Data 100, Spring 2023

Exam Prep Section #5 Solutions

Linear Regression Fundamentals

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

   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 $\vec{x}$, such a model makes predictions $\hat{y} = \hat{\theta} \cdot \vec{x} = \hat{\theta}_0 + \hat{\theta}_1 x_1 + \hat{\theta}_2 x_2 + \ldots + \hat{\theta}_p x_p$. Assume that the design matrix and observations used to construct $\hat{\theta}$ are $\mathbb{X}$ and $\mathbb{Y}$.

   (a) Suppose $\hat{\theta} = [2, 0, 1]$ and we receive an observation $\vec{x_1} = [1, 2, 5]$. What $\hat{y_1}$ value will this model predict for the given observation?

   Solution: $\hat{y_1} = \hat{\theta} \cdot \vec{x_1} = (2)(1) + (0)(2) + (1)(5) = 7$

   (b) Suppose the true y_1 was 3.5. What will be the L_2 loss for our prediction $\hat{y_1}$ from the previous part?

   Solution: $(y_1 - \hat{y_1})^2 = (3.5 - 7)^2 = 12.25$

   (c) Which of the following are true?

   - ☐ A. $\vec{x_1}^T \hat{\theta} = 0$
   - ☐ B. $\vec{x_1}^T \hat{\theta} = \hat{y_1}$
   - ☐ C. $\vec{x_1}^T \hat{\theta} - y_1 = 0$
   - ☐ D. $\mathbb{X}\hat{\theta} - \mathbb{Y} = 0$
   - ☐ E. $\mathbb{X}^T(\mathbb{X}\hat{\theta} - \mathbb{Y}) = 0$

   (d) (T/F) Define the residuals of this model as $e_i = y_i - \hat{y}_i$. For all data points x_i and y_i in $\mathbb{X}$, $\mathbb{Y}$, the sum of residuals $\sum_i e_i = 0$. Justify why.

   Solution: Since it has an intercept term, the residuals are orthogonal to all columns in $\mathbb{X}$, including the bias column. Hence, $\vec{e}^T \vec{1} = \sum_i e_i = 0$. This is true.

   (e) Suppose we arbitrarily removed a feature from the design matrix. Which of the following could happen to the new optimal loss compared to the old optimal loss?

   - ◯ A. The optimal loss must decrease.
   - ◯ B. The optimal loss may decrease, but it must not increase.

◯ C. The optimal loss may increase, but it must not decrease.

◯ D. The optimal loss must increase.

(f) Regardless of the previous part, suppose we arbitrarily added a feature to the design matrix, and we know that the rank of the design matrix increased. However, the optimal loss remains the exact same. Explain how this could occur.

Solution: This could occur if we can't "reach" $\mathbb{Y}$ any better given this new feature. As an example, suppose our first feature vector is $[1, 0, 0]^T$. We are suddenly given another feature $[0, 1, 0]^T$.

Suppose our $\mathbb{Y} = [0, 0, 1]^T$. There is no way we can construct the 1 in the last dimension using either of our features, before or after we received the new feature. The best we can do with the first feature alone or both features is to simply set $\theta = 0$.

In general, if the new feature is orthogonal to the residuals, it will be of no use!

Weighted Least Squares

Shiangyi wants to extend her multiple linear regression modeling framework to incorporate more explicit outlier desensitization for predicting housing prices. One of the ways to do this is to remove outliers, but instead of removing them entirely, perhaps we can choose to "care" less about them through our loss function.

In other words, we can change our loss function slightly to assign *less* of a weighting to the loss of these outliers. To do this, she decides to weight each sample by a particular amount α_i in the calculation of the loss function. In other words, we augment the loss function as follows:

$$L(\theta) = \sum_i \alpha_i (y_i - x_i^T \theta)^2$$

(a) Show that the augmented loss function can be written as follows in matrix/vector notation (i.e. without any summations) for some matrix A that you will find. Assume that α is a vector such that the ith element contains α_i.

$$L(\theta) = \|A(y - X\theta)\|_2^2$$

Solution:

Our loss is written as a sum of a bunch of squared terms, so we should be able to write this as the norm-squared of some vector. Now, it's our business to find that vector. First, we know that $y - X\theta$ will give us an $n \times 1$ dimensional vector, where the ith component is $y_i - x_i^T\theta$. We now need to extend this to find a vector such that the ith component is $\sqrt{\alpha_i}(y_i - x_i^T\theta)$. In this way, when we take the norm-squared of that vector, we will obtain $L(\theta)$. Note that $\text{diag}(\sqrt{\alpha})(y - X\theta)$ accomplishes this, since the ith component of this vector is $\sqrt{\alpha_i}(y_i - x_i^T\theta)$. Taking the norm-squared of this quantity:

$$\|A(y - X\theta)\|_2^2 = \sum_{i=1}^{n} \alpha_i (y_i - x_i^T\theta)^2 = L(\theta)$$

where

$$A = \text{diag}(\sqrt{\alpha}) = \begin{bmatrix} \sqrt{a_1} & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{a_2} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sqrt{a_n} \end{bmatrix}$$

(b) Using the loss vector specified in matrix/vector notation, derive the optimal solution for θ in terms of the appropriate variables (i.e. X, y, α).

Hint: You should not be doing any optimization (i.e. calculus) in this part!

Solution:

$$L(\theta) = \|A(y - X\theta)\|_2^2$$
$$L(\theta) = \|Ay - (AX)\theta\|_2^2$$

Set $y' = Ay = (\text{diag}\sqrt{\alpha})y$ and $X' = Ax = (\text{diag}\sqrt{\alpha})X$. Then:

$$L(\theta) = \|(y' - X'\theta)\|_2^2$$

The OLS solution to the above is simply: $\hat{\theta} = (X'^T X')^{-1} X'^T y'$. Substituting the appropriate quantities in terms of α will yield the solution. More specifically, we have:

$$\hat{\theta} = ((AX)^T(AX))^{-1}(AX)^T Ay$$
$$\ldots = (X^T A^T AX)^{-1} X^T A^T Ay$$
$$\ldots = (X^T A^2 X)^{-1} X^T A^2 y$$

where we use the fact that $(AX)^T = X^T A^T$ and that $A^T = A$ since A is diagonal.

(c) True/False: The weighting function $\alpha_i = f(y_i)$ must be linear in terms of X and y for the optimal solution derived to hold. Why or why not?

Solution: False. The problem to minimize the loss with respect to θ depends only on θ being linear, hence we can have any non-linearities in terms of X and y. That is how linearizing transformations work!

(d) Suggest a usage for the following weighting functions $\alpha_i = f(y_i)$:

$$f(y_i) = \frac{e^{y_i}}{\sum_j e^{y_j}}$$

$$f(y_i) = \frac{e^{-y_i}}{\sum_j e^{-y_j}}$$

Solution: The first weighting function overweights extremely positive outliers, so we will assign the biggest weights to the loss value on outliers corresponding to the highest-priced houses.

The second weighting is the opposite - it assigns the highest weight to the losses on the lowest-priced houses and the lowest weight to the highest-priced houses.

(e) The weighting function $\alpha_i = f(...)$ can be a function of the following variables while being a linear model:

- ☐ A. X
- ☐ B. y
- ☐ C. θ

Solution: Anything in terms of θ will introduce a non-linear term, but anything involving X and y is fine since neither are involved in the linearity of the model in terms of θ.