Simple Equations Worth Thinking About Again

Raymond Heberer
ray.heberer@gmail.com

May 5, 2018

1 Introduction

Ideas from mathematics underly virtually every technique and concept in Machine Learning (ML). While possessing a rigorous understanding of all these ideas is certainly not required, it can be beneficial. For the beginning practitioner, the scope of the math that forms the foundation of ML is intimidating, and many choose to reassure themselves by observing that modern programming frameworks mean that the ability to implement an algorithm is far removed from the ability to understand its mathematical underpinnings.

This "top-down" approach, emphasizing results over theory, is both effective and marketable. However, there are certain mathematical foundations of ML that are worth becoming familiar with.

In what follows, I state some of the most important equations a beginner should seek to understand when approaching ML. I then attempt to offer commentary on each that provokes some thought. While the choice of topics was motivated by which ideas a fundamental to ideas in ML, they are not unique to it, and can be understood independently, without motivation to study ML.

The two main groups of objects from which these fundamentals are taken will involve **Function and Derivatives** and **Vectors and Matrices**. In addition, I will select and explore a couple statistical techniques. For these, some understanding of linear algebra and probability will be needed in order to benefit fully from their analysis.

I hope to show two things by the end of this series. First, that some ideas that seem very basic are worth revisiting even if they appear familiar or trivial. Second, that focusing in on selected topics or techniques that appear challenging can be a valuable exercise, even if there is not enough time for a practitioner to analyze every tool in their arsenal with such depth.

2 Functions and Derivatives

2.1 Variables and Function Notation

$$y = f(x)$$

The result of some secondary school math training is that the above statement appears unremarkable, uninformative, or confusing. One might recall a moment of confusion in that pre-calculus course where the vertical axes of all the graphs in assignments went from being labeled as y to $f(x)$ (or perhaps the other way around) without the instructor acknowledging the change, and the students feeling too naive to ask "so... are they just the same thing?"

In my view, $y = f(x)$ should be viewed with a certain amount of respect for how much information can be condensed into modern notation. On one level, it is just saying "we have some variable named x, and a function of that variable, f, and the output of that function can be assigned to another variable, which we will name y." But think of just what an algebraic variable is. Think of what a function is, and how it can be characterized by things such as its domain and range.

Could the output of that function of x have been given a different name, say, z? Certainly. Could we for the sake of economy not dedicate a new variable to the output, and always refer to it as $f(x)$? Of course.

It's with simple, almost redundant mathematical statements such as $y = f(x)$ that the amount of information contained in a few strokes of ink becomes clear, as does the interface between ideas, the conventions set to facilitate their communication and manipulation, and the history of those conventions. Do not scoff at them unless you can explain the essence of their contents to a child.

2.2 Equation of a Line

$$y = mx + b$$

We've seen how $y = f(x)$ could be motivation to pause and consider, among other things, the nature of algebraic variables and how their values can sometimes be intertwined in ways exactly captured by this idea of functions. The familiar, generic equation of a line then is an opportunity to consider how functions themselves are characterized and grouped.

In other words, if before we were pondering the x and the y, now is a time to look more closely at the m and the b. What roles do they play? How do they show up in different forms and families of functions? What even is meant by "families" of functions?

Looking at a line, one can characterize it completely by noting how it's tilted (slope), and identifying one point it passes through (the y-intercept lends itself to conveniently to form a simple equation). Though the two variables that are related linearly range across the real numbers, the two defining features are fixed for a single line.

Yet there is the sense that these two quantities which characterize lines could have taken on different values, and indeed they can. What will result from this will be more lines. Different lines for different values for the two characteristics, but all lines, no squiggly curves. So m and b are in a sense parameters that, when assigned, completely designate one instance of a line, and in their generality and potential to take on any out of some set of numbers, define a "family" of functions.

Now, this understandably seems like I'm making much ado about nothing, but I think the idea becomes powerful when one thinks about how functions are built. There are only really three ingredients: numbers, variables, and other functions. There are also just three ways of combining them: adding them together, multiplying them or, given two or more functions, composing them (taking one function of another function).

Objects like x and y are one ingredient - the variables - and now we can see that m and b are another ingredient - the numbers. The notion of a "family" of functions can then be thought of as the collection of functions that can be made given some amount of variables in a particular arrangement, allowing the numbers these variables are added to and multiplied by to vary arbitrarily.

2.3 The Power Rule

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(x^n\right) = nx^{n-1}$$

Consider the function $f(x) = x^2$. One way to visualize this is to think of f as the area of a square as a function of its side-length x. Now consider: can we define a new function that represents the change in f when x is increased by a small, yet nonzero, amount dx. This will also be a function of x; let's call it df. The precise value of dx doesn't matter, but it is fixed; think of it like a parameter, like b in the equation for a line.

This new function will now be given by $df(x) = f(x + dx) - f(x) = (x + dx)^2$. Expanding, we get:

$$df = (x^2 + 2xdx + dx^2) - x^2 = 2xdx + dx^2$$

Thinking back to our square, dx is then a little piece of additional length added to both the width and height. That makes df the total additional area added to the square. This area consists of two long, thin components - a vertical strip and a horizontal strip, with width and height dx respectively - as well as a small square in the corner with side length dx. You can see how these correspond to the elements of our equation.

Now consider another, related function of x of the form $\frac{\mathrm{d}f}{\mathrm{d}x}(x)$. We'll call it a finite difference quotient. It is like the difference function considered above, except this difference is scaled by dx. It denotes the slope of a line that passes through $f(x)$ and $f(x + dx)$. For $f(x) = x^2$, it equals $2x + dx$.

The magic of calculus happens when we consider the consequences of letting dx get infinitely close to zero. When this happens, $\frac{\mathrm{d}f}{\mathrm{d}x} = 2x$ becomes a better and better approximation of the slope of a line tangent to $f(x)$ at the point x. Since we can bring dx arbitrarily close to zero, our approximation can become arbitrarily accurate. That is to say, exact.

The power rule comes about because for functions of higher powers of x, $f(x + dx) - f(x)$ will equal $x^n + nx^{n-1}dx + \ldots + dx^n - x^n$ for finite dx, where the intermediate terms all contain dx^2 or some higher order dx term. The finite difference quotient will then leave nx^{n-1} as the only dx free term, making it the only one to survive when we let dx approach zero.

To think deeply about the power rule is to respect that approximations that can be brought infinitely close to complete accuracy reveal the exact form of relationships, like that of functions to their "rates of change." It is to recognize that in its widely used form, while the notation $\frac{\mathrm{d}f}{\mathrm{d}x}$ no longer represents a fraction, it does hint at one path of logical thought involving fractions, and eventually limits, that can be used to recover its meaning.

2.4 Finite Differences and Difference Quotients

$$\Delta f(x) = f\left(x + \Delta x\right) - f(x)$$

$$\frac{\Delta f(x)}{\Delta x} = \frac{f\left(x + \Delta x\right) - f(x)}{\Delta x}$$

I already have invoked finite differences and different quotients in tracing a path of reasoning motivating the Power Rule in derivative calculus. However, finite differences are useful not just as an intermediate logical step, but as computational results with real, practical value. In practice, analytical expressions for derivatives are rarely used in computing, and when dealing with real-world data, it will be necessarily discontinuous.

So noticing that expressions in mathematics can live these dual lives, existing both in the service of precisely defined, abstract ideas, and as practically useful tools in their own rights is one result of pondering finite differences in and of themselves. Another is considering that there are different kinds of them: forward, backward, and central to name the most conceptually distinct. What we

considered so far are forward differences. Here are the other two:

$$\nabla f(x) = f(x) - f\left(x - \Delta x\right)$$

$$\delta f(x) = f\left(x + \frac{1}{2}\Delta x\right) - f\left(x - \frac{1}{2}\Delta x\right)$$

Can we have others? Of course! Why not go one quarter of a Δx forward and three quarters backward? What's interesting is that in the infinitesimal limit, all of these objects will converge to the derivative. To me, this reinforces the point I made above about the nature of approximations which can become arbitrarily accurate, but it suggests something else under the lens of computational practicality.

It is not always the case that different forms of finite differences will behave identically on all collections of pairs of purportedly related points - that is to say, on all data. In fact, the error of the central finite difference approaches zero faster than either the left or the right, but it may not be applicable to, say, a signal arriving in real-time (since we don't know the future).

Some mathematics is intended to, and may also actually do stuff in the "real" world. Calling it impure won't make this fact go away. For people like data scientists, you would think this would be reinforced day-to-day, but I find that reminding oneself that some ideas exist as more than abstract entities is still fruitful in small, deliberate doses.

3 Vectors and Matrices

$$\vec{a}, \vec{b} \in \mathbb{R}^n$$

3.1 Vector Arithmetic

$$\vec{a} + \vec{b} = (a_1 + b_1, a_2 + b_2, \ldots, a_n + b_n)$$

$$\lambda\vec{a} = (\lambda a_1, \lambda a_2, \ldots, \lambda a_n)$$

3.2 Vector Magnitude

$$|\vec{a}| = \sum_{i=1}^{n} \sqrt{a_i^2} = \sqrt{a_1 + a_2 + \ldots + a_n}$$

3.3 L1 and L2 Norms

$$||\vec{a}||_1 = \sum_{i=1}^{n} |a_i|$$

$$||\vec{a}||_1 = \sum_{i=1}^{n} \sqrt{a_i^2}$$

3.4 Matrix Multiplication

$$\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{B} \in \mathbb{R}^{n \times p}$$

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

$$\mathbf{AB} \in \mathbb{R}^{m \times p}$$

$$\mathbf{C} = \mathbf{AB} = \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1p} \\ c_{21} & c_{22} & \dots & c_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \dots & c_{mp} \end{pmatrix}$$

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \ldots + a_{im}b_{mj} = \sum_{k=1}^{m} a_{ik}b_{kj}$$

3.5 Matrix Inversion

$$\mathbf{A} \in \mathbb{R}^{n \times n}$$

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_n$$

$$\mathbf{A}^{-1} \in \mathbb{R}^{n \times n}$$

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

4 Statistical Relationships and Procedures

4.1 Correlation

$$\rho_{\mathbf{X},\mathbf{Y}} = \frac{\mathbf{cov}(\mathbf{X}, \mathbf{Y})}{\sigma_{\mathbf{X}}\sigma_{\mathbf{Y}}}$$

$$= \frac{E[(\mathbf{X} - \mu_{\mathbf{X}})(\mathbf{Y} - \mu_{\mathbf{Y}})]}{\sigma_{\mathbf{X}}\sigma_{\mathbf{Y}}}$$

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

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

4.2 Principal Components Analysis

$$\mathbf{w_{(1)}} = \arg\max_{||\mathbf{w}||=1} \left\{ ||\mathbf{X}\mathbf{w}||^2 \right\} = \arg\max_{||\mathbf{w}||=1} \left\{ ||\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}||^2 \right\}$$

$$\mathbf{w_{(1)}} = \arg\max \left\{ \frac{||\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}||^2}{\mathbf{w}^T\mathbf{w}} \right\}$$

$$\mathbf{W}_k = \arg\max_{||\mathbf{w}||=1} \left\{ ||\hat{\mathbf{X}}_k\mathbf{w}||^2 \right\}$$

$$\hat{\mathbf{X}}_k = \mathbf{X} - \sum_{s=1}^{k-1} \mathbf{X}\mathbf{w}_{(s)}\mathbf{w}_{(s)}^T$$

4.2.1 Singular-Value Decomposition (SVD)

$$\mathbf{M} \in \mathbb{R}^{m \times n}$$

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$$

- The nonzero singular values are the square roots of the nonzero eigenvalues of $\mathbf{M}^*\mathbf{M}$ or $\mathbf{M}\mathbf{M}^*$
- $\mathbf{\Sigma}$ is a diagonal $(m \times n)$ matrix of non-negative real numbers, known as the **singular values** of $\mathbf{M}$
- The columns of $\mathbf{V}$ (right singular vectors) are eigenvectors of $\mathbf{M}^*\mathbf{M}$
- The columns of $\mathbf{U}$ (left singular vectors) are eigenvectors of $\mathbf{M}\mathbf{M}^*$