DSC 40A - Homework 02

Due: Friday, January 24, 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.

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. Plot $L(h, y)$ as a function of h.

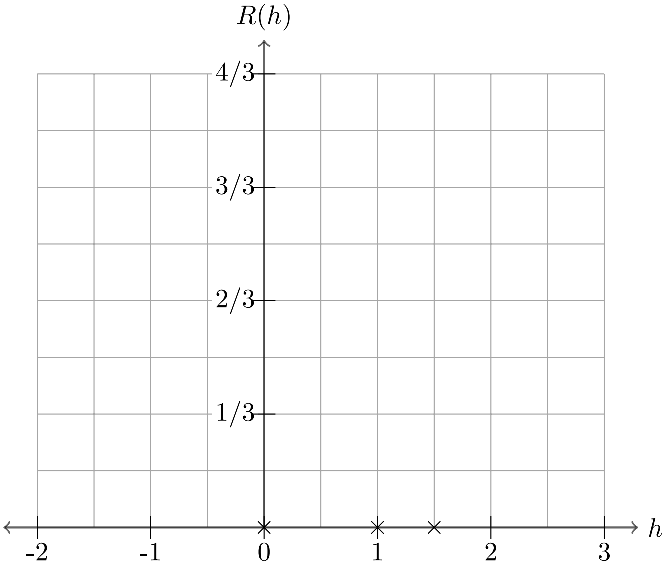
b) Suppose that we have the following data:

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

Plot the empirical risk

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

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



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), $R(h) = \frac{1}{n}\sum L(h, y)$.

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 tax of α-percent has been imposed. That is, the new avocado prices are $(1 + \alpha)y_1, (1 + \alpha)y_2, \ldots, (1 + \alpha)y_n)$. Is $(1 + \alpha)h^*$ sill 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.

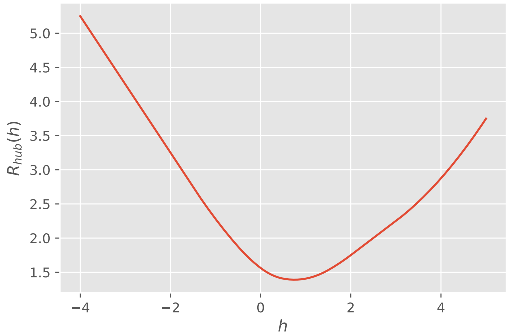
Problem 2.

The *Huber Loss* is a mixture between the square loss and the absolute loss. It is 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) What is the derivative of L_{hub} with respect to h? Your answer should also be a piecewise function.

b) Suppose $\{-\frac{1}{2}, \frac{1}{2}, 1, 4\}$ is a data set. The plot of the empirical risk, $R_{\text{hub}}(h) = \frac{1}{n}\sum L_{\text{hub}}(h, y)$ 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. To help the graders track your progress, include a table with the value of h at each iteration, such as below:

$$\begin{aligned} h_0 &= 5 \\ h_1 &=? \\ h_2 &=? \\ h_3 &=? \\ &\vdots \end{aligned}$$

Problem 3.

We have so far been concerned with predicting numerical quantities, like salaries. Now suppose we want to predict which college an incoming UCSD student will be assigned to (e.g., Warren, Sixth, Muir, etc.). Predicting a discrete category (as opposed to a number) is an important machine learning task called *classification.*

We can use empirical risk minimization to make classifications, too. Suppose we have gathered a data set of n previous students and their colleges:

Warren
Sixth
Warren
Muir
Marshall
Warren
Warren
⋮

The first step is to choose a number which will uniquely represent each college. For instance:

$$\begin{aligned} \text{Revelle} &\to 1 \\ \text{Muir} &\to 2 \\ \text{Marshall} &\to 3 \\ \text{Warren} &\to 4 \\ \text{Sixth} &\to 5 \end{aligned}$$

We then map each instance of a college to its corresponding number, giving us a new data set of numbers. For instance, the data above becomes:

4
5
4
2
3
4
4
⋮

a) Now that we have converted converted the data to a list of numbers, we can make a prediction by minimizing the mean absolute loss. Explain why this is not a good idea.

b) The *zero-one* loss is defined as follows:

$$L_{01}(h, y) = \begin{cases} 0, & h = y, \\ 1, & h \neq y. \end{cases}$$

As usual, define the risk to be:

$$R_{01}(h) = \frac{1}{n} \sum_{i=1}^{n} L_{01}(h, y)$$

Notice that $R_{01}(h)$ can be interpreted as the misclassification rate. That is, if $R_{01}(h) = .7$, then predicting h would result in the wrong answer for 70% of the data points. Given the data set $\{1, 1, 1, 2, 2, 3, 4, 4, 4, 4\}$, plot the empirical risk $R_{01}(h)$ for $h \in [0, 5]$.

Hint: the function should have point discontinuities.

c) We have seen that the median minimizes the risk when the absolute loss is used, and that the mean minimizes the risk when the square loss is used. What quantity minimizes the risk when the zero-one loss is used?

d) Is gradient descent useful for minimizing the risk with zero-one loss? Why or why not? Make reference to your plot of the risk in your answer.

Hint: the risk is indeed non-convex, but gradient descent can still be useful for minimizing non-convex functions. Is there some other reason?

Problem 4.

The gradient descent update rule for minimizing a function $R(h)$ is:

$$h_{\text{next}} = h_{\text{prev}} - \alpha \frac{dR}{dh}(h_{\text{prev}}).$$

We said in class that the sign of dR/dh is meaningful: if it is positive we should move to the left, and if it negative we should move to the right.

Why is the *magnitude* of the derivative useful, too? That is, what is wrong with using the update rule:

$$h_{\text{next}} = h_{\text{prev}} - \alpha \cdot \operatorname{sign}\left(\frac{dR}{dh}(h_{\text{prev}})\right),$$

where $\operatorname{sign}(\cdot)$ returns the sign of its argument as either zero or one. For instance, $\operatorname{sign}(-4) = -1$ and $\operatorname{sign}(42) = 1$.

Problem 5.

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.

Suppose that $f_1(x)$ and $f_2(x)$ are convex functions defined on all real numbers. We wish to show that their sum, $f(x) = f_1(x) + f_2(x)$, is also a convex function.

One way to show that f is convex is to prove that it satisfies the definition. That is, it is sufficient to show that for any real numbers a and b and for all $t \in (0, 1)$,

$$f(ta + (1-t)b) \leq tf(a) + (1-t)f(b).$$

Prove that this holds by using a chain of inequalities.

Hint: when proving something like this, first identify the special things that we know about the important entities in the problem. In this case, we know that 1) f is the sum of f_1 and f_2; and 2) f_1 and f_2 are convex functions. We will need to use both pieces of information in our proof. Which should we use first? If you get stuck, ask yourself: have I used all of these pieces of information yet?

Problem 6.

Remember that there are several ways of showing that a function is convex:

1. From the definition.
2. Use the second derivative test for convexity.
3. Show that the function is built from other convex functions using one or more of the properties mentioned in lecture (i.e., it is the sum, composition, or pointwise maximum of convex functions).

In each of the problems below, use one of the above justifications to prove that the function is convex.

a) $f(x) = x$

b) $f(x) = e^x$

c) $f(x) = |x|$

d) The mean absolute error (as a function of h; consider the y_i as fixed):

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

Problem 7.

Linear algebra plays an important role in machine learning, and we will be using it soon. These questions will help you remember some of the basics from your course on linear algebra.

a) Define the matrix X and the vector $\vec{a}$ as below:

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{pmatrix}, \qquad \vec{a} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}.$$

One way to interpret the result of matrix-vector multiplication is that it is a mixture of the columns of X. To see this, show that $X\vec{a} = a_1\vec{x}^{(1)} + a_2\vec{x}^{(2)} + a_3\vec{x}^{(3)}$, where

$$\vec{x}^{(1)} = \begin{pmatrix} x_{11} \\ x_{21} \\ x_{31} \end{pmatrix}, \qquad \vec{x}^{(2)} = \begin{pmatrix} x_{12} \\ x_{22} \\ x_{32} \end{pmatrix}, \qquad \vec{x}^{(3)} = \begin{pmatrix} x_{13} \\ x_{23} \\ x_{33} \end{pmatrix}$$

are the columns of X.

b) Recall that the dot product of two vectors $\vec{u} = (u_1, u_2, u_3)^\intercal$ and $\vec{v} = (v_1, v_2, v_3)^\intercal$, written $\vec{u} \cdot \vec{v}$, is the number $u_1v_1 + u_2v_2 + u_3v_3$.

Let $\vec{u}$ and $\vec{v}$ be as above, and let $\vec{w} = (w_1, w_2, w_3)^\intercal$. Show that

$$\vec{u} \cdot (\vec{v} + \vec{w}) = \vec{u} \cdot \vec{v} + \vec{u} \cdot \vec{w}$$

by writing out the left hand side and the right hand side in terms of the elements of each vector and showing that they are equal.