DS 100/200: Principles and Techniques of Data Science **Date: March 13, 2020**

Discussion #8 Solutions

Name:

Linear Regression Fundamentals

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

   (a) Suppose we create a linear model with parameters $\hat{\theta} = [\hat{\theta}_0, \ldots, \hat{\theta}_p]$. As we saw in lecture, given an observation $\boldsymbol{x}$, such a model makes predictions $\hat{y} = \hat{\theta} \cdot \boldsymbol{x}$.
   Suppose $\hat{\theta} = [2, 0, 1]$ and we receive an observation $\boldsymbol{x_1} = [1, 2, 3]$. What $\hat{y_1}$ value will this model predict for the given observation?

   Solution: $\hat{y_1} = \hat{\theta} \cdot \boldsymbol{x_1} = (2)(1) + (0)(2) + (1)(3) = 5$

   (b) Suppose the correct y_1 was 3.5. What will be the L_2 loss for our prediction $\hat{y_1}$ from question 1a?

   Solution: $(y_1 - \hat{y_1})^2 = 2.25$

   (c) Suppose we receive another observation $\boldsymbol{x_2} = [-2, 5, 1]$. What $\hat{y_2}$ value will this model predict for the given observation?

   Solution: $\hat{y_2} = \hat{\theta} \cdot \boldsymbol{x_2} = (2)(-2) + (0)(5) + (1)(1) = -3$

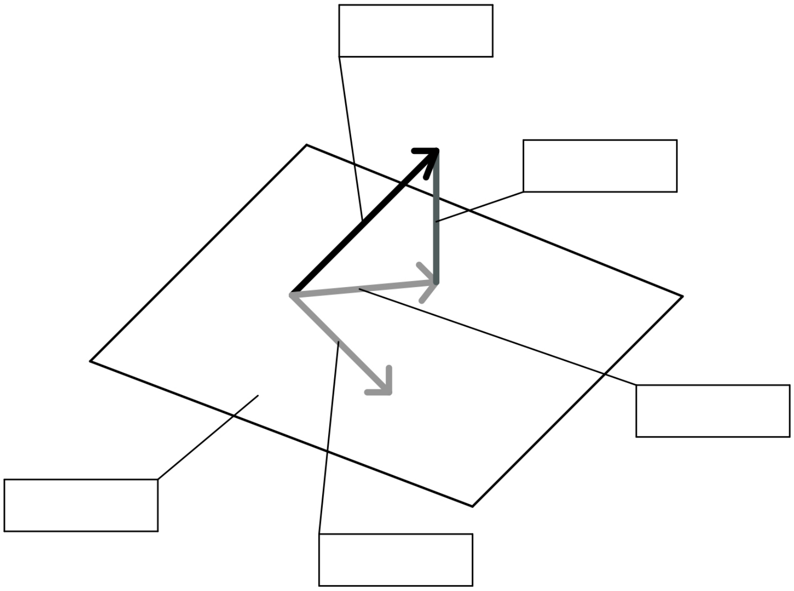
   (d) Suppose the correct y_2 for 1c was was -4. What will be the mean squared error of the $\hat{\theta}$ from 1a given the two observations (from 1b and 1c)?

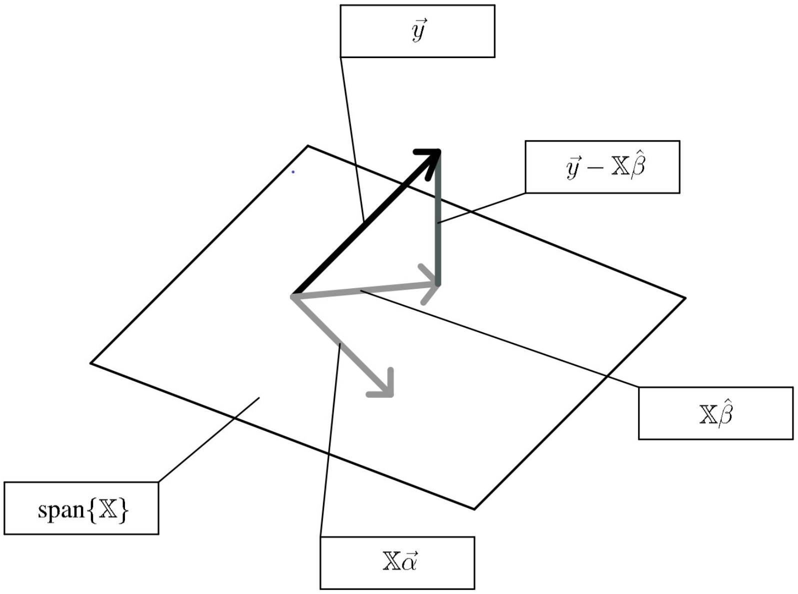
   Solution: We want to find the L_2 loss for $\hat{y_2}$. Once we find that, we take the average with the L_2 loss for our prediction $\hat{y_1}$ that we found in part 1b.

   $$(-3 - (-4))^2 = 1$$
   $$\frac{1 + 2.25}{2} = 1.625$$

Geometry of Least Squares

2. Suppose we have a dataset represented with the design matrix $\text{span}(\mathbb{X})$ and response vector $\vec{y}$. We use linear regression to solve for this and obtain optimal weights as $\hat{\beta}$. Draw the geometric interpretation of the column space of the design matrix $\text{span}(\mathbb{X})$, the response vector $\vec{y}$, the residuals $\vec{y} - \mathbb{X}\hat{\beta}$, and the predictions $\mathbb{X}\hat{\beta}$.




$\vec{y}$
$\vec{y} - \mathbb{X}\hat{\beta}$
$\mathbb{X}\hat{\beta}$
$\text{span}\{\mathbb{X}\}$
$\mathbb{X}\vec{\alpha}$

Solution:

(a) What is always true about the residuals in least squares regression? Select all that apply.

□ A. They are orthogonal to the column space of the design matrix.

□ B. They represent the errors of the predictions.

□ C. Their sum is equal to the mean squared error.

□ D. Their sum is equal to zero.

□ E. None of the above.

Solution: (A), (B)

(C): (C) is wrong because the mean squared error is the *mean* of the sum of the *squares* of the residuals.

(D): A counter-example is: $\mathbb{X} = \begin{bmatrix} 2 & 3 \\ 1 & 5 \\ 2 & 4 \end{bmatrix}, \vec{y} = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}$. After solving the least square problem, the sum of the residuals is -0.0247, which is not equal to zero. However, note that this statement is in general true if every feature contains the same constant intercept term.

(E): is wrong since (C) is wrong.

(b) Which are true about the predictions made by OLS? Select all that apply.

- □ A. They are projections of the observations onto the column space of the design matrix.
- □ B. They are linear in the features.
- □ C. They are orthogonal to the residuals.
- □ D. They are orthogonal to the column space of the features.
- □ E. None of the above.

Solution: (A), (B), (C)

(A) is correct because they are linear projections onto the column space. This fact also makes (C) correct, (E) incorrect, (D) incorrect.

(B) is correct. Even in, for example, polynomial regression the resulting predictions are linear in the new/transformed features.

Modeling

1. We wish to model exam grades for DS100 students. We collect various information about student habits, such as how many hours they studied, how many hours they slept before the exam, and how many lectures they attended and observe how well they did on the exam. Propose a model to predict exam grades and a loss function to measure the performance of your model on a single student.

   Solution: Example solution: Let us choose our parameters to be $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3]$. Let x_i be the i^{th} student. Let x_{i1}, x_{i2}, x_{i3} correspond to hours studied, hours slept and number of lectures attended respectively by the i^{th} student. Let y be the actual student's score on the exam.

   $$\begin{aligned} f(x_i) &= \theta_1 * x_{i1} + \theta_2 * x_{i2} + \theta_3 * x_{i3} \\ L(\boldsymbol{\theta}, x_i, y_i) &= (f(x_i) - y_i)^2 \end{aligned}$$

2. Suppose we collected even more information about each student, such as their eye color, height, and favorite food. Do you think adding these variables as features would improve our model?

   Solution: These features are most likely not going to improve our model. This problem is meant to emphasize overparameterization/overfitting using too many features that do not contribute to the performance of the model. Overfitting, bias variance and feature engineering will be discussed later on in the semester.