DSC 40A - Homework 01

Due: Friday, January 17, 2020

Write your solutions to the following problems by either typing them up or handwriting them on another piece of paper. Unless otherwise noted by the problem's instructions, show your work or provide some justification for your answer. Homeworks are due via Gradescope on Friday afternoon at 5:00 p.m.

Problem 1.

Which of the following equations involving summation notation are actually wrong? Write the letters of all which are **incorrect**; you do not need to show your work. You can assume that c is a constant, n is a positive integer, k is a positive integer that is less than n, and that $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ are real numbers.

(a) $\displaystyle\sum_{i=1}^{n} c \cdot x_i = c \sum_{i=1}^{n} x_i$

(b) $\displaystyle 4\sum_{i=1}^{n} (x_i + y_i) = 4\sum_{i=1}^{n} x_i + 4\sum_{i=1}^{n} y_i$

(c) $\displaystyle\sum_{i=1}^{n} x_i \cdot y_i = \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$

(d) $\displaystyle\sum_{i=1}^{10} x_i = \sum_{i=1}^{7} x_i + \sum_{i=8}^{10} x_i$

(e) $\displaystyle\sum_{i=1}^{n} c = c \cdot n$

(f) $\displaystyle\sum_{i=k}^{n} 5 = 5(n-k)$

(g) $\displaystyle\sum_{i=1}^{n} x_i = \sum_{j=1}^{n} x_j$

(h) $\displaystyle\sum_{i=1}^{n} i = n$

Problem 2.

In lecture, we argued that a good prediction h is one which has a small mean error:

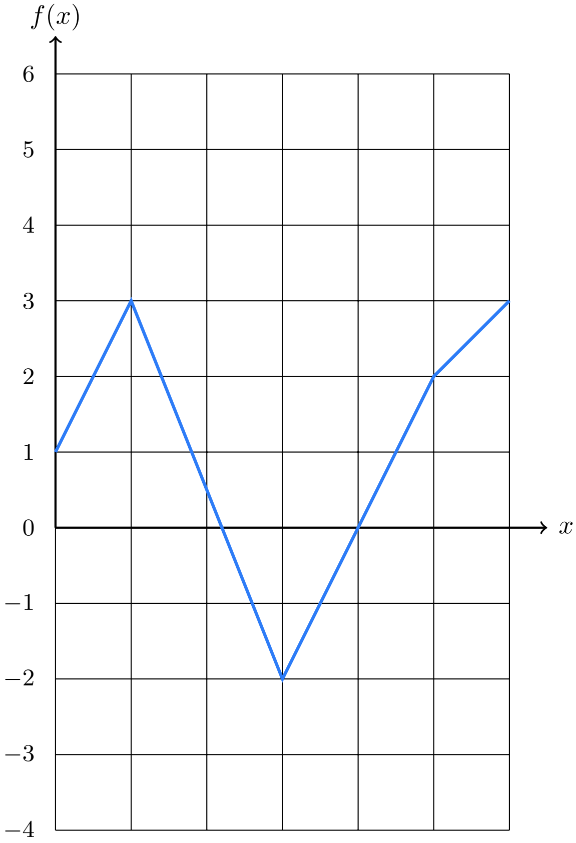
$$R(h) = \frac{1}{n}\sum_{i=1}^{n} |y_i - h|$$

We saw that the median of $y_1, \ldots, y_n$ is the prediction with the smallest mean error. But your friend Zelda thinks that instead of minimizing the mean error, it is better to minimize the *total error*:

$$T(h) = \sum_{i=1}^{n} |y_i - h|$$

In this problem, we'll see if Zelda has a good idea.

a) Consider the function f plotted below:



Draw the function $g(x) = 2 \cdot f(x)$.

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 $f(x_{\text{min}}) \leq f(x)$, no matter what x is.

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 $g(x) = c \cdot f(x)$. Argue that if x_{min} is a minimizer of f, then it is also a minimizer of g.

Hint: what does it mean (formally) for x_{min} to be a minimizer of g? Try to show that $g(x_{\text{min}}) \leq g(x)$, whatever x may be, using a chain of inequalities.

c) Zelda suggested that we minimize the total error instead of the mean error. What is the minimizer of the total error?

Problem 3.

Suppose that $y_1, \ldots, y_n$ are all real numbers, with y_1 being the smallest and y_n being the largest. Argue that $\text{Mean}(y_1, \ldots, y_n)$ falls somewhere within the interval $[y_1, y_n]$. That is, prove:

$$y_1 \leq \text{Mean}(y_1, \ldots, y_n) \leq y_n.$$

Hint: try to construct two chains of inequalities which show that $\text{Mean}(y_1, \ldots, y_n) \leq y_n$ and $\text{Mean}(y_1, \ldots, y_n) \geq y_1$.

Problem 4.

The world's richest person, Jeff Bezos (net worth: \$110 billion), has decided that his true calling is data science and has enrolled in the program here at UCSD. Of course, the data science major is capped at 635 students, and Bezos had to per\$uade someone to drop out so that he could take their place. Assume that he replaced the student who previously had the highest net worth. There were 635 students before Bezos enrolled, and, because one dropped out, there are still 635 after (including Bezos).

a) By how much did the *median* net worth of DSC students increase when Bezos enrolled?

b) Assume that the student who Bezos replaced had a net worth of \$50,000. By how much did the *mean* net worth of DSC students increase when Bezos enrolled?

Problem 5.

The National Weather Service of the United States and the Servicio Meteorológico Nacional of Mexico are collaborating to predict the weather at the border between San Diego and Tijuana. To predict the temperature in January, the National Weather Service has collected n temperatures $t_1, \ldots, t_n$ in degrees Fahrenheit and computed the mean and median temperature. But because Mexico is not one of the three countries in the world that still use Fahrenheit, the Servicio Meteorológico Nacional would rather the predicted temperature be stated in degrees Celsius.

For this problem, let $g(t)$ be the function which takes in a temperature in degrees Fahrenheit and outputs the temperature in Celsius. That is, $g(t) = \frac{5}{9} \times (t - 32)$.

a) As chief data scientist at the Servicio Meteorológico Nacional, you're tasked with finding the median temperature in Celsius. Instead of first converting each temperature $t_1, \ldots, t_n$ to Celsius and finding the median of the resulting numbers, you instead simply convert $\text{Median}(t_1, \ldots, t_n)$ to Celsius. Is it true that both approaches give the same result? That is, is it the case that

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

Give a short justification of your answer. For simplicity, you may assume that there are an odd number of temperatures; this doesn't change the answer.

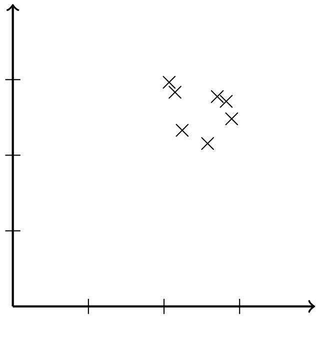
b) It is indeed true that converting the mean temperature from Fahrenheit to Celsius is the same as converting each temperature to Celsius and finding the mean. That is,

$$\text{Mean}(g(t_1), \ldots, g(t_n)) = g\left(\text{Mean}(t_1, \ldots, t_n)\right)$$

Prove mathematically that this is the case.

Problem 6.

SpaceX is trying to land their Falcon 9 rocket on a landing pad, but it keeps missing. Engineers have gathered a list $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ of the coordinates of the previous landings, where $(0, 0)$ is the center of the launchpad. When plotted, the previous landings are distributed as shown below:



The engineers are trying to predict where the next landing will be. Their prediction will be in the form of a pair of numbers, (h_x, h_y), describing the predicted horizontal and vertical position. In order to make a prediction, the engineers choose to minimize the mean squared error of their prediction:

$$\begin{aligned}
\text{MSE}(h_x, h_y) &= \text{mean squared error} \\
&= \frac{1}{n}\sum_{i=1}^{n}(\text{distance between prediction } (h_x, h_y) \text{ and } i\text{th landing } (x_i, y_i))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\sqrt{(h_x - x_i)^2 + (h_y - y_i)^2}\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[(h_x - x_i)^2 + (h_y - y_i)^2\right]
\end{aligned}$$

To minimize $\text{MSE}(h_x, h_y)$, take partial derivatives with respect to h_x and h_y, set both partial derivatives to zero, and solve for h_x and h_y in the resulting system of equations. Show that the prediction which minimizes the mean squared error is:

$$\begin{aligned}
h_x &= \frac{1}{n}\sum_{i=1}^{n} x_i = \text{Mean}(x_1, \ldots, x_n) \\
h_y &= \frac{1}{n}\sum_{i=1}^{n} y_i = \text{Mean}(y_1, \ldots, y_n)
\end{aligned}$$

Problem 7.

Consider the piecewise function:

$$f(x) = \begin{cases} \frac{1}{2}x^2 + \frac{1}{2}, & |x| \le 1 \\ |x|, & |x| > 1 \end{cases}$$

We will see next week that this function has a name, and that it plays a role in statistics.

Recall from calculus that a univariate function g is said to be *differentiable* if there is a function g' which gives the slope of f at *every* point. Even though the function f above is piecewise, it is still differentiable, as you will now show.

a) What is the slope of f at any point $x < -1$? Your answer should be a constant.

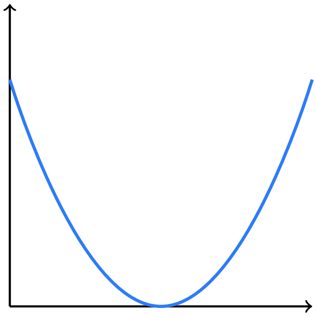
b) What is the slope of f at any point $x > 1$? Your answer should be a constant.

c) Give a formula for the slope of f that works for all x between -1 and 1.

d) What is $f'(x)$? Your answer should be a piecewise function.

Problem 8.

Suppose that f is an unknown function that is shaped like a bowl. For instance, f might be:



Suppose that at a point x_0, the slope of f is negative. Is the minimizer of f to the left of x_0 or to the right of x_0? Why?