Understanding the Basics of Gradients

Rishabh Iyer

March 19, 2024

Introduction

The journey into machine learning is as much about understanding the algorithms as it is about grasping the fundamental mathematics that powers these algorithms. Central to this mathematical foundation are derivatives, gradients, and the chain rule. These concepts not only illuminate the path to optimization but also unlock the deeper mechanics of learning from data. This article aims to demystify these concepts, offering both beginners and enthusiasts a clearer view of the mathematical landscape of machine learning.

1 The Power of Derivatives

What is a Derivative? A derivative captures the idea of how a function's output changes as its input changes, offering a precise measurement of this change at any given point. It's akin to understanding the slope of the ground under your feet: knowing whether you're ascending, descending, or on level ground.
Example: Consider the function $f(x) = x^2$. Its derivative, $f'(x) = 2x$, tells us that for every point x, the slope of the tangent to the curve at that point is $2x$.

In machine learning, derivatives help us minimize loss functions by indicating the direction to adjust our model's parameters. This is akin to finding the lowest point in a valley by following the steepest path downward.

2 Navigating with Gradients

What are Gradients: The gradient extends the concept of a derivative to functions with multiple inputs, offering a vector that points in the direction of the steepest ascent of the function.
Consider a function $f : \mathbb{R}^n \to \mathbb{R}$. For a function $f(x_1, \cdots, x_n)$, the gradient is defined as:

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n} \right)$$

This vector ∇f represents the gradient of the function f, where each component $\frac{\partial f}{\partial x_i}$ is the partial derivative of f with respect to the variable x_i. The partial derivative $\frac{\partial f}{\partial x_i}$ measures how f changes as x_i changes, holding all other variables constant.
Examples: For a 3-dimensional function $f(x, y, z) = x^2 + 2y^2 + 3z^2$, the gradient is $\nabla f = [2x, 4y, 6z]$, indicating how f changes in the x, y, and z directions. In a 4-dimensional space, for $f(w, x, y, z) = w^2 + x^2 + y^2 + z^2$, the gradient is $\nabla f = [2w, 2x, 2y, 2z]$, providing a roadmap for navigating the function's landscape in four dimensions.

Understanding the Gradient as a Dot Product: The gradient's relationship with the dot product reveals its true power: the dot product $\nabla f \cdot \vec{v}$ computes the rate of change of f in the direction of a vector $\vec{v}$, highlighting the gradient's role in pointing out the direction of steepest ascent.

Gradients are the backbone of the gradient descent algorithm, guiding the iterative adjustment of parameters to minimize the loss function, akin to descending a mountain in the path of steepest descent.

3 Sub-Gradients

A subgradient at a point on a non-differentiable function is a vector that generalizes the concept of a gradient. Unlike differentiable functions that have a unique gradient at each point, non-differentiable functions can have multiple subgradients at points of non-differentiability.

3.1 Computing Subgradients for Specific Functions

Let's consider the computation of subgradients for two commonly encountered non-differentiable functions:

Absolute Value Function $f(x) = |x|$

- For $x > 0$, $f(x) = x$, and the subgradient is 1.
- For $x < 0$, $f(x) = -x$, and the subgradient is -1.
- At $x = 0$, the function is not differentiable, but the subgradients can be any value in the interval $[-1, 1]$.

ReLU Function $f(x) = \max(x, 0)$

- For $x > 0$, $f(x) = x$, and the subgradient is 1.
- For $x < 0$, $f(x) = 0$, and the subgradient is 0.
- At $x = 0$, the function is not differentiable, but the subgradients can be any value in the interval $[0, 1]$.

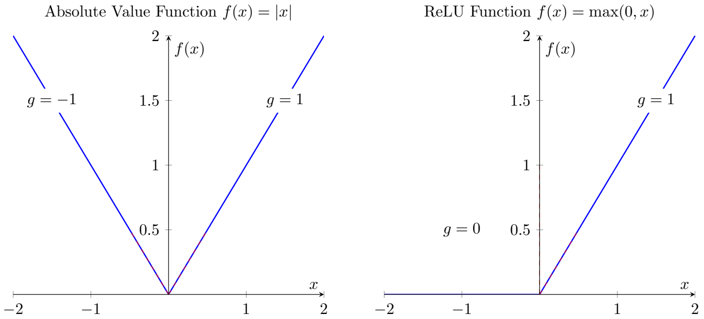

Figure 1: Illustration of subgradients for the absolute value function and the ReLU function. For the absolute value function, the subgradient is -1 for $x < 0$, varies between $[-1, 1]$ at $x = 0$, and is 1 for $x > 0$. For the ReLU function, the subgradient is 0 for $x < 0$, varies between $[0, 1]$ at $x = 0$, and is 1 for $x > 0$.

4 The Chain Rule: Connecting Functions

Basics of the Chain Rule: The chain rule allows us to compute the derivative of composite functions, revealing how changes in one function affect the output of another.
Mathematical Insight: For composite functions $f(g(x))$, where $f(x) = x^2$ and $g(x) = x + 1$, the chain rule gives us the derivative as $2(x + 1)$, multiplying the derivatives of f and g.
The Chain Rule Explained: The chain rule is a fundamental derivative rule that allows us to calculate the derivative of composite functions. In simple terms, if you have two functions $f(x)$ and $g(x)$, and you want to find the derivative of their composition $f(g(x))$, the chain rule states:

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

This means you first take the derivative of the outer function f with respect to its input, which is $g(x)$, and multiply it by the derivative of the inner function g with respect to x.

The chain rule is crucial for training neural networks through a process called backpropagation. It allows us to compute how the loss function changes with respect to each weight in the network, even in complex architectures, enabling precise adjustments to minimize the loss.

5 Calculating Gradients: Practical Examples

5.1 Gradient of the Squared Norm of a Vector

Consider the squared norm of a vector $\mathbf{w}$ defined as $||\mathbf{w}||^2 = \sum_{i=1}^{n} w_i^2$. To understand why the gradient of $||\mathbf{w}||^2$ with respect to $\mathbf{w}$ is $2\mathbf{w}$, let's look at the partial derivative in each dimension.

The squared norm of $\mathbf{w}$ is a sum of the squares of its components:

$$||\mathbf{w}||^2 = w_1^2 + w_2^2 + \ldots + w_n^2$$

To find the gradient of $||\mathbf{w}||^2$ with respect to $\mathbf{w}$, we compute the partial derivative of $||\mathbf{w}||^2$ with respect to each component w_i of $\mathbf{w}$. The gradient is a vector of these partial derivatives:

$$\nabla||\mathbf{w}||^2 = \left[\frac{\partial||\mathbf{w}||^2}{\partial w_1}, \frac{\partial||\mathbf{w}||^2}{\partial w_2}, \ldots, \frac{\partial||\mathbf{w}||^2}{\partial w_n}\right]$$

Computing the Partial Derivatives: The partial derivative of $||\mathbf{w}||^2$ with respect to w_i is:

$$\frac{\partial||\mathbf{w}||^2}{\partial w_i} = \frac{\partial}{\partial w_i}(w_1^2 + w_2^2 + \ldots + w_i^2 + \ldots + w_n^2)$$

Since w_i^2 is the only term in the sum that depends on w_i, the derivative of all other terms with respect to w_i is zero. Thus, we have:

$$\frac{\partial||\mathbf{w}||^2}{\partial w_i} = 2w_i$$

The Gradient Vector: Combining these results, the gradient of $||\mathbf{w}||^2$ is:

$$\nabla||\mathbf{w}||^2 = [2w_1, 2w_2, \ldots, 2w_n] = 2\mathbf{w}$$

This gradient $2\mathbf{w}$ indicates the direction and rate at which the squared norm $||\mathbf{w}||^2$ increases the fastest. For each dimension i, the rate of increase in $||\mathbf{w}||^2$ with respect to a small change in w_i is proportional to $2w_i$, explaining why the gradient of the squared norm is $2\mathbf{w}$.

5.2 Gradient of the Dot Product

The dot product between two vectors $\mathbf{w}$ and $\mathbf{x}$, where both vectors are of dimension n, is defined as $\mathbf{w} \cdot \mathbf{x} = \sum_{i=1}^{n} w_i x_i$. To find the gradient of the dot product with respect to the vector $\mathbf{w}$, we look at how this dot product changes as each component of $\mathbf{w}$ changes.

The dot product can be expressed as:

$$\mathbf{w} \cdot \mathbf{x} = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n$$

The gradient of the dot product with respect to $\mathbf{w}$ involves computing the partial derivative of $\mathbf{w} \cdot \mathbf{x}$ with respect to each component of $\mathbf{w}$:

$$\nabla_{\mathbf{w}}(\mathbf{w} \cdot \mathbf{x}) = \left[\frac{\partial(\mathbf{w} \cdot \mathbf{x})}{\partial w_1}, \frac{\partial(\mathbf{w} \cdot \mathbf{x})}{\partial w_2}, \ldots, \frac{\partial(\mathbf{w} \cdot \mathbf{x})}{\partial w_n}\right]$$

Computing the Partial Derivatives: To compute each partial derivative, consider how the dot product changes with a small change in w_i, while keeping all other components of $\mathbf{w}$ constant. For any i, the only term in the sum $\mathbf{w} \cdot \mathbf{x}$ that depends on w_i is $w_i x_i$. Therefore, the partial derivative of $\mathbf{w}\cdot\mathbf{x}$ with respect to w_i is simply x_i:

$$\frac{\partial(\mathbf{w} \cdot \mathbf{x})}{\partial w_i} = x_i$$

The Gradient Vector: Thus, the gradient of the dot product $\mathbf{w} \cdot \mathbf{x}$ with respect to $\mathbf{w}$ is the vector $\mathbf{x}$ itself:

$$\nabla_{\mathbf{w}}(\mathbf{w} \cdot \mathbf{x}) = [x_1, x_2, \ldots, x_n] = \mathbf{x}$$

This result tells us that the gradient of the dot product points in the direction of $\mathbf{x}$, indicating how the dot product $\mathbf{w} \cdot \mathbf{x}$ increases as $\mathbf{w}$ moves in the direction of $\mathbf{x}$. The magnitude of the change in the dot product for a small change in $\mathbf{w}$ is directly proportional to the components of $\mathbf{x}$, reflecting the linear relationship between $\mathbf{w}$ and $\mathbf{w} \cdot \mathbf{x}$ in each dimension.

5.3 Gradient of the Linear Regression Loss Function

TODO: Reference the article that goes over Linear Regression in detail.

The linear regression model predicts an output $\hat{y}_i$ for each input vector $\mathbf{x}_i$ using the linear combination $\hat{y}_i = \mathbf{w} \cdot \mathbf{x}_i + b$, where $\mathbf{w}$ represents the weight vector, b the bias, and $\cdot$ denotes the dot product. The mean squared error (MSE) loss function for N observations is given by:

$$L(\mathbf{w}, b) = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{w} \cdot \mathbf{x}_i + b - y_i)^2$$

To find the gradient of $L(\mathbf{w}, b)$ with respect to the weights $\mathbf{w}$ and bias b, we apply the chain rule, which allows us to decompose the derivative of a composite function into the product of derivatives.

Note that:

$$L(\mathbf{w}, b) = \sum_{i=1}^{N} f(g_i(\mathbf{w}, b))$$

where $f(x) = x^2$ and $g_i(\mathbf{w}, b) = \mathbf{w} \cdot \mathbf{x}_i + b - y_i$.

Gradient with Respect to $\mathbf{w}$: The partial derivative of L with respect to each weight w_j in $\mathbf{w}$ is computed as follows using the chain rule (see Section above).

$$\frac{\partial L}{\partial w_j} = \frac{1}{N} \sum_{i=1}^{N} 2(\mathbf{w} \cdot \mathbf{x}_i + b - y_i) \cdot \frac{\partial}{\partial w_j}(\mathbf{w} \cdot \mathbf{x}_i + b)$$

Since $\frac{\partial}{\partial w_j}(\mathbf{w}\cdot\mathbf{x}_i+b)=x_{ij}$ (where x_{ij} is the j-th component of $\mathbf{x}_i$), we have:

$$\frac{\partial L}{\partial w_j}=\frac{2}{N}\sum_{i=1}^{N}(\mathbf{w}\cdot\mathbf{x}_i+b-y_i)\cdot x_{ij}$$

Thus, the gradient of L with respect to $\mathbf{w}$ is:

$$\nabla_{\mathbf{w}}L=\frac{2}{N}\sum_{i=1}^{N}(\mathbf{w}\cdot\mathbf{x}_i+b-y_i)\cdot\mathbf{x}_i$$

Gradient with Respect to b: Similarly, the derivative of L with respect to b is:

$$\frac{\partial L}{\partial b}=\frac{1}{N}\sum_{i=1}^{N}2(\mathbf{w}\cdot\mathbf{x}_i+b-y_i)\cdot\frac{\partial}{\partial b}(\mathbf{w}\cdot\mathbf{x}_i+b)$$

Since $\frac{\partial}{\partial b}(\mathbf{w}\cdot\mathbf{x}_i+b)=1$, we get:

$$\frac{\partial L}{\partial b}=\frac{2}{N}\sum_{i=1}^{N}(\mathbf{w}\cdot\mathbf{x}_i+b-y_i)$$

The gradient of the MSE loss function with respect to $\mathbf{w}$ and b points in the direction of steepest increase of the loss. To minimize the loss, one must move in the opposite direction of the gradient, adjusting $\mathbf{w}$ and b accordingly. This is the essence of gradient descent optimization in linear regression.

6 Conclusion

Derivatives, gradients, and the chain rule form the mathematical bedrock upon which machine learning algorithms stand. By understanding these concepts, you gain insights into how algorithms learn and adapt, making informed decisions based on the landscape of data they navigate. This exploration into the mathematics of machine learning not only demystifies the field but also empowers you to engage with it more deeply, laying the groundwork for further discovery and innovation.