DSC 40A - Midterm 01

February 4, 2020

Name: SOLUTIONS

PID:

By signing below, you are agreeing that you will behave honestly and fairly during and after this exam. You should not discuss any part of this exam with anyone enrolled in the course who has not yet taken the exam (this includes posting questions about the exam on Piazza!)

Signature:

Name of student to your **left**:

Name of student to your **right**:

(Write "N/A" if a wall/aisle is to your left/right.)

Instructions:

- Write your solutions to the following problems in the boxes provided.
- Scratch paper is provided at the end of the exam.
- No calculators are permitted, but a cheat sheet is.
- Write your name or PID at the top of each sheet in the space provided.

Tips:

- Show your work to receive partial credit.
- Look through the entire exam before starting.
- Make good use of all assumptions given to you.

(Please do not open your exam until instructed to do so.)

Problem 1. (25 points) *Miscellaneous.*

a) True or False. If a function f is convex, then gradient descent is guaranteed to find the global minimum of f no matter what learning rate is chosen.

○ True ● False

Solution: The value of the function can increase between steps of gradient descent if the function is non-convex or the learning rate is chosen to be too large.

b) True or False. When predicting a single number from a data set of numbers $y_1, \ldots, y_n$, the "mean absolute error" is minimized by the mean of $y_1, \ldots, y_n$.

○ True ● False

Solution: It is minimized by the *median.*

c) True or False. The mean squared error of the least squares regression line of a data set $\{(x_1, y_1), (x_2, y_2)\}$ with two points must be zero.

● True ○ False

Solution: Given any two points, there is a straight line that goes directly through the two points. This line has a zero SSE, and so the least squares solution must, too.

d) True or False. Let w_1 be the slope of the least squares regression line of a data set (x_i, y_i). Then the slope of the regression line for the data set $(2x_i, y_i)$ is $4 \cdot w_1$.

○ True ● False

Solution: The slope should *decrease* when the scale of the x's is increased.

e) Suppose we collect a data set of 1000 temperatures in Fahrenheit. We will run gradient descent to find the prediction minimizing the mean squared error. What is a reasonable choice of initial hypothesis, h_0, and learning rate parameter, α?

$h_0 =$ 70 $\alpha =$ 1/20

Solution: A reasonable starting location is a typical temperature – so, around 70 in San Diego.

We want the size of our steps to be small enough that we'll converge. "Small enough” isn’t a technical term, but we’ll say a step of size 1 degree is relatively small.

The size of the step is $\alpha \cdot |dR/dh|$, where α is the learning rate. We want this to be around 1, so we have to solve for a learning rate that will yield such a small step size.

The mean squared error is $\frac{1}{n}\sum_{i=1}^n (h - y_i)^2$. The derivative of this is

$$\frac{dR}{dh} = \frac{2}{n}\sum_{i=1}^n (h - y_i)$$

If we have a data set of temperatures in San Diego, they range from about 60 to 80. So a safe guess for the magnitude of $h - y_i$ is around 20 degrees. The derivative is twice the average difference, so we’ll end up with something like

$$\left|\frac{dR}{dh}\right| = 40$$

This means that α should be around 1/40.

There are a few places here where our guesstimates could be different. For instance, maybe our starting location isn’t chosen well. Then the magnitude of $h - y_i$ could be considerably larger – say, 100 degrees. Then the magnitude of the derivative is around 200, and a good step size is about 1/200. This was our lower bound on an acceptable answer.

What is the largest learning rate we’d accept? Suppose a good step size was 2 (this is probably too large, but we’ll work with it.) If we say that a typical magnitude of $h - y_i$ is 2 degrees, then the magnitude of the derivative is 4, and we need $\alpha = 1/2$.

Problem 2. (25 points) *Convexity.*

Determine whether each of the below functions is convex. If the function **is** convex, give a short justification for why (pictures alone will receive partial, but not full, credit). If the function is **non**-convex, you do not need to provide justification.

a) $f(x) = \begin{cases} e^x, & x > 0 \\ 1 - x, & x \leq 0 \end{cases}$

● convex ○ non-convex

Solution: Both e^x and $1 - x$ are convex from the second derivative test. f is the elementwise maximum of these functions, and is therefore convex.

b) $f(x) = x^2 + \cos x$

● convex ○ non-convex

Solution: Second derivative test. Second derivative of x^2 is 2. Second derivative of cos is something like cos or negative cos... I can't remember, but whatever it is, it is bigger than -1. Since $2 - 1 \geq 0$, this is convex.

c) $R_{\text{UCSD}}(h) = \frac{1}{n} \sum_{i=1}^{n} \left[1 - e^{-(h-y)^2/\sigma^2} \right]$, where the y_i's are salaries and $\sigma = 1{,}000{,}000$.

○ convex ● non-convex

Solution: The $\sigma = 1{,}000{,}000$ clue is a red-herring. It doesn't matter how large we make σ, if we zoom out enough, we will see that R will be non-convex.

d) $R(h) = \frac{1}{n} \sum_{i=1}^{n} |h - y_i|^3$

● convex ○ non-convex

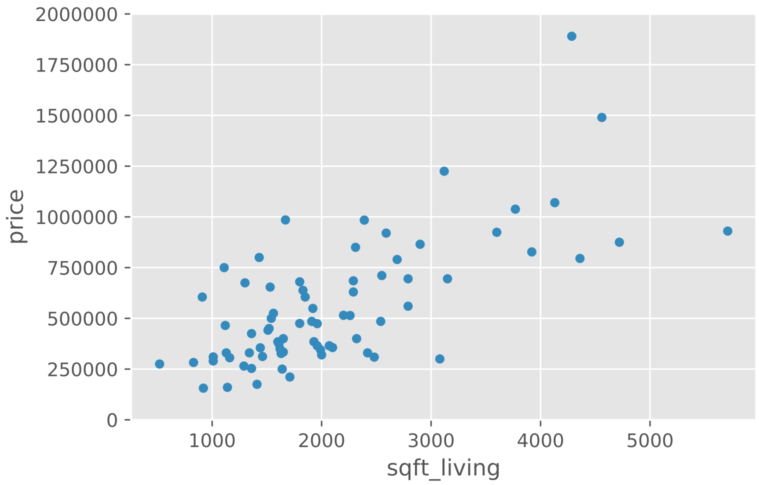
Solution: $|h - y_i|^3$ is convex from the second derivative test. Its second derivative is a piecewise function:

$$\begin{cases} -6(h - y_i), & h < y_i \\ 6(h - y_i), & h \geq y_i \end{cases}$$

R is the sum of convex functions, and so it is convex.

Problem 3. (25 points) *Regression.*

The plot below shows the relationship between the price and the size (in square feet) of 75 homes in Seattle.



a) Suppose least squares is used to fit a linear prediction rule of the form $H(x) = w_1 x + w_0$ to this data, where x represents the size of a home in square feet. Give reasonable guesses for the value of the parameters w_1 and w_0 which minimize the mean squared error.

$w_0 =$ 100,000

$w_1 =$ 250

Solution: It looks as though the regression line would go from around 250,000 at 1,000 sqft to around 1.5 million at 6,000 sqft. This is a slope of 1.25 million / 5,000 = 250. The intercept is somewhat less clear; it looks to be anywhere from zero to 250,000.

b) Suppose that fitting a straight line to the data shown above results in a mean squared error of c_1. Now suppose you add another data point and fit another line to the new data. Suppose that the mean squared error of this line is c_2. Is it possible that $c_2 < c_1$? Give a short justification.

● Yes, c_2 can be smaller than c_1. ○ No, $c_2 \geq c_1$.

Solution: Suppose that our new point is placed at $(\bar{x}, \bar{y})$. This causes no change in the slope or intercept of our regression line, and since the regression line passes through $(\bar{x}, \bar{y})$, there is no error associated with it. Hence the *total* squared error is the same as before, but n is one larger, so the mean squared error is smaller.

c) The scatter plot shows only a small sample of 75 homes from a much larger data set of 21,000 homes. Suppose you fit a line to the full data set and find that w_1 is 4.2 (much smaller than you were expecting it to be). Assuming that the overall trend in house prices is still evident in the larger data set, give a reasonable explanation for why the slope of the fitted line might be so much smaller than expected. Give a short informal justification of your reason, making reference to the least squares equation for the slope:

$$w_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Solution: There might be an outlier in home size. For instance, suppose that there is a home listed at 999,999 sq ft.

With 999,999 as one of the x values, $\bar{x}$ becomes quite large: it is on the scale of 500,000. Therefore, the magnitude of $(x_i - \bar{x})$ will typically be on the order of 500,000. $(y_i - \bar{y})$ will also typically be on the order of several hundred thousand dollars. Therefore the numerator is on the order of n times $500{,}000^2$, which is the same as the denominator. So we expect to see an answer on the order of 1.

An outlier such as this can occur in a real data set if it is messy. For instance, sometimes people (unfortunately) use “impossible” numbers like 999,999 to denote a missing value.

Problem 4. (25 points) *Gradient Descent.*

Suppose we have collected data $\{0, 1, 2, 5\}$. Let $L(h, y) = (h - y)^4$, and define the risk to be: $R(h) = \frac{1}{n}\sum_{i=1}^{n} L(h, y_i) = \frac{1}{4}\sum_{i=1}^{n}(h - y_i)^4$. Run **one** iteration of gradient descent on R using the data above, a learning rate of $\alpha = 1/36$ and an initial prediction of $h_0 = 2$. Partial credit will be assigned for attempting to perform each part of the gradient descent update, so if you get stuck in one part, make sure to try and do the rest.

Solution:

First, we need the derivative of R. We have:

$$\frac{dR}{dh} = \frac{d}{dh}\frac{1}{n}\sum_{i=1} L(h, y_i) = \frac{1}{n}\sum_{i=1}\frac{d}{dh}L(h, y_i)$$

We'll take the derivative of the loss now:

$$\begin{aligned}\frac{d}{dh}L(h, y_i) &= \frac{d}{dh}(h - y_i)^4 \\ &= 4(h - y_i)^3\end{aligned}$$

So

$$\begin{aligned}\frac{d}{dh}R(h) &= \frac{1}{n}\sum_{i=1}^{n}\frac{dL}{dh}(h) \\ &= \frac{1}{n}\sum_{i=1}^{n} 4(h - y_i)^3\end{aligned}$$

If we use the fact that $n = 4$, then:

$$\frac{d}{dh}R(h) = \sum_{i=1}^{4}(h - y_i)^3$$

Now we run gradient descent. The new position is:

$$\begin{aligned}
h_0 - \alpha \frac{dR}{dh}(h_0) &= h_0 - (1/36)\sum_{i=1}^{4}(h - y_i)^3 \\
&= 2 - \frac{1}{36}\left((2-0)^3 + (2-1)^3 + (2-2)^3 + (2-5)^3\right) \\
&= 2 - \frac{1}{36}(8 + 1 + 0 - 27) \\
&= 2 - \frac{1}{36}(-18) \\
&= 2 + \frac{18}{36} \\
&= 2.5
\end{aligned}$$

Prediction after first iteration: 2.5

Before turning in your exam, please check that your name is on every page.