Data100 Sp22 Disc 7
Gradient Descent

Attendance:

https://tinyurl.com/disc7michelle

Announcements

Due Dates

- Lab 7 due Tues, March 8

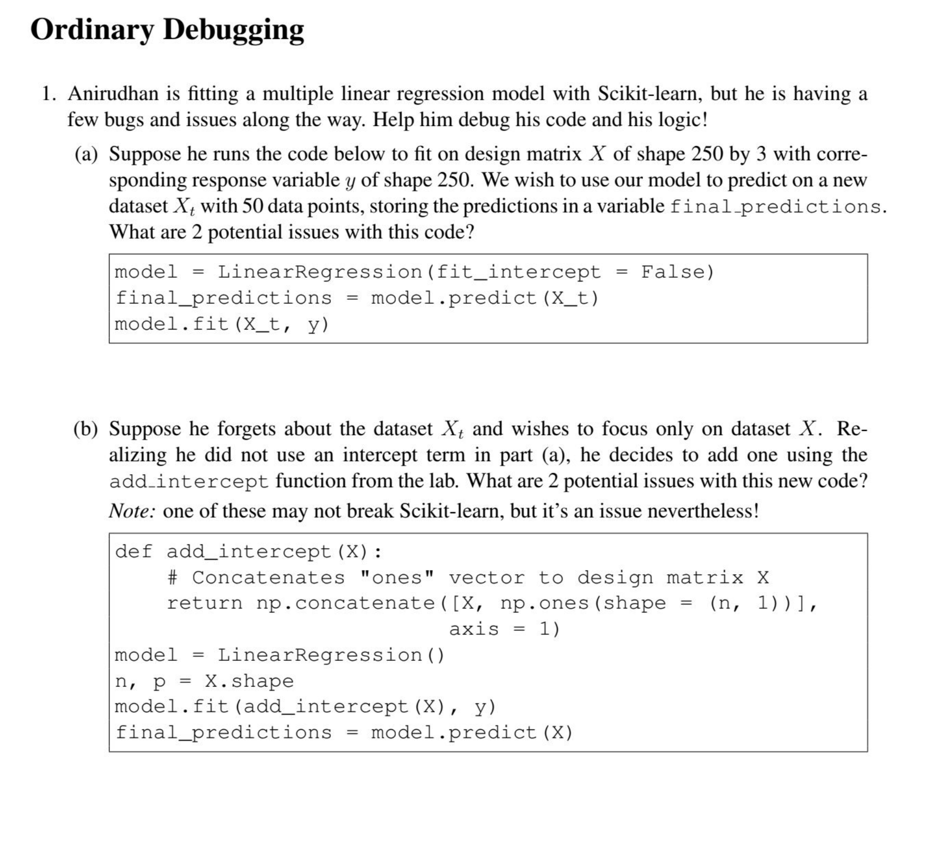
- Proj 1A (Housing) due Thurs, March 10

Other

-Midterm scores are out, regrade requests due Thursday, March 10th, at midnight

Debugging

1



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)
```

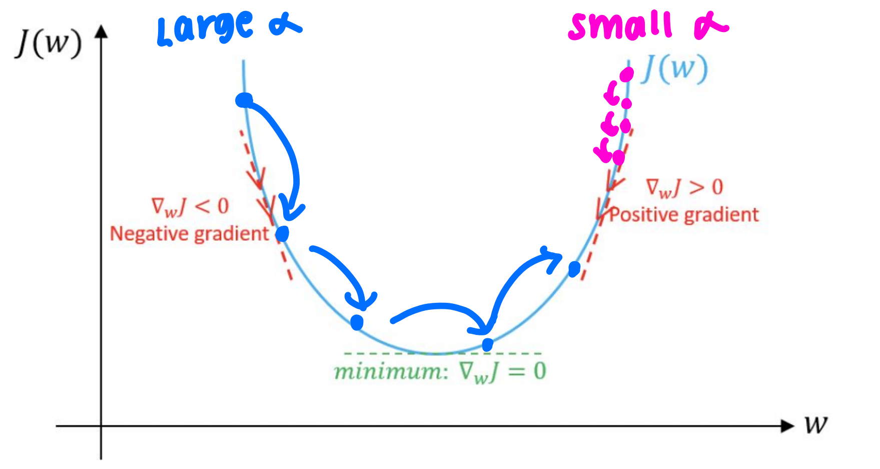
Gradient Descent

2

Why use Gradient Descent?

- So far, we have covered models that have an analytic, closed-form solution
- OLS: $\hat{\theta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$
- This is not always the case - we often need to slowly build our way towards an optimal solution.
- Gradient descent is one such optimization algorithm

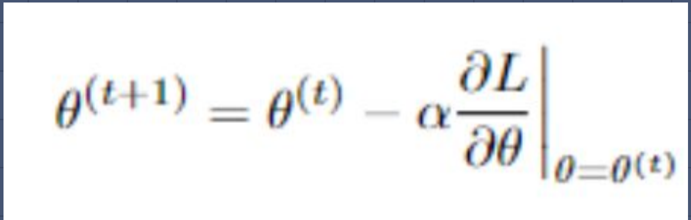
Gradient Descent Overview



Large α: Doesn't converge (sometimes)

small α: Takes long to converge

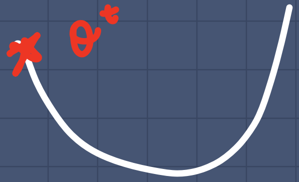
Gradient Descent Algorithm



$$\theta^{(t+1)} = \theta^{(t)} - \alpha \frac{\partial L}{\partial \theta}\bigg|_{\theta=\theta^{(t)}}$$

t : timestep

Gradient Descent Algorithm



$$\theta^{(t+1)} = \theta^{(t)} - \alpha \frac{\partial L}{\partial \theta}\bigg|_{\theta=\theta^{(t)}}$$

← update

← derivative

- Theta at (t+1) = theta at (t) - (learning rate)*gradient of L evaluated at theta(t)

α

Gradient Descent Example

$$\theta^{(t+1)} = \theta^{(t)} - \alpha \frac{\partial L}{\partial \theta}\bigg|_{\theta=\theta^{(t)}}$$

Suppose $$L(\theta, x, y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta x_i)$$

Gradient Descent Example

$$\theta^{(t+1)} = \theta^{(t)} - \alpha \frac{\partial L}{\partial \theta}\Big|_{\theta=\theta^{(t)}}$$

Suppose $L(\theta, x, y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta x_i)$ Find gradient = $\frac{\partial L}{\partial \theta} = \frac{1}{n}\sum_{i=1}^{n}(-x_i)$

Gradient Descent Example

$$\theta^{(t+1)} = \theta^{(t)} - \alpha \frac{\partial L}{\partial \theta}\bigg|_{\theta=\theta^{(t)}}$$

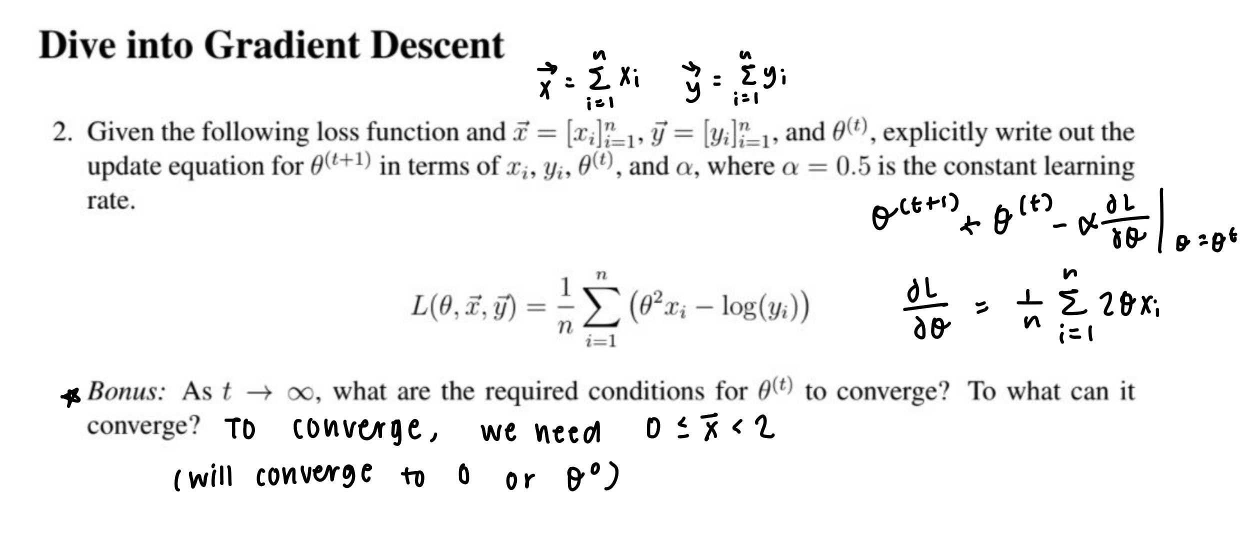
Suppose $L(\theta, x, y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta x_i)$

Find gradient = $\boxed{\frac{\partial L}{\partial \theta} = \frac{1}{n}\sum_{i=1}^{n}(-x_i)}$

Plug into equation to get = $\theta^{(t+1)} = \theta^{(t)} - \alpha\frac{1}{n}\sum_{i=1}^{n}(-x_i)$

How to choose α: cross validation

↑ will learn about later ◡̈



Dive into Gradient Descent

$\vec{x} = \sum_{i=1}^{n} x_i \quad \vec{y} = \sum_{i=1}^{n} y_i$

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.

$\theta^{(t+1)} \leftarrow \theta^{(t)} - \alpha \frac{\partial L}{\partial \theta}\Big|_{\theta=\theta^t}$

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

$\frac{\partial L}{\partial \theta} = \frac{1}{n}\sum_{i=1}^{n} 2\theta x_i$

* *Bonus:* As $t \to \infty$, what are the required conditions for $\theta^{(t)}$ to converge? To what can it converge? To converge, we need $0 \le \bar{x} < 2$

(will converge to 0 or θ^0)

$\theta^{(t+1)} = \theta^t - 0.5 \cdot \frac{1}{n}\sum_{i=1}^{n} 2\theta^{(t)} x_i$

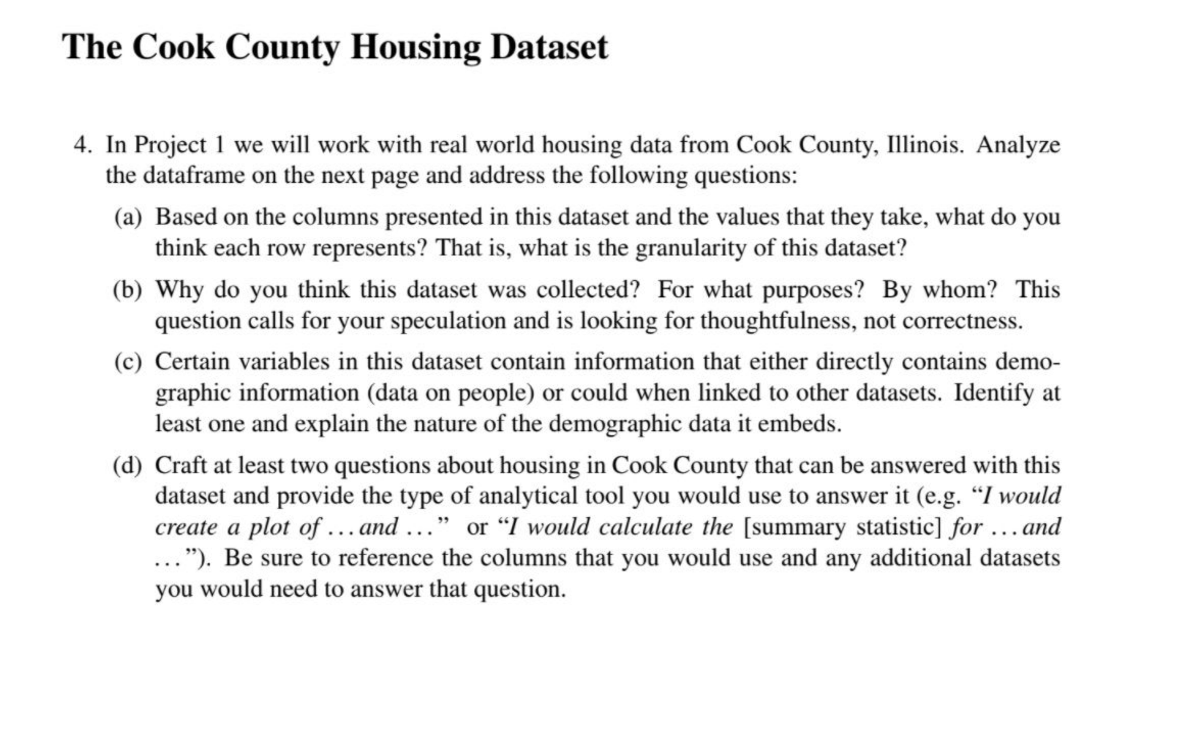
$\theta^{(t+1)} = \theta^t - \frac{1}{n}\sum_{i=1}^{n} \theta^{(t)} x_i$ ← mean

$= \theta^t\left(1 - \frac{1}{n}\sum_{i=1}^{n} x_i\right) = \theta^t(1 - \bar{x})$

$\boxed{\theta^{(t+1)} = \theta^{(t)}(1 - \bar{x})}$

HCE

3



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.