DS 100/200: Principles and Techniques of Data Science **Date: November 13, 2019**

Midterm 2 Review: Linear Regression & Feature Engineering

Name: Raguvir Kunani

1

Fall 2018 Midterm Q10-11: Linear Models

1. Recall from lecture that a linear model is defined as a model where our prediction is given by the equation below, where p is the number of parameters in our model and $\phi(x)$ is some transformation on x:

$$\hat{y} = f_\beta(x) = \sum_{j=1}^{p} \beta_j \phi_j(x)$$

linear means every term in the equation is β_i times some function of x

Which of the following models are linear? **Select all the apply.**

- ☒ A. $f_\beta(x) = \beta_1 x + \beta_2 \sin(x)$
- ☒ B. $f_\beta(x) = \beta_1 x + \beta_2 \sin(x^2)$
- ☒ C. $f_\beta(x) = \beta_1$
- ☒ D. $f_\beta(x) = (\beta_1 x + \beta_2)x$ $= \beta_1 x^2 + \beta_2 x$
- ☐ E. $f_\beta(x) = \ln(\beta_1 x + \beta_2) + \beta_3$

2. Suppose we have data about 5 people shown below:

name	level	trials	phase
Magda	1	10	1
Valerie	5	20	-1
Kumar	2	15	1
Octavia	6	30	1
Dorete	6	5	-1

(a) Suppose we want to model the **level** of each person, and use the following constant model: $f_\beta(\mathbf{x}) = \beta_1$. What is $\hat{\beta}_1$, the value that minimizes the average L2 loss?

$\hat{\beta}_1$ = mean (y) = 4

↑ cheat sheet

(only true for constant model)

(b) We can also compute $\hat{\beta}$ from the previous part by using the normal equation $\hat{\beta} = (\Phi^T\Phi)^{-1}\Phi^T Y$. If we use the normal equation to compute $\hat{\beta}$, how many rows and columns are in the feature matrix Φ? Write your answer in the form **rows** $\times$ **columns**, e.g. 1×1.

$\hat{y} = X\beta$ is our model in general, where X is our feature matrix and β is our parameters. In this case there is only one parameter so β is a scalar. Thus, X must only have 1 column. X has as many rows as there are data points $\Rightarrow$ X has **5 rows × 1 column**

(c) What is $(\Phi^T\Phi)^{-1}\Phi^T Y$ from the previous part? Write your answer in the form of a **Python list**, e.g. [1, 2, 3].

shouldn't be a Y here

we know from (a) that $\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} y_i = 4$. We want $(\Phi^T\Phi)^{-1}\Phi^T y = \frac{1}{n}\sum_{i=1}^{n} y_i$

Note $\vec{1}^T y = \sum_{i=1}^{n} y_i$, where $\vec{1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$. Then $\frac{1}{n}\vec{1}^T y = \frac{1}{n}\sum_{i=1}^{n} y_i$, which is what we want. Thus,

$(\Phi^T\Phi)^{-1}\Phi^T = \frac{1}{n}\vec{1}^T = [1/5, 1/5, 1/5, 1/5, 1/5]$

2

Fall 2018 Midterm Q14: Feature Engineering

3. Recall the *tips* dataset from lab and homework, which contains records about tips, total bills, and information about the person who paid the tip. There are a total of 244 records in **tips**. In addition, you can assume that there are no missing or NaN values in the dataset. The first 5 rows of the **tips** DataFrame are shown below, where *sex* takes on values $\in \{"Male", "Female"\}$, *smoker* takes on values $\in \{"Yes", "No"\}$, *day* takes on values from Monday to Sunday as strings, and *time* takes on values $\in \{"Breakfast", "Lunch", "Dinner"\}$.

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
3	23.68	3.31	Male	No	Sun	Dinner	2
4	24.59	3.61	Female	No	Sun	Dinner	4

(a) Suppose we use `pd.get_dummies` to create a one-hot encoding of only our `sex` column. This yields a feature matrix Φ_{q1} with **exactly 2 columns** `sex_Male`, `sex_Female`, where values can be either 0 or 1 in each column.

Which of the following are true? **Select all that apply.**

☐ A. Φ_{q1} has 244 rows.

☐ B. Φ_{q1} has full column rank.

☐ C. $(\Phi_{q1}^T\Phi_{q1})$ is invertible.

☐ D. None of the above

(b) Suppose we use `pd.get_dummies` to create a one-hot encoding of only our `sex` and `smoker` columns. This yields a feature matrix Φ_{q2} with 4 columns.

Which of the following are true? **Select all that apply.**

- ☐ A. Φ_{q2} has 244 rows.
- ☐ B. Φ_{q2} has full column rank.
- ☐ C. $(\Phi_{q2}^T \Phi_{q2})$ is invertible.
- ☐ D. None of the above

(c) Suppose we use `pd.get_dummies` to create a one-hot encoding of only our `sex` and `smoker` columns, and also include a bias column. This yields a feature matrix Φ_{q3} with 5 columns.

Which of the following are true? **Select all that apply.**

- ☐ A. Φ_{q3} has 244 rows.
- ☐ B. Φ_{q3} has full column rank.
- ☐ C. $(\Phi_{q3}^T \Phi_{q3})$ is invertible.
- ☐ D. None of the above

(d) For the `day` column, we can either use a one-hot encoding or an integer encoding. By integer encoding, we mean mapping Monday to 1, Tuesday to 2, and so on. Which of the following statements are true? **Select all that apply.**

- ☐ A. One-hot encoding creates fewer columns than integer encoding.
- ☐ B. One-hot encoding gives all days of the week the same weight, while integer encoding gives certain days of the week higher weight than others.
- ☐ C. The columns generated by the one-hot encoding of the days of the week are linearly independent of each other.
- ☐ D. None of the above

3

Fall 2018 Midterm Q21-22: Linear Regression

4. Suppose in some universe, the true relationship between the measured luminosity of a single star Y can be written in terms of a single feature ϕ of that same star as

$$Y = \beta^* \phi + \epsilon$$

where $\phi \in \mathbb{R}$ is some non-random scalar feature, $\beta^* \in \mathbb{R}$ is a non-random scalar parameter, and ϵ is a random variable with $\mathbb{E}[\epsilon] = 0$ and $\text{var}(\epsilon) = \sigma^2$. For each star, you have a set of features $\mathbf{\Phi} = \begin{bmatrix} \phi_1 & \phi_2 & \dots & \phi_n \end{bmatrix}^T$ and luminosity measurements $\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \dots & y_n \end{bmatrix}^T$ generated by this relationship. Your Φ may or may not include the feature ϕ described above. The ϵ_i for the various y_i have the same probability distribution and are independent of each other.

(a) What is $\mathbb{E}[Y]$?

$\mathbb{E}[\beta^* \phi + \epsilon] = \mathbb{E}[\beta^* \phi] + \mathbb{E}[\epsilon] = \beta^* \phi + 0$

○ A. 0 ● B. $\beta^* \phi$ ○ C. $\phi(\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi}^T \mathbf{y}$
○ D. β^* ○ E. None of the above

(b) What is $\text{var}(Y)$?

$\text{Var}(\beta^* \phi + \epsilon) = \text{Var}(\epsilon) = \sigma^2$

○ A. $\frac{\sigma^2}{n}$ ○ B. $\frac{\sigma^2}{n^2}$ ○ C. 0
○ D. $\frac{1}{n-1} \sum_{i=1}^{n} \left(y_i - \frac{1}{n} \sum_{i=1}^{n} y_i \right)^2$ ● E. None of the above

(c) Suppose you have information about the exact ϕ value for each star, but try to fit a linear model for Y that includes an intercept term β_0.

$$Y = \beta_0 + \beta_1 \phi$$

Note the true relationship has no intercept term, so our model is not quite correct. Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the values that minimize the average L_2 loss. Let $\mathbf{y}$ be the actual observed data and $\hat{\mathbf{y}} = \hat{\beta}_0 + \hat{\beta}_1 \boldsymbol{\phi}$ be the fitted values.

Cheat sheet

$\sum_i e_i = 0$ when model has intercept

i. Which of the following could possibly be the value of $\hat{\beta}_0$ after fitting our model? **Select all that apply; at least one is correct.**

$\hat{\beta}_0$ could be any value depending on the training data

☒ A. -1 ☒ B. 0 ☒ C. 1 ☒ D. 10

ii. Which of the following could possibly be the residual vector for our model? **Select all that apply; at least one is correct.**

☒ A. $\begin{bmatrix} -2 & -4 & 6 \end{bmatrix}^T$ ☐ B. $\begin{bmatrix} 0.0001 & 0.0003 & -0.0005 \end{bmatrix}^T$
☐ C. $\begin{bmatrix} 3 & 12 & -9 \end{bmatrix}^T$ ☐ D. $\begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^T$

5. Suppose we create a new loss function called the OINK loss, defined as follows for a single observation:

$$L_{OINK}(\beta, x, y) = \begin{cases} a(f_\beta(x) - y) & f_\beta(x) \geq y \\ b(y - f_\beta(x)) & f_\beta(x) < y \end{cases}$$

You decide to use the constant model (given on the left) and average OINK loss (given on the right).

$$f_\beta(x) = \beta \qquad L(\beta, \mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} L_{OINK}(\beta, x_i, y_i)$$

The data are given below. Find the optimal $\hat{\beta}$ that minimizes the loss.

x	3	1	5	4	2	0	6
y	40	0	50	30	20	60	10

(a) when $a = b = 1$

(b) when $a = 1, b = 5$

(c) when $a = 3, b = 6$

See the last page for the solution

.

4

Fall 2018 Final Q26,29-32: Linear Regression

6. What is always true about the residuals in least squares regression? Assume our model includes a bias term. Select all that apply.

Synonymous

- ■ A. They are orthogonal to the column space of the features.
- ■ 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.

↗ L_2 loss, no regularization

7. Which are true about the predictions made by OLS (ordinary least squares, no regularization)? Assume our model includes a bias term. Select all that apply.

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

8. Which of the following would be true if you chose mean absolute error (L1) instead of mean squared error (L2) as your loss function when making a linear model? Select all that apply.

- ☐ A. The results of the regression would be more sensitive to outliers.

can still use gradient descent with L_1 since gradient exists everywhere except the minimum

☐ B. You would not be able to use gradient descent to find the regression line.

■ C. You would not be able to use the normal equation to calculate your parameters.

☐ D. The sum of the residuals would now be zero.

☐ E. None of the above.

9. Let $\hat{y} \in \mathbb{R}^n$ be the vector of fitted values in the ordinary least squares regression of $y \in \mathbb{R}^n$ on the full column-rank feature matrix $\Phi \in \mathbb{R}^{n \times p}$ with n much larger than p. Denote the fitted coefficients as $\hat{\beta} \in \mathbb{R}^p$ and the vector of residuals as $e \in \mathbb{R}^n$.

   (a) What is $\Phi(\Phi^T\Phi)^{-1}\Phi^T y$?

   ◯ A. 0 ◯ B. $\hat{y}$ ◯ C. $\mathbf{e}$ ◯ D. $\hat{\beta}$ ◯ E. 1 ◯ F. None of the above

   (b) What is $\Phi(\Phi^T\Phi)^{-1}\Phi^T \hat{y}$? (Notice the hat in $\hat{y}$)

   ◯ A. 0 ◯ B. $\hat{y}$ ◯ C. $\mathbf{e}$ ◯ D. $\hat{\beta}$ ◯ E. 1 ◯ F. None of the above

   (c) Suppose $\boldsymbol{e} \neq \mathbf{0}$. Define a new feature matrix $\boldsymbol{\Psi}$ by appending the residual vector $\boldsymbol{e}$ to the feature matrix $\boldsymbol{\Phi}$. In other words,

$$\boldsymbol{\Psi} = \begin{bmatrix} | & | & \vdots & | & | \\ \boldsymbol{\Phi}_{:,1} & \boldsymbol{\Phi}_{:,2} & \cdots & \boldsymbol{\Phi}_{:,d} & \boldsymbol{e} \\ | & | & \vdots & | & | \end{bmatrix}$$

   We now want to fit the model $\boldsymbol{y} = \boldsymbol{\Psi\gamma} = \gamma_1\boldsymbol{\Phi}_{:,1} + \gamma_2\boldsymbol{\Phi}_{:,2} + \cdots + \gamma_p\boldsymbol{\Phi}_{:,p} + \gamma_{p+1}\boldsymbol{e}$ by choosing $\hat{\boldsymbol{\gamma}} = [\hat{\gamma}_1 \ldots \hat{\gamma}_{d+1}]^T$ to minimize the L_2 loss. What is $\hat{\gamma}_{p+1}$?

   ◯ A. 0 ◯ B. 1 ◯ C. $\boldsymbol{e}^T\boldsymbol{y}$ ◯ D. $1 - \hat{\boldsymbol{\beta}}^T\hat{\boldsymbol{\beta}}$
   ◯ E. $(\Phi^T\Phi)^{-1}\Phi^T$ ◯ F. None of the above

10. We collect some data $\mathcal{D} = \{(x_1, y_1), ..., (x_n, y_n)\}$ and decide to model the relationship between $\boldsymbol{X}$ and $\boldsymbol{y}$ as

$$\boldsymbol{y} = \beta_1\boldsymbol{\Phi}_{:,1} + \beta_2\boldsymbol{\Phi}_{:,2}$$

    where $\boldsymbol{\Phi}_{i,:} = \begin{bmatrix} 1 & x_i \end{bmatrix}$ We found the estimates $\hat{\beta}_1 = 2$ and $\hat{\beta}_2 = 5$ for the coefficients by minimizing the L_2 loss. Given that $\boldsymbol{\Phi}^T\boldsymbol{\Phi} = \begin{bmatrix} 4 & 2 \\ 2 & 5 \end{bmatrix}$, answer the following problems. If not enough information is given, write "Cannot be determined."

    (a) What was the sample size n? Hint: Consider the form of the feature matrix.

    (b) What must $\boldsymbol{\Phi}^T\boldsymbol{y}$ be for this data set?

OINK Loss Problem

$$L_{OINK}(\beta, x, y) = \begin{cases} a(f_\beta(x) - y) & f_\beta(x) \geq y \\ b(y - f_\beta(x)) & f_\beta(x) < y \end{cases}$$

$\Rightarrow$ always positive, so this almost L_1 loss

(i) $a = b = 1 \Rightarrow L_{OINK}$ becomes L_1 loss! Our goal is to find which value of β minimizes L_1 loss. What follows is a derivation of the minimizing β. Define the data as (x_i, y_i). Our model is constant, so it always predicts β regardless of the input x.

$$\text{Avg. } L_1 \text{ loss} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \beta|$$

$$\frac{d}{d\beta} \text{Avg. } L_1 \text{ loss} = \frac{d}{d\beta} \left[\frac{1}{n} \sum_{i=1}^{n} |y_i - \beta| \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{d}{d\beta} \left[|y_i - \beta| \right]$$

split $|x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0 \end{cases}$

$$= \frac{1}{n} \sum_{i=1}^{n} \begin{cases} \frac{d}{d\beta}[y_i - \beta] & \text{if } y_i \geq \beta \\ \frac{d}{d\beta}[\beta - y_i] & \text{if } y_i < \beta \end{cases}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \begin{cases} -1 & \text{if } y_i \geq \beta \\ 1 & \text{if } y_i < \beta \end{cases}$$

To minimize, we set the derivative equal to 0 and solve.

$$\frac{d}{d\beta} \text{Avg. } L_1 \text{ loss} = \frac{1}{n} \sum_{i=1}^{n} \begin{cases} -1 & \text{if } y_i \geq \beta \\ 1 & \text{if } y_i < \beta \end{cases} = 0$$

$$\sum_{i=1}^{n} \begin{cases} -1 & \text{if } y_i \geq \beta \\ 1 & \text{if } y_i < \beta \end{cases} = 0$$

$$\sum_{y_i \geq \beta} -1 + \sum_{y_i < \beta} 1 = 0$$

$$\sum_{y_i < \beta} 1 = -\sum_{y_i \geq \beta} -1$$

$$\sum_{y_i < \beta} 1 = \sum_{y_i \geq \beta} 1$$

the minimizing β satisfies this condition.

$$\left(\#\, y_i < \beta\right) = \left(\#\, y_i \geq \beta\right)$$

By definition, $\beta = \text{median}(y)$ is the only β that satisfies the above condition. Thus, $\hat{\beta} = \text{median}(y)$ is the minimizing β.

For our dataset, the median of y is 30,

so $\boxed{\hat{\beta} = 30}$.

(ii) $a = 1$, $b = 5$

$$L_{OINK} = \begin{cases} f_\beta(x) - y & \text{if } f_\beta(x) \geq y \\ 5(y - f_\beta(x)) & \text{if } f_\beta(x) < y \end{cases}$$

Want to minimize Avg. OINK Loss:

$$\frac{d}{d\beta}\left[\text{Avg. OINK Loss}\right] = \frac{d}{d\beta}\left[\frac{1}{n}\sum_{i=1}^{n} \text{OINK Loss}\right]$$

↓ plug in OINK loss

$$= \frac{d}{d\beta}\left[\frac{1}{n}\sum_{i=1}^{n} \begin{cases} f_\beta(x_i) - y_i & \text{if } f_\beta(x) \geq y_i \\ 5(y_i - f_\beta(x_i)) & \text{if } f_\beta(x) < y_i \end{cases}\right]$$

↓ plug in constant model $f_\beta(x) = \beta$

$$= \frac{d}{d\beta}\left[\frac{1}{n}\sum_{i=1}^{n} \begin{cases} \beta - y_i & \text{if } \beta \geq y_i \\ 5(y_i - \beta) & \text{if } \beta < y_i \end{cases}\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} \begin{cases} \frac{d}{d\beta}[\beta - y_i] & \text{if } \beta \geq y_i \\ \frac{d}{d\beta}[5(y_i - \beta)] & \text{if } \beta < y_i \end{cases}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \begin{cases} 1 & \text{if } \beta \geq y_i \\ -5 & \text{if } \beta < y_i \end{cases}$$

To minimize, set the derivative to 0 and solve:

$$\frac{1}{n}\sum_{i=1}^{n} \begin{cases} 1 & \text{if } \beta \geq y_i \\ -5 & \text{if } \beta < y_i \end{cases} = 0$$

$$\sum_{i=1}^{n} \begin{cases} 1 & \text{if } \beta \geq y_i \\ -5 & \text{if } \beta < y_i \end{cases} = 0$$

$$\sum_{\beta \geq y_i} 1 + \sum_{\beta < y_i} -5 = 0$$

$$\sum_{\beta \geq y_i} 1 = -\sum_{\beta < y_i} -5$$

$$\sum_{\beta \geq y_i} 1 = 5 \sum_{\beta < y_i} 1$$

Minimizing β satisfies this equation

$$(\#\, y_i \leq \beta) = 5\,(\#\, y_i > \beta)$$

This implies that the minimizing β is such that the number of y_i less than or equal to β is 5 times the number of y_i greater than β. For our dataset, the value of β that achieves this is $\boxed{\hat{\beta} = 50}$.

(iii) Same process as (ii), but in the end you'll get:

$$3\,(\#\, y_i < \beta) = 6\,(\#\, y_i > \beta)$$

$$(\#\, y_i < \beta) = 2\,(\#\, y_i > \beta)$$

The value of β for our dataset that satisfies this equation is $\boxed{\hat{\beta} = 40}$.