Midterm Exam - DSC 40A, Fall 2022

Instructions

- This exam consists of 5 questions (including Question 0), worth a total of 100 Points. **Advice: Read all of the problems before starting to work.**
- You have **50 minutes** to answer all the questions.
- **This exam is open-book, but collaboration is strictly forbidden.**
- Show your work for all questions. Correct answers with no work shown may not receive full credit.
- Please write neatly.

---

0. Statement of Academic Integrity[1 Point]

Please copy, sign, and date the following statement on your exam page. You must do this even if you are doing the exam on a separate piece of paper. **Your exam will not be graded if you do not complete this question.**

As a member of the UC San Diego community, I act with honesty, integrity, and respect for others. I will neither give nor receive assistance while taking this exam.

1. Gradient Descent [20 Points]

Suppose that we are given $f(x) = x^3 + x^2$ and learning rate $\alpha = 1/4$.

a) [5 Points] First of all, write down the updating rule for gradient descent in general and for this function.

Solution: In general, the updating rule for gradient descent is:

$$x_{i+1} = x_i - \alpha \nabla f(x_i) = x_i - \alpha \frac{\partial f}{\partial x}(x_i),$$

where $\alpha \in \mathbb{R}_+$ is the learning rate or step size. For this function, since f is a single-variable function, we can write down the updating rule as:

$$x_{i+1} = x_i - \alpha \frac{df}{dx}(x_i) = x_i - \alpha f'(x_i).$$

We also have:

$$\frac{df}{dx} = f'(x) = 3x^2 + 2x,$$

thus the updating rule can be written down as:

$$x_{i+1} = x_i - \alpha(3x_i^2 + 2x_i) = -\frac{3}{4}x_i^2 + \frac{1}{2}x_i.$$

b) [5 Points] If we start at $x_0 = -1$, should we go left or right? Can you verify mathematically? What is x_1? Can gradient descent converge? If so, where it might converge to, given appropriate step size?

Solution: We have

$$f'(x_0) = f'(-1) = 3(-1)^2 + 2(-1) = 1 > 0,$$

so we go left, and

$$x_1 = x_0 - \alpha f'(x_0) = -1 - \frac{1}{4} = -\frac{5}{4}.$$

Intuitively, the gradient descent cannot converge in this case because

$$\lim_{x \to -\infty} f(x) = -\infty,$$

We need to find all local minimums and local maximums. First, we solve the equation $f'(x) = 0$ to find all critical points. We have:

$$f'(x) = 0 \Leftrightarrow 3x^2 + 2x = 0 \Leftrightarrow x = -\frac{2}{3} \text{ and } x = 0.$$

Now, we consider the second-order derivative:

$$f''(x) = \frac{d^2 f}{dx^2} = 6x + 2.$$

We have $f''(x) = 0$ only when $x = -1/3$. Thus, for $x < -1/3$, $f''(x)$ is negative or the slope $f'(x)$ decreases; and for $x > -1/3$, $f''(x)$ is positive or the slope $f'(x)$ increases. Keep in mind that $-1 < -2/3 < -1/3 < 0 < 1$. Therefore, f has a local maximum at $x = -2/3$ and a local minimum at $x = 0$. If the gradient descent starts at $x_0 = -1$ and it always goes left then it will never meet the local minimum at $x = 0$, and it will go left infinitely. We say the gradient descent cannot converge, or is divergent.

c) [5 Points] If we start at $x_0 = 1$, should we go left or right? Can you verify mathematically? What is x_1? Can gradient descent converge? If so, where it might converge to, given appropriate step size?

Solution: We have

$$f'(x_0) = f'(-1) = 3 \cdot 1^2 + 2 \cdot 1 = 5 > 0,$$

so we go left, and

$$x_1 = x_0 - \alpha f'(x_0) = 1 - \frac{1}{4} \cdot 5 = -\frac{1}{4}.$$

From the previous part, function f has a local minimum at $x = 0$, so the gradient descent can converge (given appropriate step size) at this local minimum.

d) [5 Points] Write down 1 condition to terminate the gradient descent algorithm (in general).

Solution: There are several ways to terminate the gradient descent algorithm:

- If the change in the optimization objective is too small, i.e. $|f(x_i) - f(x_{i+1})| < \epsilon$ where ϵ is a small constant,
- If the gradient is close to zero or the norm of the gradient is very small, i.e. $\|\nabla f(x_i)\| < \lambda$ where λ is a small constant.

2. Regression [15 Points]

Solve the linear regression for this dataset $\mathcal{D} = \{(0,0),(4,2),(5,1)\}$.
$\bar{x} =$
$\bar{y} =$
$w_1^* =$
$w_0^* =$

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
0	0				
4	2				
5	1				

Finally, what can we say about the correlation r and the slope for this dataset (mathematically)?

Solution:

$$\bar{x} = \frac{1}{3}(0+4+5) = 3$$

$$\bar{y} = \frac{1}{3}(0+2+1) = 1$$

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
0	0	-3	-1	3	9
4	2	1	1	1	1
5	1	2	0	0	4

$$w_1^* = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{3+1+0}{9+1+4} = \frac{4}{15}$$

$$w_0^* = \bar{y} - w_1^*\bar{x} = 1 - \frac{4}{15} \cdot 3 = 1 - \frac{4}{5} = \frac{1}{5}$$

Because $w_1^* = r\frac{\sigma_y}{\sigma_x}$ where the standard deviations σ_y and σ_x are non-negative, and $w_1^* = 4/15 > 0$, thus the correlation r is positive and the slope is positive.

3. Predicting Baklava Sales [39 Points]

Mahdi runs a local pastry shop near UCSD and sells traditional desert called Baklava. He bakes Baklavas every morning to keep his promise of providing fresh Baklavas to his customers daily. Here is the amount of Baklava he sold each day during last week in pounds(lb):

$$y_1 = 100, y_2 = 110, y_3 = 75, y_4 = 90, y_5 = 105, y_6 = 90, y_7 = 25$$

Mahdi needs your help as a data scientist to suggest the best constant prediction h^* of daily sales that minimizes the empirical risk using $L(h, y)$ as the loss function. Answer the following questions and give a **brief justification** for each part. **This problem has many parts, if you get stuck, move on and come back later!**

a) [2 Points] Let $L(h, y) = |y - h|$. What is h^*? We refer to this prediction as h_1^*.

> **Solution:** As we have seen in lectures, the median minimizes the absolute loss risk function.
>
> $$h_1^* = \text{Median}(y_1, \cdots, y_7).$$

b) [2 Points] Let $L(h, y) = (y - h)^2$. What is h^*? We refer to this prediction as h_2^*.

> **Solution:** As we have seen in lectures, the median minimizes the absolute loss risk function.
>
> $$h_2^* = \text{Mean}(y_1, \cdots, y_7).$$

c) [2 Points] **True** or **False**: Removing y_1 and y_3 from the dataset does not change h_2^*.

> **Solution: False**. It changes the mean from 85 to 84. (However, the median is not changed.)

d) [2 Point] Mahdi thinks that y_7 is an outlier. Hence, he asks you to remove y_7 and update your predictions in parts a and b accordingly. Without calculating the new predictions, can you justify which prediction changes *more*? h_1^* or h_2^*?

> **Solution:** The mean squared loss is more sensitive to outliers and removing data changes the mean more. Therefore, removing y_7 affects h_2^* more than h_1^*.

e) [2 Points] **True** or **False**: Let $L(y, h) = |y - h|^3$. You can use the Gradient descent algorithm to find h^*.

> **Solution: False**. The function $|y - h|^3$ is not differantiable everywhere so we can not use the gradient descent to find the minimum.

f) [3 Points] **True** or **False**: Let $L(y, h) = \sin(y - h)$. The Gradient descent algorithm is guaranteed to converge, provided that a proper learning rate is given.

Solution: False. The function is not convex, so the gradient descent algorithm is not guaranteed to converge.

Mahdi has noticed that Baklava daily sale is associated with weather temperature. So he asks you to incorporate this feature to get a better prediction. Suppose the last week's daily temperatures are $x_1, x_2, \cdots, x_7$ in Fahrenheit (F). We know that $\bar{x} = 65$, $\sigma_x = 8.5$ and the best linear prediction that minimizes the mean squared error is $H^*(x) = -3x + w_0^*$.

g) [8 Points] What is the correlation coefficient (r) between x and y? What does that mean?

Solution: We know $w_1^* = \frac{\sigma_y}{\sigma_x} r$. We know that $\sigma_x = 8.5$ and $w_1^* = -3$. We can find σ_y as follows:

$$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{1}$$

$$= \frac{1}{7}[(100-85)^2 + (110-85)^2 + (75-85)^2 + (90-85)^2 + (105-85)^2 + (90-85)^2 + (25-85)^2] \tag{2}$$

$$= \frac{1}{7}[15^2 + 25^2 + 10^2 + 5^2 + 20^2 + 5^2 + 60^2] = 714.28 \tag{3}$$

Then, $\sigma_y = 26.7$ which results in $r = -0.95$. That means the weather tempreture inversly affect Baklava sale, i.e., they are highly correlated but in the reverse direction.

h) [2 Points] **True** or **False**: The unit of r is $\frac{lb}{F}$ (Pound per Fahrenheit).

Solution: False. The correlation coefficient has no unit. (It is a unitless number in $[-1, 1]$ range.)

i) [4 Points] Find w_0^*.

Solution: Note that $H(\bar{x}) = \bar{y}$. Therefore,

$$H(65) = -3 \times 65 + w_0^* = 85 \rightarrow w_0^* = 280. \tag{4}$$

j) [6 Points] What the best linear prediction $H^*(x)$ would be if we multiply all x_i's by 2?

Solution: The standard deviation also scales by a factor of 2, i.e., $\sigma_x' = 2 \times \sigma_x$. The same is true for the mean, i.e., $\bar{x}' = 2 \times \bar{x}$. The correlation r does not change, as well as σ_y. Thererfore, $w_1'^* = \frac{w_1^*}{2}$.

Also,

$$\bar{y} = H(\bar{x}') = \frac{w_1^*}{2}(2\bar{x}) + w_0'^* = w_1^*\bar{x} + w_0'^*$$

Therefore, $w_0'^* = w_0^* = 280$.

k) [6 Points] What the best linear prediction $H^*(x)$ would be if we add 20 to all x_i's?

Solution: All parameters remain unchanged except $\bar{x}' = \bar{x} + 20$. Since r, σ_x and σ_y are not changed, so does w_1^*. Then, one can write,

$$\bar{y}' = H(\bar{x}') \Rightarrow 85 = -3(65 + 20) + w_0^*$$

which results in $w_0^* = 340$.

4. Midterm Loss Function [25 Points]

Consider a dataset that consists of $y_1, \cdots, y_n$. In class, we used calculus to minimize mean squared error, $R_{sq}(h) = \frac{1}{n}\sum_{i=1}^{n}(h - y_i)^2$. In this problem, we want you to apply the same approach to a slightly different loss function defined beelow:

$$L_{\text{midterm}}(y, h) = (\alpha y - h)^2 + \lambda h$$

a) [5 Points] Write down the empiricial risk $R_{\text{midterm}}(h)$ by using the above loss function.

Solution:

$$R_{\text{midterm}}(h) = \frac{1}{n}\sum_{i=1}^{i=n}[(\alpha h - y_i)^2 + \lambda h] = [\frac{1}{n}\sum_{i=1}^{i=n}(\alpha h - y_i)^2] + \lambda h$$

b) [20 Points] The mean of dataset is $\bar{y}$, i.e. $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Find h^* that minimizes $R_{\text{midterm}}(h)$ using calculus. Your result should be in terms of $\bar{y}$, α and λ.

Solution:

$$\frac{d}{dh}R_{\text{midterm}}(h) = [\frac{2\alpha}{n}\sum_{i=1}^{i=n}(\alpha h - y_i)] + \lambda \tag{5}$$

$$= 2\alpha^2 h - 2\alpha\bar{y} + \lambda. \tag{6}$$

By setting $\frac{d}{dh}R_{\text{midterm}}(h) = 0$ we get

$$2\alpha^2 h^* - 2\alpha\bar{y} + \lambda = 0 \Rightarrow h^* = \frac{\bar{y}}{\alpha} - \frac{\lambda}{2\alpha^2}.$$