Data 100, Spring 2022 **March 11, 2022**

Discussion #8

Fairness in Housing Appraisal

Recall that Project 1's dataset comes from the Cook County Assessor's Office (CCAO) in Illinois, a government institution that determines property taxes across most of Chicago's metropolitan area and its nearby suburbs. In the United States, all property owners are required to pay property taxes, which are then used to fund public services including education, road maintenance, and sanitation.

1. "How much is a house worth?" If you were a homeowner, why would you want your property to be valued high? Why would you want your property to be valued low?

2. Which of the following scenarios strike you as unfair and why? You can choose more than one. There is no single right answer but you must explain your reasoning.

    A. A homeowner whose home is assessed at a higher price than it would sell for.
    B. A homeowner whose home is assessed at a lower price than it would sell for.
    C. An assessment process that systematically overvalues inexpensive properties and undervalues expensive properties.
    D. An assessment process that systematically undervalues inexpensive properties and overvalues expensive properties.

3. Imagine your home is assessed at a higher value than you believe it would sell for on the market. What might that concretely mean to you, as an individual homeowner?

Dummy Variables/One-hot Encoding

In order to include a qualitative variable in a model, we convert it into a collection of dummy variables. These dummy variables take on only the values 0 and 1. For example, suppose we have a qualitative variable with 3 possible values, call them A, B, and C, respectively. For concreteness, we use a specific example with 10 observations:

$$[A, A, A, A, B, B, B, C, C, C]$$

We can represent this qualitative variable with 3 dummy variables that take on values 1 or 0 depending on the value of this qualitative variable. Specifically, the values of these 3 dummy variables for this dataset are x_A, x_B, and x_C, arranged from left to right in the following design matrix, where we use the following indicator variable:

$$x_{k,i} = \begin{cases} 1 & \text{if } i\text{-th observation has value } k \\ 0 & \text{otherwise.} \end{cases}$$

This representation is also called one-hot encoding. It should be noted here that $\vec{x_A}$, $\vec{x_B}$, and $\vec{x_C}$ are all vectors.

$$\mathbb{X} = \begin{bmatrix} | & | & | \\ \vec{x_A} & \vec{x_B} & \vec{x_C} \\ | & | & | \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

We will show that the fitted coefficients for $\vec{x_A}$, $\vec{x_B}$, and $\vec{x_C}$ are $\bar{y}_A$, $\bar{y}_B$, and $\bar{y}_C$, the average of the y_i values for each of the groups, respectively.

4. Show that the columns of $\mathbb{X}$ are orthogonal, (i.e., the dot product between any pair of column vectors is 0).

5. Show that

$$\mathbb{X}^T\mathbb{X} = \begin{bmatrix} n_A & 0 & 0 \\ 0 & n_B & 0 \\ 0 & 0 & n_C \end{bmatrix}$$

   Here, n_A, n_B, n_C are the number of observations in each of the three groups defined by the levels of the qualitative variable.

6. Show that

$$\mathbb{X}^T\mathbb{Y} = \begin{bmatrix} \sum_{i \in A} y_i \\ \sum_{i \in B} y_i \\ \sum_{i \in C} y_i \end{bmatrix}$$

   where i is an element in group A, B, or C.

7. Use the results from the previous questions to solve the normal equations for $\hat{\theta}$, i.e.,

$$\begin{aligned} \hat{\theta} &= [\mathbb{X}^T\mathbb{X}]^{-1}\mathbb{X}^T\mathbb{Y} \\ &= \begin{bmatrix} \bar{y}_A \\ \bar{y}_B \\ \bar{y}_C \end{bmatrix} \end{aligned}$$

8. *(Extra)* Show that if you augment your $\mathbb{X}$ matrix with an additional $\vec{1}$ bias vector as shown below, $\mathbb{X}^T\mathbb{X}$ is not full rank. Conclude that the new $\mathbb{X}^T\mathbb{X}$ is not invertible, and we cannot use the least squares estimate in this situation.

   Hint: Use the original computation of this matrix from question 6 to help you!

$$\mathbb{X} = \begin{bmatrix} | & | & | & | \\ \vec{1} & \vec{x_A} & \vec{x_B} & \vec{x_C} \\ | & | & | & | \end{bmatrix}$$

Ridge and LASSO Regression

9. Earlier, we posed the linear regression problem as follows: Find the θ value that minimizes the average squared loss. In other words, our goal is to find $\hat{\theta}$ that satisfies the equation below:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, L(\theta) = \underset{\theta}{\operatorname{argmin}}\, \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2$$

Here, $\mathbb{X}$ is a $n \times (p+1)$ matrix, θ is a $(p+1) \times 1$ vector and $\mathbb{Y}$ is a $n \times 1$ vector. Recall that the extra 1 in $(p+1)$ comes from the intercept term. As we saw in lecture, the optimal $\hat{\theta}$ is given by the closed form expression $\hat{\theta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$.

To prevent overfitting, we saw that we can instead minimize the sum of the average squared loss plus a regularization term $\lambda\mathcal{R}(\theta)$.

- If use the function $\mathcal{R}(\theta) = \sum_{j=0}^{p} \theta_j^2 = ||\theta||_2^2$, we have "ridge regression". Recall that $\mathcal{R}$ is the ℓ_2 norm of θ, so this is also referred to as "ℓ_2 regularization".
- If we use the function $\mathcal{R}(\theta) = \sum_{j=0}^{p} |\theta_j| = ||\theta||_1$, we have "LASSO regression". Recall that $\mathcal{R}$ is the ℓ_1 norm of θ, so this is also referred to as "ℓ_1 regularization".

Note that in both of the above formulations, we are regularizing the intercept term to simplify the mathematical formulation of ridge and LASSO regression. In practice, we would not actually want to regularize the intercept term.

For example, if we choose $\mathcal{R}(\theta) = ||\theta||_2^2$, our goal is to find $\hat{\theta}$ that satisfies the equation below:

$$\begin{aligned}\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, L(\theta) &= \underset{\theta}{\operatorname{argmin}}\, \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2 + \lambda||\theta||_2^2 \\ &= \underset{\theta}{\operatorname{argmin}}\, \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbb{X}_{i,\cdot}^T\theta)^2 + \lambda\sum_{j=0}^{p}\theta_j^2\end{aligned}$$

Recall that λ is a hyperparameter that determines the impact of the regularization term. Like ordinary least squares, we can also find a closed form solution to ridge regression: $\hat{\theta} = (\mathbb{X}^T\mathbb{X} + n\lambda\mathbf{I})^{-1}\mathbb{X}^T\mathbb{Y}$. It turns out that $\mathbb{X}^T\mathbb{X} + n\lambda\mathbf{I}$ is guaranteed to be invertible (unlike $\mathbb{X}^T\mathbb{X}$ which might not be invertible).

(a) As model complexity increases, what happens to the bias and variance of the model?

(b) In ridge regression, what happens if we set $\lambda = 0$? What happens as λ approaches ∞?

(c) If we have a large number of features (10,000+) and we suspect that only a handful of features are useful, which type of regression (Lasso vs Ridge) would be more helpful in interpreting useful features?

(d) What are the benefits of using ridge regression over OLS?

Weighted Least Squares for Housing (Bonus)

10. Anirudhan wants to extend his 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, he 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.

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

    (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!

    (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?

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

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

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

    (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