MATH 342W / 650.4 Spring 2024 Homework #3

Benjamin Minkin

Monday 18th March, 2024

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.

Weather is a dynamic system where small errors in measurement can cause devastating effects in prediction. The problem with predicting weather is getting the very precise measurements throughout the predicting process.

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

Weathermen lie because errors in one direction are treated more harshly. Predicting rain before clear skies goes unadmonished whereas predicting clear skies before rain causes mistrust. This leads them to have a bias in favor of predicting rain. To get an accurate prediction go to the government website weather.gov instead of weather.com.

(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.

Predicting earthquakes is different than weather because we do not understand the underlying causes behind earthquakes. It is hard to predict reality when there is no model predicting when earthquakes arrive.

(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?

A predictor that predicts the combination of the lock based on its color that was trained from a dataset containing two locks.

(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?

Overfitting can cause the model to match the data very closely. What I take from this quote is to focus on creating models using relevant variables instead of trying to maximize the R^2.

(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.

Sometimes, politicians focus on minimizing unemployment. This can lead unemployment to go down faster than expected. This human intervention makes the unemployment metric to be a bad predictor on macroeconomic 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.

I mostly agree with this, Choosing good proxies is key to creating a good model.

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 $[\boldsymbol{H}]$?

(p+1)

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

$$H = X(X^TX)^{-1}X^T$$
$$HX = X(X^TX)^{-1}X^TX = XI = X$$

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

A projection onto itself or a bigger space is itself.

(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 $[X']$.

$$H = X(X^TX)^{-1}X^T$$
$$H' = X'(X^{T'}X')^{-1}X^{T'}$$

because x is parallel to x', they will have the same projection matrix as magnitude is irrelevant.

(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 $\forall n$.

Idempotency

$$II = I$$

Symmetric

$$I^T = I$$

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

Beacause I is full rank it will project onto $\mathbb{R}^n$ space

(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?

There are P degrees of freedom because a P+1 design matrix will return an r^2 of 1 because each y has an x to account for it. This is absolute overfitting and will be bad at prediction.

(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?

$$\text{Proj}_{\text{colsp}[X]}[\boldsymbol{y}] = X(X^TX)^{-1}X^Ty$$
$$OLS = (X^TX)^{-1}X^Ty$$

This is the same as OLS except that here there is an X in front.

(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.

No it will not work. This is because e does not intersect and does not contain any more information. Regression again will not be any better than the first regression so it is not an iterative process.

(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.

$$Q^T Q = Q^{-1} Q \rightarrow I = I$$

(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.

Normalize each row

$$\boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$$
$$= \frac{1}{||X||} * ||X^T X|| * \frac{1}{||X^T||} * V(V^T V)^{-1} V^T$$
$$= V(V^T V)^{-1} V^T = H = QQ^T$$

(n) [difficult] [MA] This problem is independent of the others. Let H be an orthogonal projection matrix. Prove that rank $[\boldsymbol{H}]$ = 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 colsp $[\boldsymbol{Q}]$ is the same as the sum of the projections onto each column of $\boldsymbol{Q}$.

Due to being orthogonal each column is in an independent space and has no overlap. This means that the sum of the column projections is equivalent to the whole projection at once.

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

$$SST = \sum y_i - \bar{y}^2$$

Adding to x will not affect this.

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

$$\sum_{j=1}^{p} ||proj_j(y^>)||^2$$

As you add columns, the column space increases which increases the amount projected raising SSR

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

Adding more columns will always raise SSR even when it is garbage. This means that you can add garbage to increase the R^2 fit of your model. However, this is misleading becauseb it is resulting from garbage in. This is overfitting.

(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.

Adding more columns will always raise SSR even when it is garbage. This means that you can add garbage to increase the R^2 fit of your model. However, this is misleading becauseb it is resulting from garbage in. This is overfitting.

(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.

To help, you can partition the data into two categories. One will be the training data and the other will be the validation data. This will help with understanding the underlying relationships instead of focusing on fitting the sample 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?

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?

K is the amount of training sets partitioned from the data. Here K=1 because there is one training set.

(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.

There would be some benefit to increasing K. It would reduce the variance of the predictions from the model.

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

It tries to solve the issue of training on all of the data while having a way to validate without causing overfitting. Without K-fold CV, the training data would also be used to validate the data leading to possible overfitting. K-fold CV also outputs a prediction with less variance.

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