DSC 40A - Homework 1

Due: Sunday, April 10, 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. Salary Statistics

During the class, I presented the salaries of some of my friends working as data scientists and software developers. Among the 7 software developers, we have computed that they have an average salary of $210K$.

a) 🥑🥑🥑 Last night, I noticed that I copied one number wrong. One particular friend of mine's entry level salary as a software developer is in fact $270K$, instead of the $200K$ that I presented in class. From this information alone (without access to the full data set), can you correct the mistake and find the correct average salary for the group of the software developers?

b) 🥑🥑🥑 In general, are you able to correct an average if one of the data values changes? What about the median? The mode?

Problem 2. Averages

Which of the following statements *must* be true? Remember to justify all answers.

a) 🥑🥑 Some of the numbers in the data set must be smaller than the average.

b) 🥑🥑 At least half of the numbers in a data set must be smaller than the average.

c) 🥑🥑 Exactly half of the numbers in a data set must be smaller than the average.

d) 🥑🥑 Not all of the numbers in the data set can be smaller than the average.

Problem 3. Fahrenheit or Celsius?

The other day, you went online to look at how other cities around the world compare to San Diego in terms of the climate. You looked up Osaka, Japan and found that their temperature is recorded in Celsius. The table below shows the monthly temperatures in these two places.

San Diego, US (°F)	66	66	67	69	69	72	76	77	77	74	70	66
Osaka, Japan (°C)	9	10	14	20	25	28	32	33	29	23	18	12

You want to compare the median monthly temperatures in San Diego and Osaka. One way is to convert all the degrees in Fahrenheit to Celsius before computing the median temperature for San Diego. This way, you compare the median temperatures in both places using the same units.

You friend Skip is lazier and thinks you can skip some of that work: "Why don't we find the median temperature in San Diego in Fahrenheit first, and then only convert the median value to Celsius? That way we only need to do the conversion once or twice."

a) 🥑🥑🥑 Is Skip correct that you'll get the same median temperature either way? Show your work.

b) 🥑🥑🥑🥑 More generally, let $g(t)$ be the function which takes in a temperature in degrees Fahrenheit and outputs the temperature in Celsius. That is, $g(t) = \frac{5}{9} \times (t - 32)$. Then this problem can be formulated mathematically as the validity of the following equation:

$$\text{Median}(g(t_1), \ldots, g(t_n)) = g(\text{Median}(t_1, \ldots, t_n)).$$

Prove this equation or disprove it by counter-example.

c) 🥑🥑🥑 Finally, you find that the median temperatures for San Diego and Osaka are quite close to each other. Can you say that the climate at the two cities are similar to each other? Where do you think the median temperature is more representative of the overall climate, San Diego or Osaka? What is your reasoning?

Problem 4. Minimize risk or maximize likelihood?

In our lecture, we argued that one way to make a good prediction h is to minimize the mean absolute error associated with data set $D = \{x_1, \ldots, x_n\}$:

$$R_{ab}(h; D) = \frac{1}{n} \sum_{i=1}^{n} |h - x_i|.$$

We saw that the median of $x_1, \ldots, x_n$ is the prediction with the smallest mean error. Your friend Max thinks that instead of minimizing the mean error, it is better to maximize the following quantity:

$$M(h) = \prod_{i=1}^{n} e^{-|h - x_i|}.$$

The above formula is written using product notation, which is similar to summation notation, except terms are multiplied and not added. For example,

$$\prod_{i=1}^{n} a_i = a_1 \cdot a_2 \cdot \ldots \cdot a_n.$$

Max's reasoning is that for some models, $e^{-|h - x_i|}$ is used to compute how likely the predicted value h will appear given the observation x_i – hence it is called "likelihood." Then, we should attempt to maximize the chance of getting the prediction h, given the set of observations. In this problem, we'll see if Max has a good idea.

a) 🥑🥑🥑 For an arbitrary fixed value of x_i, sketch a graph of the basic shape of the likelihood function $M(h) = e^{-|h-x_i|}$. Explain, based on the graph, why larger values of $M(h)$ correspond to better predictions h.

b) 🥑🥑🥑🥑 Informally, a minimizer of a function f is an input x_{min} where f achieves its minimum value. More formally, x_{min} is a minimizer of f if $f(x_{\text{min}}) \leq f(x)$ for all values of x. In the same way, x_{max} is a maximizer of f if $f(x_{\text{max}}) \geq f(x)$ for all values of x.

Suppose that f is some unknown function which takes in a real number and outputs a real number. Suppose that c is an unknown positive constant, and define the function $g(x) = e^{-c \cdot f(x)}$. Prove that if x_{min} is a minimizer of f, then it is also a maximizer of g.

c) 🥑🥑🥑🥑 At what value h^* is $M(h)$ maximized? Is this a reasonable prediction? Discuss the pros and cons of using Max's prediction strategy, and describe scenarios where this gives a good prediction and where this gives a bad prediction, in your opinion.

Problem 5. Empirical risk with quadratic loss

During the group work session (in the last question), you have chosen a likely depiction of the empirical risk function with quadratic loss, $R_{sq}(h, D)$, for a dataset $D = \{100K, 200K, 300K, 400K\}$. Now let's prove your intuition.

a) 🥑🥑🥑🥑🥑 Prove that for any data set $D = \{x_1, \ldots, x_n\}$, the empirical risk:

$$R_{sq}(h, D) = \frac{1}{n}\sum_{i=1}^{n}(h - x_i)^2$$

can be simplified as:

$$R_{sq}(h, D) = (h - y)^2 + z.$$

Express quantities y and z in terms of $x_1, \ldots, x_n$ and n. (Hint: expand the squares in $R_{sq}(h, D)$. Use induction if you find it useful.)

b) 🥑🥑 Using the above result, minimize $R_{sq}(h, D)$ (write down what's $\arg\min_{h \in \mathbb{R}} R_{sq}(h, D)$ and what's $\min_{h \in \mathbb{R}} R_{sq}(h, D)$) without using calculus. You can use the properties of the quadratic functions.

c) 🥑🥑🥑 For an arbitrary data set $D = \{x_1, \ldots, x_n\}$ ($x_i \in \mathbb{R}$, for any $i = 1, \ldots, n$), is quantity y generally positive, negative, or zero, and why? How about quantity z and why?