CSE 151A - Homework 05

Due: Wednesday, May 6, 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 Wednesday at 11:59 p.m.

Essential Problem 1.

Consider the function $h(x) = (x^3 - x + 1)^2$. It is easy to verify that this is a non-convex function. We know that for a convex function with the appropriate choice of the learning rate, gradient descent converges to the global minimum. However for a non-convex function, since there can be multiple local minima, gradient descent can converge to any local minima. An intuitive workaround for this problem then is to perform gradient descent from multiple initial values of x and then find the minimum among all the local minima in order to get a better estimate of the global minima.
Run 5 iterations of gradient descent with learning rate 0.05 for $h(x)$ with the following two initial values of x and find the global minima of $h(x)$ (given that $h(x)$ has only two local minima):

1. $x_0 = 0$
2. $x_0 = -1$

Note: you should run at least the first two iterations of gradient descent "by hand", without using code, so as to get a feeling for the procedure. You may use code (including the code in the demo notebook posted during lecture) to do the remaining iterations.

Solution: First, we calculate the gradient of the function $h(x) = (x^3 - x + 1)^2$

$$h'(x) = 2(x^3 - x + 1)(3x^2 - 1)$$

Following the update rule $x_i = x_{i-1} - \alpha \cdot h'(x_{i-1})$ for five iterations gives us the following:

1. Starting with $x_0 = 0$

$$\begin{aligned} x_1 &= 0 - 0.05 \cdot (2(0 - 0 + 1)(0 - 1)) = 0.1 \\ x_2 &= 0.1 - 0.05 \cdot (2(0.1^3 - 0.1 + 1)(3(0.1)^2 - 1)) = 0.187 \\ &\vdots \\ x_5 &= 0.370 \end{aligned}$$

2. Starting with $x_0 = -1$

$$\begin{aligned} x_1 &= 0 - 0.05 \cdot (2(1)(2)) = -1.2 \\ x_2 &= 0.1 - 0.05 \cdot (2(0.472)(3.320)) = -1.357 \\ &\vdots \\ x_5 &= -1.305 \end{aligned}$$

If we assume both of these x-values have achieved local minima, we can find the global minimum by

looking at their function values.

$$h(0.370) = 0.463$$
$$h(-1.305) = 0.007$$

Since $h(-1.305) < h(0.370)$, we can assume that the global minimum is achieved at $x = -1.305$ with a value of 0.007.

In fact, if you run a few more iterations of gradient descent for both the above initial values, you will find that the value of x when starting with the initial value $x_0 = 0$ converges to around 0.577, and the value of x when starting from the initial value $x_0 = -1$ converges to around -1.325 with a true global minimum function value of 0.

Essential Problem 2.

Suppose you have gathered the data below:

Feature 1	Feature 2	Class
-1	2	1
2	-3	1
1	0	1
-3	1	-1
1	1	-1

This data is just randomly generated and doesn't mean anything, but you can think of each row as being a person, and each feature is a different measurement about that person (like their height or weight).

Suppose you use a logistic regression model to predict the class label. You train your model using gradient descent to find the weight vector $\vec{w}$.

Below are two vectors, $\vec{w}^{(1)}$ and $\vec{w}^{(2)}$. One of them is the weight vector found by gradient descent, and the other is a vector I made up. Which is the vector found by gradient descent? Explain how you know this.

$$\vec{w}^{(1)} = (-0.31, -1.29, 3.90)^T$$
$$\vec{w}^{(2)} = (-0.22, -1.41, 3.95)^T$$

Solution: Gradient descent (applied to $-f(x)$ in this case) should give us some 'optimal' $\vec{w}$. In the case of logistic regression, the optimal $\vec{w}$ should maximize the log-likelihood function.

$$\log L(\vec{w}) = -\sum_{i=1}^{n} \log(1 + \exp(-y_i \vec{w} \cdot \text{Aug}(\vec{x}^{(i)})))$$

We can compute the log-likelihood of $\vec{w}^{(1)}, \vec{w}^{(2)}$ as

$$\log L(\vec{w}^{(1)}) = -26.23$$
$$\log L(\vec{w}^{(2)}) = -27.07$$

Since the log-likelihood of $\vec{w}^{(2)}$ is less than that of $\vec{w}^{(1)}$, we know that it cannot have achieved a maximum value and thus isn't the result of gradient descent. Therefore, $\vec{w}^{(1)}$ must be the one found by gradient descent.

Essential Problem 3.

For each of the following functions, determine if it is convex or not using the second derivative test. Show your work in each case.

a) $f(x) = 3x^3 + 2x - 4$

Solution:

$$\frac{df}{dx} = 9x^2 + 2$$
$$\frac{d^2f}{dx^2} = 18x$$

We find that $18x \leq 0$ for $x \leq 0$, thus the function is not convex.

b) $f(x) = \dfrac{e^x + e^{-x}}{2}$

Solution:

$$\frac{df}{dx} = \frac{1}{2}(e^x - e^{-x})$$
$$\frac{d^2f}{dx^2} = \frac{1}{2}(e^x + e^{-x})$$

Since e^x can never be negative, our second derivative must always be greater than zero. Thus the function is convex.

c) $f(x, y) = x^2 + y^2 - 5xy + 10x + 12y - 42$

Solution:

$$\frac{\partial f}{\partial x} = 2x - 5y + 10$$
$$\frac{\partial f}{\partial y} = 2y - 5x + 12$$
$$H(x, y) = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} = \begin{bmatrix} 2 & -5 \\ -5 & 2 \end{bmatrix}$$

We want to check that our Hessian is positive semi-definite (PSD). So, for some arbitrary vector $\vec{z} = (z_1, z_2)^\mathsf{T}$ we need $\vec{z}^\mathsf{T} H(x, y)\vec{z} \geq 0$.

$$H(x, y)\vec{z} = \begin{bmatrix} 2z_1 - 5z_2 \\ -5z_1 + 2z_2 \end{bmatrix}$$
$$\vec{z}^\mathsf{T} H(x, y)\vec{z} = 2z_1^2 - 5z_1z_2 - 5z_1z_2 + 2z_2^2$$
$$= 2(z_1^2 - 5z_1z_2 + z_2^2)$$

Suppose $\vec{z} = (1, 1)^\mathsf{T}$, then $\vec{z}^\mathsf{T} H(x, y)\vec{z} = -6$, thus our Hessian is not PSD. Therefore the function is not convex.

d) $f(\vec{x}) = \|\vec{x}\|^2$

Solution: We can solve our derivatives for an arbitrary x_i or x_j to make generalizing easier.

$$f(\vec{x}) = x_1^2 + x_2^2 + \cdots + x_d^2$$
$$\frac{\partial f}{\partial x_i} = 0 + \cdots + 2x_i + \cdots + 0 = 2x_i$$
$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \begin{cases} 2 & \text{If } i = j \\ 0 & \text{If } i \neq j \end{cases}$$

So our Hessian takes on a familiar form,

$$H(\vec{x}) = \begin{bmatrix} 2 & 0 & \cdots & 0 \\ 0 & 2 & & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & & 2 \end{bmatrix}$$
$$= 2I_d$$

It can then be shown that for any arbitrary $\vec{z} \in R^d$, we have $\vec{z}^\intercal(2I_d)\vec{z} = 2(\vec{z}^\intercal\vec{z})$, which can be expressed as $2\|\vec{z}\|^2$. This is always greater than or equal to zero, thus our Hessian is PSD. Therefore the function is convex.

Plus Problem 1. (6 plus points)

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:

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

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

Solution: The magnitude of the derivative is useful in telling us when to stop the gradient descent procedure. If we did not use the magnitude, every step would be of size α and we would not be able to tell when we have converged. On the other hand, the magnitude of the derivative tells use the steepness of the function at the current position; if it is near zero, we may be near a local minimum.

Note: I'll offer a larger-than-average number of plus points in next week's homework.