
DSC 40A - Homework 2
Due: Sunday, April 17, 2022 at 11:59pm PDT

Write your solutions to the following problems by either typing them up or handwriting them on another piece of paper. Homeworks are due to Gradescope by 11:59pm PDT on Sunday.

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.

This policy also means that you **should not post or answer homework-related questions on Piazza**, which is a written medium. This includes private posts to instructors. Instead, when you need help with a homework question, talk to a classmate or an instructor in their office hours.

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. The point value of each problem or sub-problem is indicated by the number of avocados shown.

Problem 1. Properties of Empirical Risk

Given a data set $y_1 \leq y_2 \leq \dots \leq y_n$, define the following empirical risk functions:

$$R_{\text{ab}}(h) = \frac{1}{n} \sum_{i=1}^n |y_i - h|, \quad R_{\text{sq}}(h) = \frac{1}{n} \sum_{i=1}^n (y_i - h)^2.$$

Parts (a), (b), and (c) below concern R_{ab} . Parts (d) and (e) concern R_{sq} .

- a) 🥑🥑 Suppose $n = 10$. For an unknown number c with $y_2 < c < c + 2 < y_3$, how does $R_{\text{ab}}(c)$ compare to $R_{\text{ab}}(c + 2)$? Can you determine which is bigger, and by how much?
- b) 🥑🥑 Still suppose $n = 10$. This time, for an unknown number c with $y_5 < c < c + 2 < y_6$, how does $R_{\text{ab}}(c)$ compare to $R_{\text{ab}}(c + 2)$? Can you determine which is bigger, and by how much?
- c) 🥑🥑🥑 Now for an unknown number of data points n , suppose that $R_{\text{ab}}(c) = R_{\text{ab}}(c + 2) = V$, where V is the minimum value $R_{\text{ab}}(h)$. If we add to the data set a new data point at $c + 0.5$, what is the new minimum of $R_{\text{ab}}(h)$ and at which value of h is it achieved?
- d) 🥑🥑 For an arbitrary data set of size n and an unknown number c with $c < y_1$, how does $R_{\text{sq}}(c)$ compare to $R_{\text{sq}}(c - 1)$? Can you determine which is bigger, and by how much?
- e) 🥑🥑 Suppose an unknown element y_i from the data set $y_1 \leq y_2 \leq \dots \leq y_{10}$ is changed to $y_i + 2$. Can you determine how the value of h^* that achieves the minimum of R_{sq} changes, and by how much?

Problem 2. Mean, Variance, and Standard Deviation

- a) 🥑🥑🥑🥑 Suppose we are given a data set $\{d_1, d_2, \dots, d_n\}$ and know its mean, variance, and standard deviation to be $mean_d$, var_d , and std_d . Consider another data set $\{t_1, t_2, \dots, t_n\}$, where t_i is a linear

transformation of d_i :

$$t_i = f(d_i) = a \cdot d_i + b$$

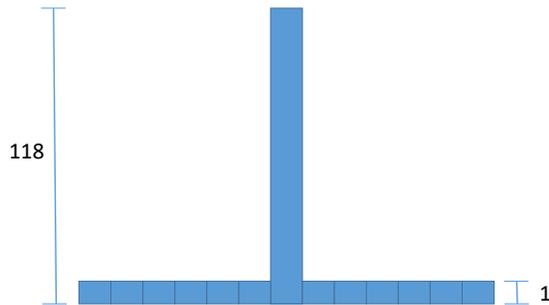
for each $i = 1, 2, \dots, n$. If the mean, variance, and standard deviation of this transformed data set are $mean_t$, var_t , std_t , then express each of these quantities in terms of $mean_d$, var_d , std_d , a , and b .

- b) 🥑🥑 Notice that for a linear transformation f , we can switch the order of computing the mean and applying f and achieve the same result. Can you say the same for the standard deviation? If the same holds, prove it. If the same does not hold, what are the restrictions on the linear transformation f so that you can switch the order of computing the standard deviation and applying f and get the same result? How about for the variance?

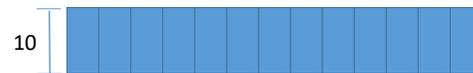
Problem 3. Spread of Distributions

🥑🥑🥑🥑 Look at the four different data distributions shown below. Each has the same x-axis, where the markings are evenly spaced, splitting the data into thirteen bins of equal size. The frequency count for each bin is shown by the height of each bar.

Distribution A:



Distribution B:



Distribution C:



Distribution D:



Can you rank these four distributions in ascending order of their mean absolute deviation from the median? Be sure to justify (in words or in numbers) every inequality or equality that you write.

Problem 4. Gradient Descent on a New Loss Function

Consider a new loss function,

$$L_p(h; y) = \begin{cases} h - y - \frac{1}{4}, & h - y > \frac{1}{2} \\ (h - y)^2, & h - y \leq \frac{1}{2} \end{cases} . \quad (1)$$

a) 🥑🥑 Fix an arbitrary value of y . Draw the graph of $L_p(h; y)$ as a function of h (label your choice of y on the graph). Note that the piecewise function connects at: $h - y = \frac{1}{2}$ (instead of at $|h - y| = \frac{1}{2}$). Also notice that the minimum of $L_p(h; y)$ is zero, regardless of the value of y .

b) 🥑🥑🥑🥑 Recall from single-variable calculus the following definitions:

- A function $f(x)$ is continuous at $x = a$ if $\lim_{x \rightarrow a} f(x) = f(a)$.
- A function $f(x)$ is differentiable at $x = a$ if $\lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$ exists. This limit is called $f'(a)$, the derivative of f at $x = a$.

Fix an arbitrary value of y . Show that as a function of h , L_p is continuous and differentiable.

Give the derivative of L_p as a piecewise function of h .

c) 🥑🥑 Plot the empirical risk associated with this loss function, on the data set $\{y_1, y_2, y_3\} = \{-2, 0, 2\}$ of size $n = 3$.

$$R_p(h) = \frac{1}{n} \sum_{i=1}^n L_p(h; y_i) \quad (2)$$

Problem 5. Bounding the derivative of R_p

For the empirical risk R_p defined in equation (2), prove the following property that we used during the class:

$$\left(\frac{d}{dh} R_p(h) \right)^2 \leq 4(R_p(h) - R_p(h^*)),$$

where h^* is the minimum of R_p . In particular, follow the steps laid out below.

a) 🥑🥑🥑🥑🥑 First use the property of the loss function $L_p(h; y_i)$ to prove that for any h and $h' \in \mathbb{R}$,

$$\left(\frac{d}{dh} L_p(h; y_i) - \frac{d}{dh} L_p(h'; y_i) \right)^2 \leq 4(L_p(h; y_i) - L_p(h'; y_i)) + 4 \frac{d}{dh} L_p(h'; y_i)(h' - h). \quad (*)$$

Check the 4 cases where $(h - y_i), (h' - y_i) \leq 1/2$, $(h - y_i) \leq 1/2$ while $(h' - y_i) > 1/2$, $(h - y_i) > 1/2$ while $(h' - y_i) \leq 1/2$, and $(h - y_i), (h' - y_i) > 1/2$, respectively.

You will in fact find that in 3 out of the 4 cases, *equality* is achieved in the above expression (*). Could you identify the case where the equality does not hold in general? Prove that in this particular case, even if the equality does not hold, the above *inequality* (*) still holds. (Hint: if $T_1 - T_2 \leq 0$, then $T_1 \leq T_2$.)

b) 🥑🥑🥑🥑 Use what we proved in the class that $(\frac{1}{n} \sum_{i=1}^n a_i)^2 \leq \frac{1}{n} \sum_{i=1}^n a_i^2$ (no need to prove this inequality again) to obtain:

$$\left(\frac{d}{dh} R_p(h) - \frac{d}{dh} R_p(h') \right)^2 \leq \frac{1}{n} \sum_{i=1}^n \left(\frac{d}{dh} L_p(h; y_i) - \frac{d}{dh} L_p(h'; y_i) \right)^2 .$$

Then use linearity of the differential operator $\frac{d}{dh}$ as well as the fact that $\frac{d}{dh} R_p(h^*) = 0$ when h^* is the minimum of R_p to finish the proof that: $(\frac{d}{dh} R_p(h))^2 \leq 4(R_p(h) - R_p(h^*))$.