DS 100/200: Principles and Techniques of Data Science **Date: 10/25/19**

Homework #6 Solutions

Due Date: Friday, 11/1/19 at 11:59 PM

- **You will turn in this homework by uploading your answers in PDF format to Gradescope.** You may turn in your answer as a scan or good quality camera phone picture of handwritten sheets (e.g. CamScanner), or you may turn it in as a PDF generated from typeset math (e.g. using LaTeX or Microsoft Word).
- For this homework we have provided a companion notebook on DataHub. It is also linked on the course website. Problems 1, 8, and 9 explicitly ask you to run and interpret code in this notebook. You may also find the notebook helpful for problems 2 through 6. **You will not need to turn in any .ipynb files.**
- Due to resource constraints, we may elect to only grade a subset of the problems. We also may grade a subset of problems on completion, rather than on correctness. Since you dont know which problems these are, though, its in your best interest to fully attempt all of them.

Collaborators

Data science is a collaborative activity. While you may talk with others about the homework, we ask that you write your solutions individually. If you do discuss the assignments with others please include their names at the top of your submission.

Linear Regression Fundamentals and the Normal Equation

1. In this problem, we will review some of the core concepts in linear regression.

   a. Suppose we create a linear model with parameters $\vec{\hat{\beta}} = [\hat{\beta}_0, \ldots, \hat{\beta}_p]$. As we saw in lecture, such a model makes predictions $\hat{y} = \vec{\hat{\beta}} \cdot \vec{x} = \sum \hat{\beta}_i x_i$.

   Suppose $\vec{\hat{\beta}} = [2, 0, 1]$ and we receive an observation $x = [1, 2, 3]$. What $\hat{y}$ value will this model predict for the given observation?

   > **Solution:** $\hat{y} = \vec{x} \cdot \vec{\hat{\beta}} = (1)(2) + (2)(0) + (3)(1) = 5$

   b. Suppose the correct y was 3.5. What will be the L_2 loss for our prediction $\hat{y}$ from question 1a?

   > **Solution:** $(y - \hat{y})^2 = 2.25$

   c. In the companion notebook for this homework, we have provided a design matrix $\mathbb{X}$ and a vector of response variables y. These are given as variables `X` and `y` in the companion notebook. Suppose we create a linear regression model using this data. Explain briefly why $\vec{\hat{\beta}}$ will be a 6×1 vector.

   > **Solution:** Since $\hat{y} = X\vec{\hat{\beta}}$, the number of columns in X must equal the number of rows in $\vec{\hat{\beta}}$. Another way to think about this is that $\vec{\hat{\beta}}$ is a vector of coefficients, one for each feature in X. Since X has 6 features, $\vec{\hat{\beta}}$ should have dimension 6×1.

   d. Using the normal equation from lecture 16, compute the optimal $\vec{\hat{\beta}}$. Rather than giving all six values, in your answers, just tell us which of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5$ is largest, and give its value rounded to two decimal places. Hint: `np.linalg.inv` might be useful. Hint: `x.T` gives the transpose of a matrix.

   > **Solution:** $\hat{\beta}_5 = 46.26$ is the largest value.

   e. What will our model give for $\hat{y}_1$, i.e. what will it predict for the 1st observation (i.e. row one of $\mathbb{X}$). Give your answer rounded to two decimal places.

Solution: $\hat{y_1} = 18.65$. We got this value by multiplying the first row of X by $\hat{\beta}$.

f. What is the L_2 loss for this prediction? What is the residual e_1?

Solution: The true y_1 value is 18, so:

$$L_2 = (y_1 - \hat{y}_1)^2 = 0.4225$$

$$e_1 = y_1 - \hat{y}_1 = -0.65$$

.

g. In lecture we said that the $\hat{\beta}$ that results from the normal equation will "minimize the empirical risk". Does there exist a $\vec{\beta}_{\text{other}}$ that would yield a lower loss for our prediction $\hat{y_1}$. If so, explain why we don't use $\vec{\beta}_{\text{other}}$ instead. If not, explain why none exists.

Solution: There does exist a $\vec{\beta}_{\text{other}}$ that yields a lower loss for the prediction $\hat{y_1}$. In fact, we can find such a $\vec{\beta}_{\text{other}}$ that yields a loss of 0 for the prediction $\hat{y}_1$ since we are only looking to minimize the error for the first data point. Note that we don't use $\vec{\beta}_{\text{other}}$ because it only minimizes the loss for $\hat{y}_1$ and not the mean squared error over all of the data points.

Note: There are other equivalent notations for linear models. The notation we've used in this problem is consistent with what we saw in Prof. Nolan's lectures, i.e. using arrows to represent vectors, hats to represent estimates, and betas to represent parameters. Other sources use different notation, e.g. bolding for vectors or even not making any typographical distinction between vectors and scalars. Hats are often omitted. Some sources use the Greek letter θ or the English letter w instead of β.

For instance, instead of saying $\hat{y} = \vec{\hat{\beta}} \cdot \vec{x}$, the Data 100 textbook uses $f_\theta(x) = \theta \cdot \mathbf{x}$, i.e. bold to represent vectors, theta instead of beta, and $f_\theta(x)$ instead of $\hat{y}$. CS 189 uses $\hat{y}_i = w \cdot X_i$. Data 100's Spring 2019 lectures used $E[Y|X] = X^T\beta$. Even our own lectures this semester are not entirely self consistent. We know it's annoying, but you'll just have to get used to this lack of a common consistent symbolic language.

Observation Space vs. Variable Space

2. The "variable space" approach views the design matrix $\mathbb{X}$ as a collection of $n \times 1$ column vectors, one for each variable. On the other hand, the "observation space" approach considers the design matrix as a collection of $1 \times p$ row vectors, one for each observation. In this exercise, we will examine many of the terms that we have been working with in regression (e.g. $\hat{\beta}$) and connect them to their dimensions and to concepts that they represent. We will also draw connections between the observation and variable spaces.

   First, we define some notation for the vectors in these two spaces. The $n \times p$ design matrix $\mathbb{X}$ corresponds to n observations on p variables (where one of these variables might actually be the one vector $\vec{1}$, a.k.a. a bias vector). $\vec{y}$ is the response variable. We assume in this problem that we use $\mathbb{X}$ and $\vec{y}$ to compute optimal parameters $\vec{\hat{\beta}}$ for a linear model, and that this linear model generates predictions $\vec{\hat{y}}$ from $\vec{\hat{\beta}}$ and $\mathbb{X}$ as we saw in lecture and in question 1 of this homework. We introduce new notation on this homework for the row and column vectors of $\mathbb{X}$ as follows:

   $$\begin{array}{ll} \vec{x}_{*j} & j^{th} \text{ column vector in } \mathbb{X}, j = 1, \ldots, p \\ \vec{x}_{i*} & i^{th} \text{ row vector in } \mathbb{X}, i = 1, \ldots, n \end{array}$$

   Below, on the left, we have several expressions, labelled a through h, and on the right we have several terms, labelled 1 to 10. **For each expression, determine its shape (e.g., $n \times p$), and match it to one the given terms.** Terms may be used more than once or not at all. If a specific expression is nonsensical because the dimensions don't line up for a matrix multiplication, write "N/A" for both.

   a. $\mathbb{X}$
   b. $\vec{\hat{\beta}}$
   c. $\vec{x}_{*j}$
   d. $\vec{x}_{1*}\vec{\hat{\beta}}$
   e. $\vec{x}_{*1}\vec{\hat{\beta}}$
   f. $\mathbb{X}\vec{\hat{\beta}}$
   g. $(\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t\vec{y}$
   h. $(I - \mathbb{X}(\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t)\vec{y}$

   1. the residuals
   2. 0
   3. 1st response, y_1
   4. 1st predicted value, $\hat{y_1}$
   5. 1st residual, e_1
   6. the estimated coefficients
   7. the predicted values
   8. the features for a single observation
   9. the value of a specific feature for all observations
   10. the design matrix

   As an example, for 2a, you would write: "2a. **Dimension:** $n \times p$, **Term:** 10".

Solution:

a. $\mathbb{X}$ has dimension $n \times p$ and is equivalent to 10 (the design matrix).

b. $\vec{\hat{\beta}}$ has dimension $p \times 1$ and is equivalent to 6 (the estimated coefficients).

c. $\vec{x}_{*j}$ has dimension $n \times 1$ and is equivalent to 9 (the value of a specific feature for all observations).

d. $\vec{x}_{1*}\vec{\hat{\beta}}$ has dimension 1 and is equivalent to 4 (the first predicted value).

e. $\vec{x}_{*1}\vec{\hat{\beta}}$ has dimension N/A and is equivalent to N/A.

f. $\mathbb{X}\vec{\hat{\beta}}$ has dimension $n \times 1$ and is equivalent to 7 (the predicted values).

g. $(\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t\vec{y}$ has dimension $p \times 1$ and is equivalent to 6 (the estimated coefficients).

h. $(I - \mathbb{X}(\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t)\vec{y}$ has dimension $n \times 1$ and is equivalent to 1 (the residuals).

Deriving Properties of the Simple Linear Regression

In lectures 14 and 15, we spent a great deal of time talking about the simple linear regression, which you also saw in Data 8. To briefly summarize, the simple linear regression model assumes that given a vector of observations $\vec{x}$, our predicted response for each of these observations is given by $\vec{\hat{y}} = \hat{\beta}_0\vec{1} + \hat{\beta}_1\vec{x}$. Or equivalently, given a single observation x, our predicted response for this observation is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

In lecture 14, we focused on the observation space representation, and saw that the $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the L_2 loss for the simple linear regression model are:

$$\hat{\beta}_0 = \bar{y} - \beta_1\bar{x}$$
$$\hat{\beta}_1 = r\frac{SD_y}{SD_x}$$

Or, rearranging terms, our predictions $\hat{y}$ are:

$$\hat{y} = \bar{y} + rSD_y\frac{x - \bar{x}}{SD_x}$$

In lecture 15, we used the exact same model, but now switched over to the variable space representation to get some geometric intuition for what we saw in lecture 14. The key geometric insight was that if we train a model on $\vec{x}$ and $\vec{y}$ and we use this model to make a prediction on a new observation x, our predicted $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is simply the vector in $\text{span}(\vec{1}, \vec{x})$ that is closest to y.

Consider some useful properties of the simple linear regression, listed below. Use the results derived from either the variable space or the observation space representation to prove these properties.

You may find the companion notebook helpful to support your thinking. See "Properties of Simple Linear Regression With and Without a Constant Term". Note: You may not answer questions 3 - 6 by simply computing the residuals of the given dataset and noting that the sum is zero. We want you to show that these properties are true for ALL possible datasets.

3. Show that/explain why the residuals from the fit have an average of 0, i.e. $\sum e_i = 0$.

Solution: In the variable space representation, we can think of $\sum e_i$ as the dot product of the residual vector, $\vec{e}$ and the one vector, $\vec{1}$. That is,

$$\sum e_i = \vec{e} \cdot \vec{1}$$

The predicted $\vec{\hat{y}}$ is the vector closest to $\vec{y}$ in the span$(\vec{1}, \vec{x})$. In other words, $\vec{\hat{y}}$ is the projection of $\vec{y}$ into the span, and the difference $\vec{y} - \vec{\hat{y}}$ is orthogonal to any vector in the span$(\vec{1}, \vec{x})$. So, $(\vec{y} - \vec{\hat{y}})$ is orthogonal to $\vec{1}$. Orthogonality, means that the dot product between $\vec{y} - \vec{\hat{y}}$ and any vector in the span is 0. In particular,

$$(\vec{y} - \vec{\hat{y}}) \cdot \vec{1} = 0$$

Since $\vec{e} = \vec{y} - \vec{\hat{y}}$, we have shown that

$$\vec{e} \cdot \vec{1} = 0$$

This same argument can be used to establish that $\vec{x} \cdot \vec{e} = 0$ because $\vec{x}$ is also in the span$(\vec{1}, \vec{x})$.

And, again, since $\vec{\hat{y}}$ is in this span, it also follows that $\vec{\hat{y}} \cdot \vec{e} = 0$. In other words, we have just given explanations (from the variable space approach) for #3, 4, and 5.

Alternatively, we can use matrix properties to establish the validity of #3, 4, and 5. The predicted value is expressed as

$$\vec{\hat{y}} = \mathbb{X}(\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t\vec{y}$$

The quantity $\mathbb{X}(\mathbb{X}^t\mathbb{X})^{-1}\mathbb{X}^t$ is a projection matrix that projects into the span of $\mathbb{X}$. We call it H or the 'hat' matrix. Note that $H^t = H$ and $HH = H$ and $(I - H)H = 0$.

We use these properties to establish #3, 4, 5. Since $\vec{1}$ is in the span of $\mathbb{X}$, we have

$$H\vec{1} = \vec{1}$$

and since

$$\vec{e} = (I - H)\vec{y}$$

it follows that

$$\vec{e} \cdot \vec{1} = \vec{y}^t(I - H)\vec{1} = 0$$

For the same reasons $\vec{x} \cdot \vec{e} = 0$. Lastly,

$$\vec{\hat{y}} \cdot \vec{e} = y^t H^t(I - H)y = 0$$

Alternatively, we can look at the problem in observation space, and use the equation

for $\hat{y}_i$ shown above,

$$\begin{aligned}
\sum_i e_i &= \sum_i (y_i - \hat{y}_i) \\
&= \sum_i (y_i - (y + rSD_y \frac{(x_i - \bar{x})}{SD_x}) \\
&= \sum_i y_i - \sum_i \bar{y} - r\frac{SD_y}{SD_x} \sum_i (x_i - \bar{x}) \\
&= n\bar{y} - n\bar{y} - r\frac{SD_y}{SD_x}[n\bar{x} - n\bar{x}] \\
&= 0
\end{aligned}$$

4. Show that/explain why the $n \times 1$ vectors $\vec{x}$ and $\vec{e}$ are orthogonal. In other words, explain why the dot product (a.k.a. inner product) of any observation used to train the model and the residuals is 0, i.e. $\vec{x} \cdot \vec{e} = \sum(x_i e_i) = 0$.

Solution: In the previous problem, we explained why $\vec{e}$ is orthogonal to any vector in the span$(\vec{1}, \vec{x})$. Therefore it follows that the dot product of $\vec{e}$ and $\vec{x}$ is 0.

5. Show that/explain why the dot product of the residuals and $\vec{\hat{y}}$ is 0, i.e. $\vec{\hat{y}} \cdot \vec{e} = \sum(\hat{y}_i e_i) = 0$.

Solution: Note that $\vec{\hat{y}} = \hat{\beta}_0 \vec{1} + \hat{\beta}_1 \vec{x}$, which means that $\vec{\hat{y}} \in \text{Span}(\vec{1}, \vec{x})$. Hence, we know that $\vec{\hat{y}}$ is in the column space of our feature matrix, which means that it is orthogonal to the residuals. Hence, we have $\vec{\hat{y}} \cdot \vec{e} = 0$.

6. Show that/explain why $(\bar{x}, \bar{y})$ is on the regression line.

Solution: Since this question is about a particular point $(\bar{x}, \bar{y})$, it is easiest to establish this property from the observation space perspective. That is, the regression line are all points (x, y) where,

$$y = \hat{\beta}_0 + \hat{\beta}_1 x,$$

which we can rewrite (using the definition of $\hat{\beta}_0$ from above as:

$$y = \bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x.$$

When we plug $\bar{x}$ in for x, we find the righthand side becomes

$$\bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1\bar{x},$$

which reduces to $\bar{y}$. We see that $(\bar{x}, \bar{y})$ is on the regression line.

Properties of a Linear Model With No Constant Term

Suppose that we don't include the intercept term in our model, that is, our model is now simply $\vec{\hat{y}} = \hat{\gamma}\vec{x}$, where $\hat{\gamma}$ is the single parameter for our model that we need to optimize.

In this case, our least squares fit finds the γ that minimizes:

$$\sum_{i=1}^{n}(y_i - \gamma x_i)^2$$

for observed data $(x_i, y_i), i = 1, \ldots, n$.

The normal equations derived in lecture no longer hold. In this problem, we'll derive a solution to this simpler model. We'll see that the least squares estimate of the slope in this model differs from the simple linear regression model, and will also explore whether or not our properties from the previous problem still hold.

7. Use calculus to find the minimizing $\hat{\gamma}$. That is, prove that

$$\hat{\gamma} = \frac{\sum x_i y_i}{\sum x_i^2}$$

Note: This is the slope of our regression line, analogous to $\hat{\beta}_1$ from our simple linear regression model.

Solution: Differentiate the sum of squares with respect to γ to find:

$$-2\sum_{i=1}^{n}(y_i - \gamma x_i)x_1$$

Set the derivative and solve for the minimizing $\hat{\gamma}$

$$\begin{aligned} 0 &= -2\sum_{i=1}^{n}(y_i - \hat{\gamma}x_i)x_1 \\ &= \sum_{i=1}^{n} y_i x_i - \hat{\gamma}\sum_{i=1}^{n} x_i^2 \end{aligned}$$

Rearrange terms

$$\sum_{i=1}^{n} y_i x_i = \hat{\gamma}\sum_{i=1}^{n} x_i^2$$

to find,

$$\hat{\gamma} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

8. In the previous section on deriving properties of estimators for a simple linear model, you established the following properties:

   - $\sum e_i = 0$.
   - $\vec{\hat{y}}$ and $\vec{e}$ are orthogonal.
   - $\vec{x}$ and $\vec{e}$ are orthogonal.
   - $(\bar{x}, \bar{y})$ is on the regression line.

   Which of these properties are still true? Support your answers by giving the values of the following quantities as computed on the dataset given in "Properties of Simple Linear Regression With and Without a Constant Term" in the companion notebook for this homework.

   - $\bar{e} = 0$

     Solution: This property does not hold anymore. Note that our proof for question 3 requires $\vec{1}$ to be one of the columns in our feature matrix. Without an intercept term, the feature matrix might not have $\vec{1}$ as one of its columns. Hence, we do not know for sure that $\vec{1} \cdot \vec{e} = 0$.

   - $\vec{\hat{y}} \cdot \vec{e} = 0$

     Solution: This property still holds. Note that our feature matrix in this scenario is just one column, which is $\vec{x}$. Since the residuals are orthogonal to the column space of our feature matrix, we know that they are orthogonal to anything in the Span of $\vec{x}$. Note that $\vec{\hat{y}} = \hat{\gamma}\vec{x}$, which means that $\vec{\hat{y}} \in \text{Span}(\vec{x})$. Hence $\vec{\hat{y}}$ is orthogonal to the residuals, which means that $\vec{\hat{y}} \cdot \vec{e} = 0$.

   - $\vec{x} \cdot \vec{e} = 0$

     Solution: This property still holds. From the previous part, we know that $\vec{x}$ is the only column in our feature matrix. Since the residuals are orthogonal to the column space of our feature matrix, we know that $\vec{x} \cdot \vec{e} = 0$.

   - $\hat{\gamma}\bar{x}$

     Solution: This property does not hold anymore. Note that our derivation in question 6 relied on the value of $\hat{\beta}_0$. However, $\hat{\beta}_0$ does not exist anymore since we don't have an intercept term, which means that $\hat{\gamma}\bar{x}$ is not necessarily equal to $\bar{y}$.

9. Recall that we can decompose the total sum of squares into two sum of squares, one measuring the variability "explained" by the regression and the other measuring the variability of the errors (the e_i). Does this property still hold? That is,

$$\sum_i (y_i - \bar{y})^2 \stackrel{?}{=} \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2$$

Support your answer by computing these three quantities using the dataset described in the previous problem.

Solution: This property does not hold anymore. To see why this is the case, we should consider why the property holds with an intercept term. Let $\vec{1}$ be a vector of all ones in the appropriate dimension. Then, the expression is equivalent to the following:

$$\left\| \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix} \right\|^2 = \left\| \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ \vdots \\ y_n - \hat{y}_n \end{bmatrix} \right\|^2 + \left\| \begin{bmatrix} \hat{y}_1 - \bar{y} \\ \hat{y}_2 - \bar{y} \\ \vdots \\ \hat{y}_n - \bar{y} \end{bmatrix} \right\|^2$$

which is equivalent to:

$$||\vec{y} - \bar{y}\vec{1}||^2 = ||\vec{y} - \vec{\hat{y}}||^2 + ||\vec{\hat{y}} - \bar{y}\vec{1}||^2$$

In the case where we have an intercept term, we know that $\vec{1}$ is one of the columns in our feature matrix. Hence, Span($\vec{1}$) is in the column space of our feature matrix, which means that $\bar{y}\vec{1}$ is also in the column space of our feature matrix. From lecture we know that the residuals $\vec{y} - \vec{\hat{y}}$ are orthogonal to any vector in the span of the column space, which includes $\bar{y}\vec{1}$ if we have an intercept term. Note that $\vec{\hat{y}}$ is also in the column space of our feature matrix, which means that the difference $\vec{\hat{y}} - \bar{y}\vec{1}$ is also in the column space of our feature matrix. Since $\vec{y} - \vec{\hat{y}}$ is orthogonal to $\vec{\hat{y}} - \bar{y}\vec{1}$ the Pythagorean theorem tells us that $||\vec{y} - \bar{y}\vec{1}||^2 = ||\vec{y} - \vec{\hat{y}}||^2 + ||\vec{\hat{y}} - \bar{y}\vec{1}||^2$.

When we don't have an intercept term, $\vec{1}$ might not be one of the columns in our feature matrix, which means that $\bar{y}\vec{1}$ and therefore $\vec{\hat{y}} - \bar{y}\vec{1}$ might not be in the column space of our feature matrix. Hence, the Pythagorean theorem might not hold anymore because the angle between $\vec{y} - \vec{\hat{y}}$ and $\vec{\hat{y}} - \bar{y}\vec{1}$ is not necessarily 90 degrees, which means that $||\vec{y} - \bar{y}\vec{1}||^2$ is not necessarily equal to $||\vec{y} - \vec{\hat{y}}||^2 + ||\vec{\hat{y}} - \bar{y}\vec{1}||^2$.