance on statistical models. In contrast, with weather systems, we are able to measure.

(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?

The nonsense predictor is the lock combination. Just because one lock has a code does not mean all locks (of that color) have the same code.

(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?

The more parameters we have, the more we can interpret the data to our desire, similar to a puppet.

(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.

For unemployment, there are many variables that impact, including time. In addition, there is the observer bias that impacts economic performance.

(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.

Yes, because without a theoretical framework, you're not making a prediction, but rather a hypothesis a statement that simply includes your feelings.

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 $\text{rank}\,[\boldsymbol{H}]$?

$p+1$

(b) [easy] Simplify $\boldsymbol{H}\boldsymbol{X}$ 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) [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 $\forall n$.

For any n, and an $n * m$ matrix M, $In * M = M$

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

R^n

(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?

P+1 is the number of degrees of freedom. This means there are 14 linearly independent columns in the design matrix.

(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 $\boldsymbol{0}_n$ 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 $\text{colsp}[\boldsymbol{Q}] = \text{colsp}[\boldsymbol{X}]$ 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.

SST is based on the null model.

(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 fitting the residuals in our model.

(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.

Most in-sample error metrics can be manipulated to get arbitrarily low errors.

(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.

We can split our data into training and testing, learn from training and test on testing.

(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?

We can do the K-fold CV.

We partition the data into K splits. Then, we run K validations. In each validation, a different k^{th} split is chosen as the testing set.

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 the percentage of data that is split into train and test subsets. The tradeoff is the amount that the OOS metrics improve by, i.e. generalization error. If train is large, our model will be robust but we won't know a lot of what will happen in the future as the test is small, while if train is small, our model will not be as robust but we will have a better idea of what will happen in the future.

(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.

Noticeable overfitting occurs when $n \approx p + 1$. As $n >> p + 1$, overfitting will not be a problem here. As the train-test split helps in fighting overfitting, it would therefore

not be beneficial to increase K. In addition, train-test split is an expensive process, so the more we stay away from it the better.

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

If we were to train-test split once, we could be stuck with a subset of data that does not reflect the overal data (is an outlier).

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

K-fold CV ties to solve the issue of the oversized impact that certain subsets of data have on OOS metrics. by partitioning the data, and having all parts act train and test at different times, we are able to have average OOS metrics that are more stable.