Homework 2 - Written Answer Key

Question 1:

(Question)

Given $\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}$, and the following function $g(\mathbf{w}) = \frac{1}{2}(w_0 + w_1 \cdot x_1 + w_2 \cdot x_2 - y)^2$

what is the gradient $\nabla_{\mathbf{w}} g(\mathbf{w}) = \begin{bmatrix} \frac{\partial g(\mathbf{w})}{\partial w_0} \\ \frac{\partial g(\mathbf{w})}{\partial w_1} \\ \frac{\partial g(\mathbf{w})}{\partial w_2} \end{bmatrix}$?

(Answer(s))

Chain Rule: $f'(g(x)) * g'(x)$

$$\frac{\partial g(w)}{\partial w_0} = 2 * \frac{1}{2}(w_0 + w_1 x_1 + w_2 x_2 - y) * 1$$

$$\frac{\partial g(w)}{\partial w_1} = 2 * \frac{1}{2}(w_0 + w_1 x_1 + w_2 x_2 - y) * x_1$$

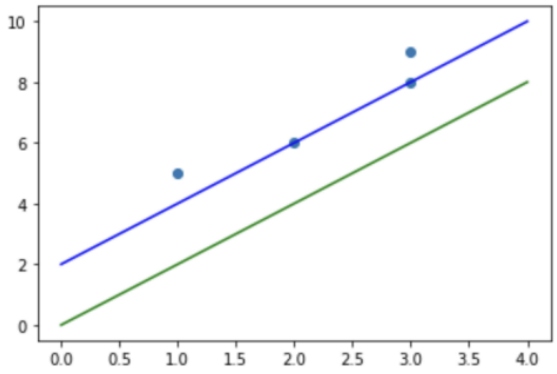
$$\frac{\partial g(w)}{\partial w_2} = 2 * \frac{1}{2}(w_0 + w_1 x_1 + w_2 x_2 - y) * x_2$$

$$\nabla_w g(w) = \begin{bmatrix} w_0 + w_1 x_1 + w_2 x_2 - y \\ (w_0 + w_1 x_1 + w_2 x_2 - y) * x_1 \\ (w_0 + w_1 x_1 + w_2 x_2 - y) * x_2 \end{bmatrix}$$

Question 2:

(Question)

Consider data $(1,5),(2,6),(3,8),(3,9)$ and regression lines: $y = 2x_1$ (the green line), $y = 2x_1 + 2$ (the blue line). (Note here $\mathbf{x} = [x_1]$.)



- What is the squared error of each of the points[1] with respect to the line $y = 2x_1$?
- The gradient of our cost function includes a sum over contributions of individual points. We could calculate the individual contributions separately. The gradient for a single $(\mathbf{x}^{(i)}, y^{(i)})$ point is:[2] $\begin{bmatrix} (w_0 + w_1 \cdot x_1^{(i)} - y^{(i)}) \\ (w_0 + w_1 \cdot x_1^{(i)} - y^{(i)}) x_1^{(i)} \end{bmatrix}$
  For the line $y = 2x_1$ what is the gradient contribution for each of the four examples?[3]
- What is the squared error of each of the points with respect to the line $y = 2x_1 + 2$? (the blue line in the graph)[4]
- What is the gradient contribution for each of the four examples to the line $y = 2x_1 + 2$?
- Which line has a smaller RSS?
- Would it be possible for a different line to have a smaller RSS?

(Answer(s))

1.

$$y = 2x$$

$$\text{Points: } (1,5), (2,6), (3,8), (3,9)$$

$$(1,5) : \hat{y} = 2(1) = 2$$

$$(\hat{y} - y)^2 = ((2) - (5))^2 = 9$$

$$(2,6) : \hat{y} = 2(2) = 4$$

$$(\hat{y} - y)^2 = ((4) - (6))^2 = 4$$

$$(3,8) : \hat{y} = 2(3) = 6$$

$$(\hat{y} - y)^2 = ((6) - (8))^2 = 4$$

$$(3,9) : \hat{y} = 2(3) = 6$$

$$(\hat{y} - y)^2 = ((6) - (9))^2 = 9$$

2.

$$y = 2x$$

$$GC = \begin{bmatrix} (w_0 + w_1 \cdot x_1^{(i)} - y^{(i)}) \\ (w_0 + w_1 \cdot x_1^{(i)} - y^{(i)}) x_1^{(i)} \end{bmatrix}$$

$$(1,5) : \begin{bmatrix} 2(1) - 5 \\ (2(1) - 5) * 1 \end{bmatrix} = \begin{bmatrix} -3 \\ -3 \end{bmatrix}$$

$$(2,6) : \begin{bmatrix} 2(2) - 6 \\ (2(2) - 6) * 2 \end{bmatrix} = \begin{bmatrix} -2 \\ -4 \end{bmatrix}$$

$$(3,8) : \begin{bmatrix} 2(3) - 8 \\ (2(3) - 8) * 3 \end{bmatrix} = \begin{bmatrix} -2 \\ -6 \end{bmatrix}$$

$$(3,9) : \begin{bmatrix} 2(3) - 9 \\ (2(3) - 9) * 3 \end{bmatrix} = \begin{bmatrix} -3 \\ -9 \end{bmatrix}$$

3.

$$y = 2x + 2$$

$$\text{Points: } (1,5), (2,6), (3,8), (3,9)$$

$$(1,5) : \hat{y} = 2(1) + 2 = 4$$

$$(\hat{y} - y)^2 = ((4) - (5))^2 = 1$$

$$(2,6) : \hat{y} = 2(2) + 2 = 6$$

$$(\hat{y} - y)^2 = ((6) - (6))^2 = 0$$

$$(3,8) : \hat{y} = 2(3) + 2 = 8$$

$$(\hat{y} - y)^2 = ((8) - (8))^2 = 0$$

$$(3,9) : \hat{y} = 2(3) + 2 = 8$$

$$(\hat{y} - y)^2 = ((8) - (9))^2 = 1$$

4.

$$y = 2x + 2$$

$$GC = \begin{bmatrix} (w_0 + w_1 \cdot x_1^{(i)} - y^{(i)}) \\ (w_0 + w_1 \cdot x_1^{(i)} - y^{(i)}) x_1^{(i)} \end{bmatrix}$$

$$(1,5) : \begin{bmatrix} 2(1) + 2 - 5 \\ (2(1) + 2 - 5) * 1 \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$$

$$(2,6) : \begin{bmatrix} 2(2) + 2 - 6 \\ (2(2) + 2 - 6) * 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$(3,8) : \begin{bmatrix} 2(3) + 2 - 8 \\ (2(3) + 2 - 8) * 3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$(3,9) : \begin{bmatrix} 2(3) + 2 - 9 \\ (2(3) + 2 - 9) * 3 \end{bmatrix} = \begin{bmatrix} -1 \\ -3 \end{bmatrix}$$

5.

$$y = 2x$$

$$RSS = 9 + 4 + 4 + 9 = 26$$

$$y = 2x + 2$$

$$RSS = 1 + 0 + 0 + 1 = 2$$

Line 2 ($y = 2x + 2$) has a smaller RSS.

6.

Yes, it is possible for a line to have a smaller RSS. Imagine a line similar to the blue line ($y = 2x + 2$) except its intercept is slightly higher. Such a line, if it was still below (1, 5) and (3, 9) but above (2, 6) and (3, 8), would have a smaller RSS.

Question 3:

(Question)

Given the following data $((x_1, x_2)^T, y)$: $((0,0)^T, 1), ((0,1)^T, 4), ((1,0)^T, 3), ((1,1)^T, 7)$

- create the design matrix X (include the column of 1's)
- create the target vector y
- write out the closed form solution for computing $\mathbf{w}$ that we discussed in class.
- compute w_0, w_1, w_2
- compute RSS
- compute TSS
- compute R^2
- what portion of variance in y explained by $\mathbf{x}$?
- predict the value of $\mathbf{x}^T = (0.5, 0.5)$

(Answer(s))

1. $$X: \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

2. $$Y: \begin{bmatrix} 1 \\ 4 \\ 3 \\ 7 \end{bmatrix}$$

3. $$w = (X^T X)^{-1} X^T y$$

4.

$$w = \left(\begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 4 \\ 3 \\ 7 \end{bmatrix}$$

$$w = \begin{bmatrix} 0.75 & -0.5 & -0.5 \\ -0.5 & 1 & 0 \\ -0.5 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 15 \\ 10 \\ 11 \end{bmatrix}$$

$$w = \begin{bmatrix} 0.75 \\ 2.5 \\ 3.5 \end{bmatrix} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}$$

5.

$$RSS = \sum_{i=1}^{N} (y^{(i)} - \hat{y}^{(i)})^2$$

$$RSS = \sum_{i=1}^{4} (y^{(i)} - (0.75 + 2.5x_1^{(i)} + 3.5x_2^{(i)}))^2$$

$$RSS = 0.25$$

6.

$$TSS = \sum_{i=1}^{N} (y^{(i)} - \overline{y})^2$$

$$\overline{y} = \frac{1 + 4 + 3 + 7}{4} = 3.75$$

$$TSS = \sum_{i=1}^{4} (y^{(i)} - 3.75)^2$$

$$TSS = 18.75$$

7.

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{0.25}{18.75} = 0.9867$$

8.

98.67% of the variance in y can be explained by x (because the percent of variance in y that can be explained by x is the same as R^2).

9.

$$\hat{y}^{(0.5,0.5)} = 0.75 + 2.5(0.5) + 3.5(0.5) = 3.75$$

Question 4:

(Question)

Suppose you were interested in crop yields and you had collected data on the amount of rainfall, the amount of fertilizer, the average temperature, and the number of sunny days.

How could you formalize this as a regression problem?

(Answer(s))

Data: the data that was collected on the amount of rainfall, the amount of fertilizer, the average temperature, and the number of sunny days makes up the X matrix, and the crop yield will be the y vector. Note that a column of 1's should be appended to the left of the X matrix for subsequent calculations.

$$X = \begin{bmatrix} (1's) & rainfall & fertilizer & avg\_temp & sunny\_days \end{bmatrix}$$

$$y = \begin{bmatrix} crop\_yield \end{bmatrix}$$

Model: the model that will be a multiple linear regression model.

$$\hat{y}^{(i)} = w_0 + w_1 x_1^{(i)} + w_2 x_2^{(i)} + w_3 x_3^{(i)} + w_4 x_4^{(i)}$$

Loss Function: the Loss Function will still be RSS

$$RSS = \sum_{i=1}^{N} (y^{(i)} - \hat{y}^{(i)})^2$$

$$RSS = \sum_{i=1}^{4} (y^{(i)} - (w_0 + w_1 x_1^{(i)} + w_2 x_2^{(i)} + w_3 x_3^{(i)} + w_4 x_4^{(i)}))^2$$

Minimizing Loss: a closed-form solution can be used to compute w.

$$w = (X^T X)^{-1} X^T y$$

Question 5:

(Question)

For the following function: $f_1(w_0, w_1) = (w_0 + 2w_1 - 4)^2 + (w_0 + 3w_1 - 3)^2$ run the gradient descent algorithm for `num_iters` $= 10$ iterations (you can use your computer to perform the calculations) where you try different learning rates. For each start with $(w_0, w_1) = (0, 0)$:

(a) learning rate of $\alpha = 0.06$

(b) learning rate of $\alpha = 0.001$

(c) learning rate of $\alpha = 0.03$

Report the value of w_0, w_1 and $f(w_0, w_1)$ at the end of each step. On one graph, plot the points (w_0, w_1) at every iteration.

Evaluate (briefly in one sentence) how each learning rate contributed or did not contribute to finding a new assignment to the parameters that decreased the value of the function.

(Answer(s))

Code:

```python
import matplotlib.pyplot as plt

decimals = 6
num_iters = 10
alphas = [0.06, 0.001, 0.03]
colors = ['red', 'blue', 'green']

for i, alpha in enumerate(alphas):
    print(f"For alpha: {alpha}")
    w0, w1 = 0, 0

    # Create lists of w0, w1 values for plotting later
    w0_list, w1_list = [], []

    for _ in range(num_iters):
        # Derivative of f1 with respect to w0 is:
        # 2(w0 + 2w1 - 4) + 2(w0 + 3w1 - 3)
        # = 4w0 + 10w1 - 14
        temp0 = w0 - (alpha * (4 * w0 + 10 * w1 - 14))

        # Derivative of f1 with respect to w1 is:
        # (2)(2)(w0 + 2w1 - 4) + (2)(3)(w0 + 3w1 - 3)
        # = 10w0 + 26w1 - 34
        temp1 = w1 - (alpha * (10 * w0 + 26 * w1 - 34))

        w0, w1 = temp0, temp1

        w0_list.append(temp0)
        w1_list.append(temp1)

        # Calculate the value of the function f1 at every iteration
        f1 = (w0 + 2 * w1 - 4) ** 2 + (w0 + 3 * w1 - 3) ** 2

        print(f"w0: {round(w0, decimals)}; w1: {round(w1, decimals)}; f1: {round(f1, decimals)}")

    # Add a newline between alphas during the print
    print()

    # Plot the w1, w0 pairs
    plt.plot(w0_list, w1_list, 'o', color=colors[i], label=f"alpha: {alpha}")

# Plot the w0, w1 pairs
plt.title("w0 vs w1")
plt.legend()
plt.grid()
plt.show()
```

1.

```
For alpha: 0.06
w0: 0.84; w1: 2.04; f1: 16.528
w0: 0.2544; w1: 0.3936; f1: 11.20073
w0: 0.797184; w1: 1.666944; f1: 7.846073
w0: 0.445693; w1: 0.628201; f1: 5.728869
w0: 0.801806; w1: 1.420791; f1: 4.388013
w0: 0.596898; w1: 0.763273; f1: 3.534296
w0: 0.835679; w1: 1.254428; f1: 2.986319
w0: 0.722459; w1: 0.836113; f1: 2.630305
w0: 0.887401; w1: 1.138301; f1: 2.394889
w0: 0.831444; w1: 0.870111; f1: 2.235306
```

2.

```
For alpha: 0.001
w0: 0.014; w1: 0.034; f1: 23.66818
w0: 0.027604; w1: 0.066976; f1: 22.414687
w0: 0.040824; w1: 0.098959; f1: 21.234913
w0: 0.053671; w1: 0.129977; f1: 20.124519
w0: 0.066156; w1: 0.160061; f1: 19.079424
w0: 0.078291; w1: 0.189238; f1: 18.095786
w0: 0.090086; w1: 0.217535; f1: 17.169987
w0: 0.10155; w1: 0.244978; f1: 16.298626
w0: 0.112694; w1: 0.271593; f1: 15.478498
w0: 0.123527; w1: 0.297405; f1: 14.706589
```

3.

```
For alpha: 0.03
w0: 0.42; w1: 1.02; f1: 2.602
w0: 0.4836; w1: 1.1184; f1: 2.340962
w0: 0.510048; w1: 1.120968; f1: 2.319589
w0: 0.532552; w1: 1.113599; f1: 2.300959
w0: 0.554566; w1: 1.105226; f1: 2.282505
w0: 0.57645; w1: 1.09678; f1: 2.2642
w0: 0.598242; w1: 1.088356; f1: 2.246042
w0: 0.619946; w1: 1.079966; f1: 2.228029
w0: 0.641563; w1: 1.071609; f1: 2.210161
w0: 0.663093; w1: 1.063285; f1: 2.192436
```

4.

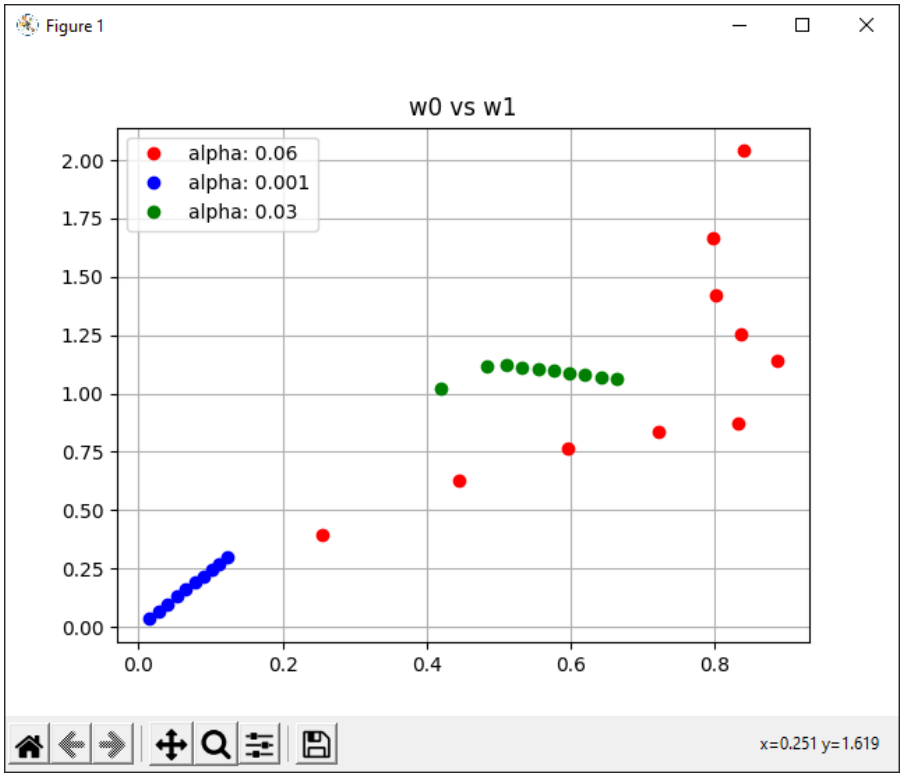


Figure 1
w0 vs w1
alpha: 0.06
alpha: 0.001
alpha: 0.03
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0.0
0.2
0.4
0.6
0.8
x=0.251 y=1.619

Question 6:

(Question)

Given the following data matrix:

$$X = \begin{bmatrix} 1 & small & Chevy & 130 \\ 1 & large & Buick & 165 \\ 1 & medium & Plymouth & 150 \\ 1 & medium & Ford & 140 \\ 1 & small & Ford & 198 \\ 1 & medium & Chevy & 150 \\ 1 & large & Buick & 225 \end{bmatrix}$$

Perform On-hot encoding on the third feature, and perform an ordinal encoding on the second feature. In your answer provide the transformed data matrix.

(Answer(s))

For one-hot encoding the third feature, there are a total of 4 possible categories: "Chevy," "Buick," "Plymouth," and "Ford." Thus, this single feature will be split into 4 total features, where each sample will have a "1" in the category that applies to it, and "0" to the other three.
For ordinal encoding the second feature, there are three total sizes: "small," "medium," and "large." We will turn this feature into an integer with "0" representing "small," "1" representing "medium," and "2" representing "large."
After applying these changes, the transformed data matrix will look like this:

$$X = \begin{bmatrix} \text{Ones} & \text{Size} & \text{Chevy} & \text{Buick} & \text{Plymouth} & \text{Ford} & \text{Numbers} \\ 1 & 0 & 1 & 0 & 0 & 0 & 130 \\ 1 & 2 & 0 & 1 & 0 & 0 & 165 \\ 1 & 1 & 0 & 0 & 1 & 0 & 150 \\ 1 & 1 & 0 & 0 & 0 & 1 & 140 \\ 1 & 0 & 0 & 0 & 0 & 1 & 198 \\ 1 & 1 & 1 & 0 & 0 & 0 & 150 \\ 1 & 2 & 0 & 1 & 0 & 0 & 225 \end{bmatrix}$$

Question 7:

(Question)

For linear regression, on a data set $X = \{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(N)}, y^{(N)})\}$, if the ith feature for every example is scaled by a constant c , does $\mathbf{w}$ change? If it does change, describe how.

(Answer(s))

Multiplying the ith feature of every example by a constant c will change w, but will not affect the predictions. To illustrate this, first consider the formula for w_1 that was covered in simple linear regression.

$$w_1 = \frac{s_{xy}}{s_{xx}} = \frac{\sum_{i=1}^{N}(x_1^{(i)} - \bar{x_1})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_1^{(i)} - \bar{x_1})^2}$$

If x_1 is multiplied by c, then this happens:

$$w_1' = \frac{s_{xy}}{s_{xx}} = \frac{\sum_{i=1}^{N}(c * x_1^{(i)} - c * \bar{x_1})(y_i - \bar{y})}{\sum_{i=1}^{N}(c * x_1^{(i)} - c * \bar{x_1})^2}$$

$$w_1' = \frac{c * \sum_{i=1}^{N}(x_1^{(i)} - \bar{x_1})(y_i - \bar{y})}{c^2 * \sum_{i=1}^{N}(x_1^{(i)} - \bar{x_1})^2}$$

$$w_1' = \frac{w_1}{c}$$

This denominator c cancels out with the c that the feature x_1 is multiplied by, so ultimately, the prediction $(w \cdot x)$ remains the same, and the feature x_i is divided by c. This can be generalized for any feature in linear regression.

Question 8:

(Question)

An online retailer like Amazon wants to determine which products to promote based on reviews. They only want to promote products that are likely to sell. For each product, they have past sales as well as reviews. The reviews have both a numeric score (from 1 to 5) and text.

(a) To formulate this as a machine learning problem, suggest a target variable that the online retailer could use.

(b) For the predictors of the target variable, a data scientist suggests to combine the numeric score with frequency of occurrence of words that convey judgement like "bad", "good", and "doesn't work." Describe a possible linear model for this relation.

(c) Now, suppose that some reviews have a numeric score from 1 to 5 and others have a score from 1 to 10. How would change your features?

(d) Now suppose the reviews have either (a) a score from 1 to 5; (b) a rating that is simply good or bad; or (c) no numeric rating at all. How would you change your features?

(e) For the frequency of occurrence of a word such as "good", which variable would you suggest to use as a predictor: (a) total number of reviews with the word "good"; or (b) fraction of reviews with the word "good"?

(Answer(s))

A.

A target variable that the online retailer could use is the percent of customers who purchase a given product after opening its product page (represented as a decimal in the range [0.00, 1.00])

B.

A possible linear model for this relation could be:

$$y = x_0 + x_1 - x_2$$

where y is the the combined score, x_0 is the numeric score of the review, x_1 number of "good judgement" words, and x_2 is the number of "bad judgement" words.

C.

If some reviews were on the 1-5 scale and others were on the 1-10 scale, I would multiply all the reviews on the 1-5 scale by 2, effectively converting them to the 1-10 scale. Then, I'd approach the problem as per usual.

D.

(a) If the reviews were scores from 1-5, I would simply convert the reviews to an integer feature.
(b) If the reviews were a good/bad rating, I would convert the reviews to a binary feature where "1" represents a "good" review and "0" represents a "bad" review.
(c) If the reviews had no numeric rating, the only features I would be able to use to predict the target variable would have to come from the text in the review. I could look for the number of occurrences of positive words, such as "good" and "helpful," and then subtract this number by the number of occurrences of negative words, such as "bad" and "broken" to create an integer value that the model can use as a normal feature.

E.

I would suggest the (b) variable that considers the fraction of reviews with the word "good." This is because some products will naturally get more exposure if they're household requirements, so the volume of reviews will be higher. Thus, the sheer number of positive reviews may be high, but could very well be out-ratio'ed by the negative reviews.