Data 100, Spring 2022	**March 4, 2022**

Discussion #7

Ordinary Debugging

1. Anirudhan is fitting a multiple linear regression model with Scikit-learn, but he is having a few bugs and issues along the way. Help him debug his code and his logic!

   (a) Suppose he runs the code below to fit on design matrix X of shape 250 by 3 with corresponding response variable y of shape 250. We wish to use our model to predict on a new dataset X_t with 50 data points, storing the predictions in a variable `final_predictions`. What are 2 potential issues with this code?

   ```
   model = LinearRegression(fit_intercept = False)
   final_predictions = model.predict(X_t)
   model.fit(X_t, y)
   ```

   (b) Suppose he forgets about the dataset X_t and wishes to focus only on dataset X. Realizing he did not use an intercept term in part (a), he decides to add one using the `add_intercept` function from the lab. What are 2 potential issues with this new code?
   Note: one of these may not break Scikit-learn, but it's an issue nevertheless!

   ```
   def add_intercept(X):
       # Concatenates "ones" vector to design matrix X
       return np.concatenate([X, np.ones(shape = (n, 1))],
                             axis = 1)
   model = LinearRegression()
   n, p = X.shape
   model.fit(add_intercept(X), y)
   final_predictions = model.predict(X)
   ```

Dive into Gradient Descent

2. Given the following loss function and $\vec{x} = [x_i]_{i=1}^n$, $\vec{y} = [y_i]_{i=1}^n$, and $\theta^{(t)}$, explicitly write out the update equation for $\theta^{(t+1)}$ in terms of x_i, y_i, $\theta^{(t)}$, and α, where $\alpha = 0.5$ is the constant learning rate.

$$L(\theta, \vec{x}, \vec{y}) = \frac{1}{n} \sum_{i=1}^{n} \left(\theta^2 x_i - \log(y_i)\right)$$

   Bonus: As $t \to \infty$, what are the required conditions for $\theta^{(t)}$ to converge? To what can it converge?

3. We want to minimize the loss function $L(\theta) = (\theta_1 - 1)^2 + |\theta_2 - 3|$. While you may notice that this function is not differentiable anywhere, we can still use gradient descent wherever the function *is* differentiable!

   Recall that for a function $f(x) = k|x|$, $\frac{df}{dx} = k$ for all $x > 0$ and $\frac{df}{dx} = -k$ for all $x < 0$.

   (a) What are the optimal values $\hat{\theta}_1$ and $\hat{\theta}_2$ to minimize $L(\theta)$? At that point $\hat{\theta}$, what is the gradient ∇L?

   (b) Suppose we initialize our gradient descent algorithm randomly at $\theta_1 = 2$ and $\theta_2 = 5$. Calculate the gradient $\nabla L = \begin{bmatrix} \frac{\partial L}{\partial \theta_1} & \frac{\partial L}{\partial \theta_2} \end{bmatrix}^T \Big|_{\theta_1=2, \theta_2=5}$ at the specified θ_1 and θ_2 values.

   (c) Apply the first gradient update with a learning rate $\alpha = 0.5$. In other words, calculate $\theta_1^{(1)}$ and $\theta_2^{(1)}$ using the initializations $\theta_1^{(0)} = 2$ and $\theta_2^{(0)} = 5$.

(d) How many gradient steps does it take for θ_1 and θ_2 to converge to their optimal values obtained in part (a) assuming we keep a constant learning rate of $\alpha = 0.5$?

Hint: After part (c), what is the derivative $\frac{\partial L}{\partial \theta_1}$ evaluated at $\theta_1^{(1)}$?

The Cook County Housing Dataset

4. In Project 1 we will work with real world housing data from Cook County, Illinois. Analyze the dataframe on the next page and address the following questions:

   (a) Based on the columns presented in this dataset and the values that they take, what do you think each row represents? That is, what is the granularity of this dataset?

   (b) Why do you think this dataset was collected? For what purposes? By whom? This question calls for your speculation and is looking for thoughtfulness, not correctness.

   (c) Certain variables in this dataset contain information that either directly contains demographic information (data on people) or could when linked to other datasets. Identify at least one and explain the nature of the demographic data it embeds.

   (d) Craft at least two questions about housing in Cook County that can be answered with this dataset and provide the type of analytical tool you would use to answer it (e.g. "*I would create a plot of ... and ...*" or "*I would calculate the* [summary statistic] *for ... and ...*"). Be sure to reference the columns that you would use and any additional datasets you would need to answer that question.

Cook County Assessor's Office Dataset (204792 rows x 62 columns): 5 randomly sampled rows

	PIN	Property Class	Neighborhood Code	Land Square Feet	Town Code	Apartments	Wall Material	Roof Material	Basement	Basement Finish	Central Heating	Other Heating
59751	20302160680000	203	121	3750.0	72	0.0	2.0	1.0	1.0	3.0	2.0	5.0
198610	6222130070000	207	13	5640.0	18	0.0	1.0	1.0	2.0	3.0	1.0	5.0
143557	13263020230000	203	70	3125.0	71	0.0	2.0	1.0	1.0	1.0	1.0	5.0
107891	30201100080000	202	101	6600.0	37	0.0	1.0	1.0	1.0	3.0	1.0	5.0
109977	5321220050000	203	150	9864.0	23	0.0	2.0	1.0	1.0	3.0	1.0	5.0

	Central Air	Fireplaces	Attic Type	Attic Finish	Design Plan	Cathedral Ceiling	Construction Quality	Site Desirability	Garage 1 Size	Garage 1 Material	Garage 1 Attachment	Garage 1 Area
	0.0	0.0	3.0	0	0.0	0.0	2.0	2.0	4.0	1.0	2.0	2.0
	1.0	0.0	3.0	0	2.0	0.0	2.0	2.0	3.0	1.0	1.0	1.0
...	1.0	0.0	3.0	0	2.0	0.0	2.0	2.0	3.0	1.0	2.0	2.0
	0.0	0.0	3.0	0	2.0	0.0	2.0	2.0	3.0	1.0	2.0	2.0
	1.0	1.0	2.0	3	2.0	2.0	2.0	2.0	2.0	2.0	1.0	2.0

	Garage 2 Size	Garage 2 Material	Garage 2 Attachment	Garage 2 Area	Porch	Other Improvements	Building Square Feet	Repair Condition	Multi Code	Number of Commercial Units	Estimate (Land)	Estimate (Building)
	7.0	0.0	0.0	0.0	3	0.0	1077.0	2.0	2.0	0.0	22500	70770
	7.0	0.0	0.0	0.0	3	0.0	1902.0	2.0	2.0	0.0	29610	212670
...	7.0	0.0	0.0	0.0	3	0.0	1260.0	2.0	2.0	0.0	43750	254010
	7.0	0.0	0.0	0.0	3	0.0	952.0	2.0	2.0	0.0	21450	61680
	7.0	0.0	0.0	0.0	3	0.0	1307.0	2.0	2.0	0.0	103570	305320

	Deed No.	Sale Price	Longitude	Latitude	Census Tract	Multi Property Indicator	Modeling Group	Age	Use	O'Hare Noise	Floodplain	Road Proximity
	1710934064	1	-87.673073	41.760296	672000.0	0	SF	58	1	0.0	0.0	0.0
	1318416011	173500	-88.188488	42.024901	804310.0	0	SF	24	1	0.0	0.0	0.0
...	1624234087	372500	-87.722925	41.930439	220702.0	0	SF	24	1	0.0	0.0	0.0
	1321946032	56000	-87.535489	41.596277	826202.0	0	SF	56	1	0.0	0.0	0.0
	1801946026	394875	-87.743717	42.072883	800900.0	0	SF	60	1	0.0	0.0	0.0

	Sale Year	Sale Quarter	Sale Half-Year	Sale Quarter of Year	Sale Month of Year	Sale Half of Year	Most Recent Sale	Age Decade	Pure Market Filter	Garage Indicator	Neigborhood Code (mapping)	Town and Neighborhood	Description	Lot Size
	2017	82	41	2	4	1	1.0	5.8	0	1.0	121	72121	This property, sold on 04/19/2017, is a one-st...	3750.0
	2013	67	34	3	7	2	1.0	2.4	1	1.0	13	1813	This property, sold on 07/03/2013, is a two-st...	5640.0
...	2016	79	40	3	8	2	1.0	2.4	1	1.0	70	7170	This property, sold on 08/29/2016, is a one-st...	3125.0
	2013	67	34	3	8	2	1.0	5.6	1	1.0	101	37101	This property, sold on 08/07/2013, is a one-st...	6600.0
	2018	85	43	1	1	1	0.0	6.0	1	1.0	150	23150	This property, sold on 01/19/2018, is a one-st...	9864.0

Dummy Variables/One-hot Encoding (Bonus)

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.

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

6. 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.

7. 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.

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

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

9. *(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!

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