DSC 40A - Group Work Session 2

due October 10, 2022 at 11:59pm

Write your solutions to the following problems by either typing them up or handwriting them on another piece of paper. You must work in a group of 2 to 4 students for at least 50 minutes to get credit for this assignment. It's best to join a discussion section if possible.

One person from each group should submit your solutions to Gradescope by 11:59pm on Thursday. Make sure to **tag all group members** so everyone gets credit. This worksheet won't be graded on correctness, but rather on good-faith effort. Even if you don't solve any of the problems, you should include some explanation of what you thought about and discussed, so that you can get credit for spending time on the assignment.

1 Empirical Risk Minimization

In class, we've seen how to minimize the empirical risk associated with certain natural loss functions, such as the absolute loss and the squared loss. There are a variety of other possible loss functions we could use instead. This problem explores empirical risk minimization with an alternate choice of loss function.

Problem 1.

In this problem, consider the loss function

$$L(h, y) = \begin{cases} 1, & |y - h| > 1 \\ |y - h|, & |y - h| \leq 1 \end{cases}.$$

a) Consider y to be a fixed number (if you want, pretend $y = 1$). Plot $L(h, y)$ as a function of h.

b) Suppose that we have the following data:

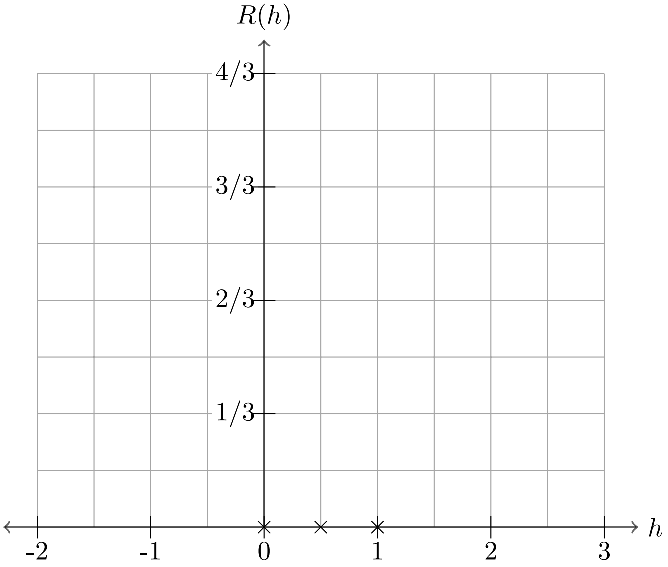
$$\begin{aligned} y_1 &= 0 \\ y_2 &= 1 \\ y_3 &= 1.5 \end{aligned}$$

Plot the empirical risk

$$R(h) = \frac{1}{n} \sum_{i=1}^{n} L(h, y_i)$$

on the domain $[-2, 3]$. It might help to use the grid on the next page; note that the vertical axis tick marks occur in increments of $1/3$ while the horizontal axis tick marks are in increments of 1.

Hint: $R(h)$ is made up of several line segments. What is the slope of each line segment?



c) Suppose that we are interested in finding the typical price of an avocado using this loss function. To do so, we have gathered a data set of n avocado prices, $y_1, \ldots, y_n$, and we found the price h^* which minimized the empirical risk (a.k.a, average loss), $R(h) = \frac{1}{n} \sum L(h, y_i)$.

Unfortunately, a flat tax of c dollars has been imposed on avocados since we performed our analysis, increasing every price in our data set by c.

Is it true that $h^* + c$ is a minimizer of R when we use the new prices, $(y_1 + c), (y_2 + c), \ldots, (y_n + c)$? Explain why or why not by explaining how the graph of R changes.

d) Suppose that instead of a flat tax, a percentage-based tax has been imposed. That is, the new avocado prices are $(1 + \alpha)y_1, (1 + \alpha)y_2, \ldots, (1 + \alpha)y_n$. Is $(1 + \alpha)h^*$ still a minimizer of R when we use the new prices? Explain why or why not.

e) Given avocado prices $\{1/4, 1/2, 3/4, 7/8, 9/8\}$, find a minimizer of R. Provide some justification for your answer.

Hint: you don't need to plot R or do any calculation to find the answer.

2 Gradient Descent

Gradient descent is used to minimize differentiable functions. In this class, we will primarily use it to minimize empirical risk.

So far, we've been making predictions using a **constant hypothesis** h, where we predict the same future value regardless of all other attributes. For example, continuing with our example of data scientist salaries, we've been predicting the same future salary for everyone, regardless of education, prior work experience, location, etc.

We are going to be changing that and instead use other attributes to help make our predictions. In this question, we'll do the same step by step.

Assume we have a dataset containing the years of experience and salaries of college graduates who are now working as data scientists. Our dataset is of the form $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ where x_i represents the number of years of experience of the i-th individual and y_i represents the salary of the i-th individual.

Suppose we want to predict future salary by taking years of experience and multiplying by some constant, w. This corresponds to using the **prediction rule** $H(x) = wx$. Our goal is to use our dataset to **learn** what the best choice of w is. To do this, suppose we choose squared loss. The corresponding empirical risk is then

$$R_{sq}(w) = \frac{1}{n}\sum_{i=1}^{n}(y_i - wx_i)^2$$

Our goal is to minimize this function with respect to w to determine the best constant multiple to make our predictions with.

Next week in class, we will walk through how to minimize such a function with algebra. But for now, we'll use it as an opportunity to practice gradient descent.

Problem 2.

a) Determine $\frac{d}{dw}R_{sq}$, that is, the derivative of $R_{sq}(w)$ with respect to w.

b) Suppose we have a dataset of just two points, $(5, 150)$ and $(7, 200)$ (assume that salaries are stored in the thousands). Determine a formula for $\frac{d}{dw}R_{sq}$ for this particular dataset. Unlike your answer to the previous part, your answer to this part should only contain a single variable, w.

c) To run gradient descent, we need to choose an "initial guess" for the best value of w, which we'll call w_0, and a "step size" α. Then, through gradient descent, we update our predictions using the rule

$$w_{i+1} = w_i - \alpha\big(\frac{d}{dw}R_{sq}(w_i)\big)$$

Suppose we choose $\alpha = \frac{1}{20}$ and that we choose an initial guess of $w_0 = 10$. Determine the values of w_1 and w_2, that is, run gradient descent for two iterations and report back the resulting values of w each time. You can use a calculator.

3 Optional: Chaining Inequalities

Note: This section is optional in that we're not expecting you to attempt it in order to get credit for the groupwork session. However, it serves as great practice for proofs!

Suppose we have collected a bunch of numbers, $y_1, \ldots, y_n$. Let's assume, too, that these numbers are in sorted order, so that $y_1 \leq y_2 \leq \ldots \leq y_n$.

The *midpoint* of $y_1, \ldots, y_n$ is the average of the smallest and largest number:

$$\text{midpoint} = \frac{y_1 + y_n}{2}.$$

Intuitively, the midpoint is at most y_n and is at least y_1; it lies somewhere in the middle of these two numbers. We can easily prove this with a *chain* of inequalities.

First, we show that the midpoint is at most y_n. We start with the definition:

$$\text{midpoint} = \frac{y_1 + y_n}{2}$$

We can do anything to the right hand side that makes it bigger, keeping in mind that we're trying to get it to look like y_n. Right now there is y_1 hanging out; can we simply change it to a y_n? Yes! Remember that $y_n \geq y_1$, so this would make the right hand side bigger. Therefore, we have to write $\leq$:

$$\leq \frac{y_n + y_n}{2}$$

We can simplify this:

$$= \frac{2y_n}{2}$$

Notice that we wrote $=$ on the last line, not $\leq$. This is because the line is indeed equal to the one before it.

$$= y_n$$

We have made a chain of inequalities and equalities; this one looks like $=, \leq, =, =$. Since $\leq$ is the "weakest link" in the chain, the strongest statement we can make is that the midpoint is $\leq y_n$, but this is what we wanted to say.

Problem 3.

Prove that the midpoint is $\geq y_1$.

Problem 4.

Suppose $y_1, \ldots, y_n$ are all positive numbers. The *geometric mean* of $y_1, \ldots, y_n$ is defined to be:

$$(y_1 \cdot y_2 \cdots y_n)^{1/n}.$$

Prove that the geometric mean is less than or equal to y_n and greater than or equal to y_1 using a chain of inequalities. (Note: Assume that the numbers are ordered)