Calculus for Machine Learning

Caroline Sun *

December 2020

1 Introduction

Calculus is going to be an *integral* part of our next few lectures regarding neural networks. As a result, this is going to be a crash course into derivatives and partials—if you'd like to get into more depth, check out the resources at the end. While it's going to be difficult to work with this material if you haven't been in a calculus class before, try to stick with it: it'll be worth it.

2 Derivatives

Simply put, a derivative is just the rate of change of a function at a given point. Even without calculus, we've dealt with rate of change and average rate of change before. Let's take a look at a basic example: lines.

2.1 Derivative and Slope

We already recognize slope as the rate of change in a line (by definition).

$$\text{slope} = \frac{\Delta y}{\Delta x} = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

In the example y = 3x + 8, let's find the rate of change given our equation above.

$$\text{slope} = \frac{\Delta y}{\Delta x} = \frac{f(1) - f(0)}{1 - 0} = \frac{11 - 8}{1 - 0} = 3$$

In a line, the rate of change at any point is constant, it's already defined by the slope! Let's take a look at more complex curves.

2.2 Curves, Secant Slopes, and Derivative

Given the parabola $y = x^2$, let's try to find the rate of change at x = 1.

*based on Kevin Fu's lecture of the same name

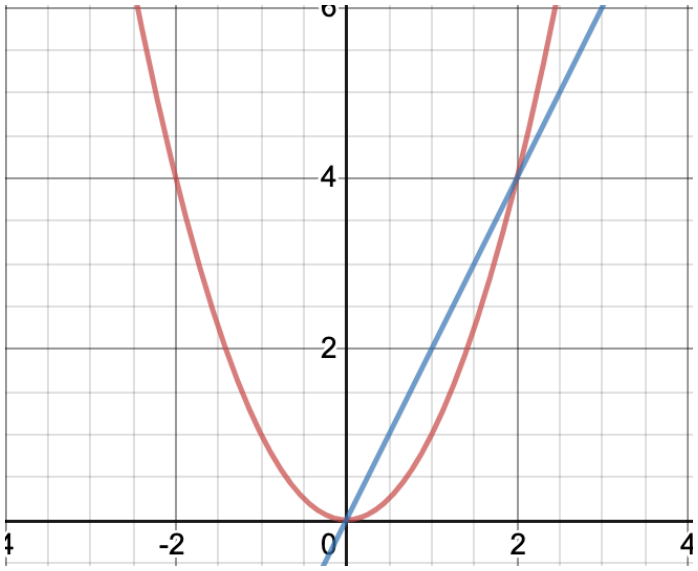

Figure 1: y = x^2 with secant y = 2x

We could approximate this value by finding the average rate of change, A, of the function between x = 1 and a close number.

$$A = \frac{f(b) - f(a)}{b - a} \tag{1}$$

Let's start by finding the average rate of change between x = 0 and x = 2, then. We would get $\frac{f(2)-f(1)}{2-1} = \frac{4-1}{2-1} = 3$. Now, let's take a closer look, and zoom into the average rate of change from x = 1 to x = 1.25. $\frac{f(1.25)-f(1)}{1.25-1} = \frac{1.5625-1}{1.25-1} = 2.25$. Even closer, let's try x = 1 to x = 1.001, which would give us a slope of 2.001. Does it look like it's approaching a number?

Essentially, we are trying to use our equation for the average rate of change, and find a b-value as close as we possibly can to our a value. Mathematically, that's a limit that we can express by the following expression.

$$\frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x} \tag{2}$$

This equation is how we can formally describe a derivative mathematically, not words. The ways we usually express derivative of function are most commonly f'(x), $\frac{dy}{dx}$, $\frac{d}{dx}$f(x), and $\frac{df}{dx}$.

As a side note, a common real world functions of derivatives includes finding the velocity of a graph at a time (by using the graph/equation of position versus time), which is often used in physics.

2.3 Run-through of Taking the Derivative of a Function

Fair warning, this section is going to be the most dense of this lecture. Granted, it's not a fully replacement of the first month of calculus, but we're going to try to distill essential information to take the derivative of the sigmoid function later.

Luckily for us, we don't have to use the limit definition of a derivative to find the derivative of a function. We have several essential derivative rules that, once established, we can use to find the derivative of complex functions.

First, let's just remember that for any equation f(x) = c, f'(x) = 0. Why? There is no change. We can circle back to the our explanation of a derivative of a line. In this function, the slope is also 0.

The first derivative rule is known as the Power Rule. This is because it applies for any function of the form $f(x) = x^n$. For any function of that form, the derivative will be $f'(x) = nx^{n-1}$. We can use the power rule on our parabola $y = x^2$ example in 2.2, and confirm that the derivative at x = 1 is 2.

A special function is $f(x) = e^x$, where the derivative will also equal $f'(x) = e^x$. The derivative of its inverse, ln(x), is $\frac{1}{x}$.

However, most functions we approach are not that simple. As a result, we have the chain rule, the sum rule, and the product/quotient rules.

The chain rule helps us find the derivative for a function containing layers of a basic function. For example, the function $f(x) = e^{3x^2}$ is a function within a function, the outermost function being e^y (where $y = 3x^2$) and the inner function being $3x^2$. We can't just use the e^x rule on this function: a change in x will also affect the inner function as well. In the chain rule, we will take the derivative of the outer function, and multiply that by the derivative of the inner function. Giving the outermost function the name f(x) and inner function g(x) in this example, the chain rule is

$$\frac{d}{dx}f(g(x)) = f'(g(x))g'(x) \tag{3}$$

Here, we would take the derivative of the big function f'(g(x)) which equals e^{3x^2} and multiply it by g'(x), which is 6x per the power rule, resulting in $\frac{d}{dx}(e^{3x^2})$ $= 6xe^{3x^2}$.

The chain rule also works for deeper functions. If we had three layers of functions, the chain rule would extend to

$$\frac{d}{dx}f(g(h(x))) = f'(g(h(x)))g'(h(x))h'(x) \tag{4}$$

Thankfully, the next rule, the sum rule, is much less complex than the chain rule. For a function f(x) = h(x) + g(x), where h(x) and g(x) are smaller functions, f'(x) = h'(x) + g'(x). Through the sum rule, we know that the derivative of $f(x) = e^x + x^2$ is $f'(x) = e^x + 2x$.

Next, we can use the product rule. For a function f(x) = h(x)g(x), its derivative is

$$f'(x) = h(x)g'(x) + h'(x)g(x) \tag{5}$$

An example is f(x) = $3x^2e^x$, where f'(x) = $6x \cdot e^x + 3x^2e^x = 3xe^x(x + 2)$.

The last relevant rule is the derivative rule is the quotient rule. For a function f(x) = $\frac{g(x)}{h(x)}$, its derivative is

$$f'(x) = \frac{h(x)g'(x) - h'(x)g(x)}{h(x)^2} \tag{6}$$

which can be remembered as "down d-up minus up d-down over down down."

This is not an extensive list of derivative rules! It's just the basics of what we need to solve our sigmoid function. For starters, it doesn't include trigonometric derivatives. If you're curious about those, check out the references section.

3 Sigmoid Derivative

The sigmoid function is

$$S(x) = \frac{1}{1 + e^{-x}}$$

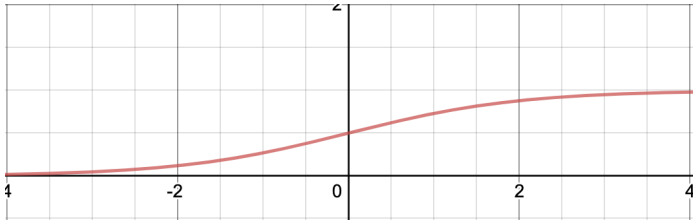

Figure 2: Sigmoid function

and will be relevant in future neural network lectures as a common activation function. We can use f'(x) = $\frac{h(x)g'(x)-h'(x)g(x)}{h(x)^2}$ as a first step to tackling this, with h(x) = 1 and g(x) = $1 + e^{-x}$. In this case, h'(x) = 0, as 1 is a constant, and g'(x) = $-e^{-x}$ per the sum rule and chain rule on $1 + e^{-x}$. Therefore, using the quotient rule we'd get S'(x) = $\frac{e^{-x}}{(1+e^{-x})^2}$.

Can we simplify this further by writing S'(x) in terms of S(x)?

let's rewrite our answer to be

$$(\frac{1}{1 + e^{-x}})\frac{e^{-x}}{1 + e^{-x}} = S(x)\frac{e^{-x}}{1 + e^{-x}}$$

and if we further split the right hand side, we get

$$= S(x)\frac{(1 + e^{-x}) - 1}{1 + e^{-x}} = S(x)(\frac{1 + e^{-x}}{1 + e^{-x}} - \frac{1}{1 + e^{-x}})$$

giving us

$$S'(x) = S(x)(1 - S(x)) \tag{7}$$

This equation fully gives us S'(x) in terms of S(x), allowing us to only use the S(x) function to find S'(x).

4 More Dimensions

Another use of derivatives in machine learning is for training neural networks, known as gradient descent. At its simplest, a neural network is a chain of weighted sums, and we use the error of our network, or the cost, to change our weighted sums to maximize accuracy. Every layer of sums affects the next, thus many variables end up changing the cost. We can't just use a derivative now, as our graph will be n-dimensional, and derivatives apply for the 2-D. Instead, we use a gradient, ∇f. This gradient is analogous to a derivative in a multi-variable context. While we won't have to manually find gradients in creating neural networks, we'll try to grasp what a gradient means within a simple example.

In the graph, $z = x^2 + y^2$, it's a parabaloid.

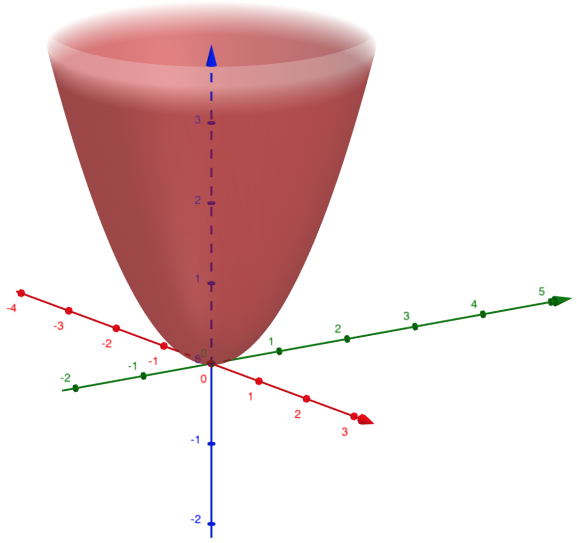

Figure 3: Paraboloid $z = x^2 + y^2$

Imagine laying a small ball onto the paraboloid: it'll roll downwards, stopping at (0, 0, 0). The negative of the gradient gives us the specific direction it will fall down to.

Finding the gradient is using the same derivatives we found earlier, except it's a vector, with each component being the derivative of the function with respect

to one variable (and treating the other variables, temporarily, as constants). These component derivatives are called partials. The partials of z with respect to x and y, are the following:

$\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}$

and our gradient is just the those put into vector form, so

$\nabla z = \langle \frac{\partial z}{\partial x}, \frac{\partial z}{\partial y} \rangle$

In our paraboloid example, the gradient would be $\langle 2x, 2y \rangle$

If we plug in a value into our gradient, we'll get the direction of the steepest incline. As a result, if we took the negation of that vector, it'd point to the direction of steepest descent. In this example, the direction of steepest descent would be at $\langle -6, -8 \rangle$ at (3, 5, 34). This can be generalized to other functions and points.

5 Closing

Hopefully, conceptually, derivatives and gradients make sense. However, further resources for both will be in the next section.

6 Resources

- https://www.khanacademy.org/math/differential-calculus/dc-diff-intro
- https://www.mathsisfun.com/calculus/derivatives-rules.html
- https://www.youtube.com/watch?v=WUvTyaaNkzMlist=PL0-GT3co4r2wlh6UHTUeQsrf3mlS2lk6x
- https://www.khanacademy.org/math/multivariable-calculus/multivariable-derivatives/partial-derivative-and-gradient-articles/a/the-gradient
- https://betterexplained.com/articles/vector-calculus-understanding-the-gradient/

7 References

Desmos and Geogebra both used to generate graphs.