DSC 40A - Homework 5

due Friday, November 14th at 11:59 PM

Homeworks are due to Gradescope by 11:59PM on the due date.

You can use a slip day to extend the deadline by 24 hours; you have four slip days to use in total throughout the quarter.

Homework will be evaluated not only on the correctness of your answers, but on your ability to present your ideas clearly and logically. You should **always explain and justify** your conclusions, using sound reasoning. Your goal should be to convince the reader of your assertions. If a question does not require explanation, it will be explicitly stated.

Homeworks should be written up and turned in by each student individually. You may talk to other students in the class about the problems and discuss solution strategies, but you should not share any written communication and you should not check answers with classmates. You can tell someone how to do a homework problem, but you cannot show them how to do it. **Only handwritten solutions will be accepted (use of tablets is permitted). Do not typeset your homework (using LaTeXor any other software)**.

For each problem you submit, you should **cite your sources** by including a list of names of other students with whom you discussed the problem. Instructors do not need to be cited.

This homework will be graded out of **69 points**. The point value of each problem or sub-problem is indicated by the number of avocados shown.

Problem 1. Mid-quarter Feedback Form

🥑🥑 Make sure to fill out this Mid-quarter Reflection and Feedback Form, linked here! Note that this form is anonymous so that you feel comfortable sharing constructive feedback. Everyone will be given two points and the course staff is trusting you to fill out the form.

Problem 2. Combinations of Convex Functions

In class, we saw that convex risk functions are nice because they are relatively easy to minimize using gradient descent. But how do we determine if our risk function is convex? One way is to show that it is built from simpler convex functions.

For each statement below, either prove the statement true using the *formal definition* of convexity from the lecture, or prove the statement false by finding a concrete counterexample. Suppose that $f_1(x)$ and $f_2(x)$ are convex functions defined on $\mathbb{R}$.

a)  $f_1(x) + f_2(x)$ is a convex function.

b)  $f_1(x) - f_2(x)$ is a convex function.

c)  If $f_1(a) = f_2(a)$, where $a \in \mathbb{R}$ is fixed, then the function $g(x)$ is also convex, where:

$$g(x) = \begin{cases} f_1(x) & x \leq a \\ f_2(x) & x > a \end{cases}.$$

Problem 3. Meet the Jensens

As we've seen, the variance of a dataset $x_1, x_2, ..., x_n$ is defined:

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

where $\bar{x} = \text{Mean}(x_1, x_2, ..., x_n)$. By expanding the summation, we find that:

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - \bar{x}^2)$$

Another way of expressing this equation is:

$$\sigma_x^2 = \text{Mean}(x_1^2, x_2^2, ..., x_n^2) - (\text{Mean}(x_1, x_2, ..., x_n))^2$$

Since $\sigma_x^2 \geq 0$, this implies that:

$$\begin{aligned} &\text{Mean}(x_1^2, x_2^2, ..., x_n^2) - (\text{Mean}(x_1, x_2, ..., x_n))^2 \geq 0 \\ &\qquad \implies \text{Mean}(x_1^2, x_2^2, ..., x_n^2) \geq (\text{Mean}(x_1, x_2, ..., x_n))^2 \end{aligned}$$

The inequality on the last line can be expressed more generally as:

$$\text{Mean}(g(x_1), g(x_2), ..., g(x_n)) \geq g\left(\text{Mean}(x_1, x_2, ..., x_n)\right)$$

The inequality above is known as Jensen's inequality, and is true for **all convex functions** g. Let's see how we can use Jensen's inequality to prove something useful!

a) 🥑🥑 Prove that the function $g(x) = -\log x$ is convex. *Hint: Use the second derivative test — it would be difficult to prove this with the formal definition of convexity.*

b) 🥑🥑🥑🥑 Using Jensen's inequality and $g(x) = -\log x$, prove that, for any dataset of positive numbers $x_1, x_2, ..., x_n$:

$$\frac{x_1 + x_2 + ... + x_n}{n} \geq (x_1 \cdot x_2 \cdot ... \cdot x_n)^{\frac{1}{n}}$$

The quantity on the left is the familiar arithmetic mean (AM), while the quantity on the right is known as the geometric mean (GM) of $x_1, x_2, ..., x_n$. The entire inequality above is known as the "AM-GM inequality." This is not your first time seeing it — you saw it in Groupwork 2, too!

c) 🥑🥑🥑🥑 Using Jensen's inequality and a convex function g, prove that the arithmetic mean is greater than or equal to the harmonic mean for any dataset of positive numbers $x_1, x_2, ..., x_n$, i.e.:

$$\frac{x_1 + x_2 + ... + x_n}{n} \geq \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}}$$

Note that you must prove your choice of function g is convex!

Hint: You can use a function that is only convex on an interval, as long as the only inputs you pass into the function come from that interval.

Problem 4. Huber Loss

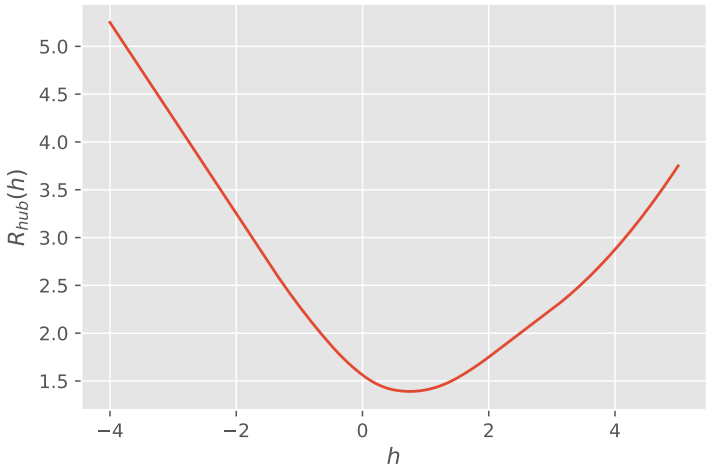
The *Huber loss* (introduced in lecture 16) is a mixture between the square loss and the absolute loss. It is defined piecewise as follows:

$$L_{\text{hub}}(h, y) = \begin{cases} |h - y|, & |h - y| > 1 \\ \frac{1}{2}(h - y)^2 + \frac{1}{2}, & |h - y| \leq 1 \end{cases}$$

a) 🥑🥑 Fix an arbitrary value of y. Draw the graph of $L_{\text{hub}}(h, y)$ as a function of h. You should notice that $L_{\text{hub}}(h, y)$ is minimized at y.

b) 🥑 What is the derivative of L_{hub} with respect to h? Your answer should also be a piecewise function.

c) 🥑🥑🥑🥑🥑 Suppose our data set is $\{-\frac{1}{2}, \frac{1}{2}, 1, 4\}$. The plot of the empirical risk, $R_{\text{hub}}(h) = \frac{1}{n}\sum_{i=1}^{n} L_{\text{hub}}(h, y)$ is shown below:



It is not possible to directly solve for the value of h which minimizes this function. Instead, run gradient descent **by hand** using an initial prediction of $h_0 = 5$ and a step size of $\alpha = 2$. Run the algorithm until it converges (it shouldn't take too many iterations). Please show your calculations, and to help the graders track your progress, include a boxed summary with the value of h at each iteration, such as below:

$$\boxed{\begin{array}{l} h_0 = 5 \\ h_1 = \ldots \\ h_2 = \ldots \\ \vdots \end{array}}$$

Problem 5. Gradient Descent in multiple dimensions

Now that we've gotten a feel for how to use gradient descent to minimize a function of a single variable, we'll use it to minimize a function of multiple variables. Let's suppose we want to fit a simple linear regression model, $H(x) = w_0 + w_1 x$, using **squared loss**. We're searching for the optimal parameters w_0^* and w_1^*, which we can be written in vector form as $\vec{w}^* = \begin{bmatrix} w_0^* \\ w_1^* \end{bmatrix}$.

We're given the following dataset of (x, y) pairs: $\{(1, 5), (2, 7)\}$.

a)  Write the empirical risk $R_{\text{sq}}(\vec{w}) = R_{\text{sq}}(w_0, w_1)$ and the gradient vector:

$$\nabla R_{\text{sq}}(\vec{w}) = \begin{bmatrix} \frac{\partial R_{\text{sq}}}{\partial w_0} \\ \frac{\partial R_{\text{sq}}}{\partial w_1} \end{bmatrix}$$

Both $R_{\text{sq}}(\vec{w})$ and $\nabla R_{\text{sq}}(\vec{w})$ should only involve the variables w_0 and w_1; everything else should be constants.

b)  Given an initial guess $\vec{w}^{(0)} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$ and a learning rate of $\alpha = \frac{1}{3}$, perform one iteration of gradient descent according to the iterative procedure:

$$\vec{w}^{(t)} = \vec{w}^{(t-1)} - \alpha \nabla R(\vec{w}^{(t-1)})$$

What are the components of $\vec{w}^{(1)}$?

Problem 6. Probability Rules for Three Events

a) 🥑🥑🥑 The multiplication rule for two events says:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B|A)$$

Use the multiplication rule for two events to prove the multiplication rule for three events:

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A) \cdot \mathbb{P}(B|A) \cdot \mathbb{P}(C|(A \cap B))$$

Hint: You can think of $A \cap B \cap C$ *as* $(A \cap B) \cap C$.

b) 🥑🥑 Suppose E, F, and G are events. Explain in words why:

$$(E \cup F) \cap G = (E \cap G) \cup (F \cap G)$$

Intuitively, the relationship between $\cap$ and $\cup$ is similar to the relationship between multiplication and addition; if e, f, g are numbers, then $(e + f) \cdot g = e \cdot g + f \cdot g$ as well.

c) 🥑🥑🥑 The general addition rule for any two events says:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$$

Use the general addition rule for two events, along with the result of part (b), to prove the general addition rule for three events:

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C)$$

This is often called the "Principle of Inclusion-Exclusion."

d) 🥑🥑🥑 To identify what students find most important in DSC 10, we want to administer a survey to the students in DSC 20, DSC 30, and DSC 40A. Consider the following information:

- There are 300 students taking at least one of DSC 20, DSC 30, or DSC 40A right now.
- 200 students are taking DSC 20 right now, and 50 students taking DSC 30 right now. There are no students taking both DSC 20 and DSC 30 right now.
- 50 students are taking both DSC 20 and DSC 40A right now, and 30 students are taking both DSC 30 and DSC 40A right now.

Suppose I choose a single student uniformly at random from the population of students taking at least one of DSC 20, DSC 30, and DSC 40A. What is the probability that they are enrolled in DSC 40A? Simplify your answer.

Problem 7. Eventful

a) 🥑🥑 Let A and B are two events such that $P(B) > 0$. What is the value of $P(A \cap \bar{B}|B)$?

b) 🥑🥑 Consider two fair 9-sided dice, each with faces numbered 1, 2, 3, ..., 9. Suppose you roll the two dice and look at one of them. You see that this one die is **less than 3**. What is the probability that the sum of the two dice rolls is **greater than 10**?

c) 🥑🥑🥑🥑 A box contains three cards. One card is red on both sides, one card is green on both sides, and one card is red on one side and green on the other. One card is selected from the box at random, and the color on one side is observed. If this side is green, what is the probability that the other side of the card is also green?

d) 🥑🥑🥑 In Texas, license plates generally consist of 3 letters followed by 4 numbers. All letters are uppercase, and repeated characters are allowed. ABC-1234 is an example of a Texas license plate.

What is the probability that a randomly generated license plate begins with a vowel or ends in a number divisible by 3? Simplify your answer.

Problem 8. Time to get competitive! 

This problem consists of a regression modeling competition on a real-world dataset known as the *Diabetes disease progression dataset*.

The Rules:

1. Your **goal** is to design and train a model with the objective of finding the best fit for the Diabetes dataset's hidden testing labels. You may design any model using techniques covered in DSC40A so far (multiple regression, data tranformations, etc.). If you wish to get more advanced, you can also use techniques covered in the course notes: in particular, you are encouraged to read over Section 3.1 and Section 3.3.
2. Following the instructions in the supplementary notebook below, you should submit your predictions to the "Modeling competition" autograder on Gradescope. You should **also submit** a copy of your completed notebook with the rest of your homework assignment as usual.
3. You are encouraged to work in pairs or teams of **at most three people**. You may submit results to Gradescope in teams but **you must submit your homework, including a copy of your code, individually as well**.
4. **To receive full credit**, you must beat the staff testing MSE of **2927.394**. **The top five teams will receive a 20% extra credit boost on this assignment, for a maximum of 120% of the total credit of this homework assignment.** The top five teams will be required to submit copies of their models and code that will be expected to reproduce the results submitted to Gradescope up to machine precision. Failure to do so, or the usage of models outside the scope of DSC40A and/or the course notes, will result in disqualification from the competition and zero points for this problem.
5. Any successful submission of code and results to Gradescope, even if it doesn't beat the staff MSE, is eligible for partial credit.

The Links:

1. Supplementary notebook
2. Training data (features and targets)
3. Testing data (features only)