Math Basics for Machine Learning

1 Why Do We Need Math for Machine Learning?

Machine learning is largely about finding patterns in numbers.

For example, suppose we want to predict the price of a house. We might have:

$$\text{size} = 2000 \text{ ft}^2, \qquad \text{bedrooms} = 3, \qquad \text{age} = 10 \text{ years}.$$

A machine learning model takes these numbers as input and produces another number:

$$\text{house features} \longrightarrow \text{model} \longrightarrow \text{predicted price}.$$

The mathematics gives us a language for describing what the model is doing.

The main mathematical ideas we need today are:

1. basic algebra,
2. vectors,
3. matrices,
4. tensors,
5. matrix operations,
6. derivatives,
7. the chain rule,
8. gradients.

2 Basic Math and Algebra

Before looking at vectors and matrices, we need some basic algebra.

2.1 Variables

A variable is simply a name for a number.

For example:

$$x = 5.$$

We can perform ordinary arithmetic with variables:

$$x + 3 = 8$$

and

$$2x = 10.$$

In machine learning, variables often represent things such as:

$$x = \text{number of bedrooms}$$

or

$$x = \text{house size}.$$

2.2 Functions

A function takes an input and produces an output.

For example:

$$f(x) = 2x + 1.$$

If $x = 3$:

$$f(3) = 2(3) + 1 = 7.$$

We can think of a function as a machine:

$$x \longrightarrow \boxed{f(x)} \longrightarrow f(x).$$

Machine learning models are essentially complicated functions.

For example, a very simple prediction model might be:

$$\text{price} = 100000 + 200(\text{square feet}).$$

The numbers in the model are called **parameters**.

2.3 Equations

An equation says that two things are equal.

For example:

$$2x + 3 = 11.$$

Subtract 3:

$$2x = 8.$$

Divide by 2:

$$x = 4.$$

This basic idea becomes important in machine learning because training a model often involves finding parameter values that make the model's predictions as accurate as possible.

3 Vectors

A **vector** is an ordered list of numbers.

For example:

$$\boldsymbol{x} = \begin{bmatrix} 3 \\ 5 \\ 2 \end{bmatrix}.$$

This vector contains three numbers.

We could use it to describe a house:

$$\boldsymbol{x} = \begin{bmatrix} \text{bedrooms} \\ \text{bathrooms} \\ \text{age} \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 10 \end{bmatrix}.$$

The important thing about a vector is that the **order matters**.

These are different vectors:

$$\begin{bmatrix} 3 \\ 2 \end{bmatrix} \neq \begin{bmatrix} 2 \\ 3 \end{bmatrix}.$$

3.1 Vector Addition

We can add two vectors by adding corresponding entries.

For example:

$$\boldsymbol{a} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \qquad \boldsymbol{b} = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}.$$

Then:

$$\boldsymbol{a} + \boldsymbol{b} = \begin{bmatrix} 1+4 \\ 2+5 \\ 3+6 \end{bmatrix} = \begin{bmatrix} 5 \\ 7 \\ 9 \end{bmatrix}.$$

We can only add vectors of the same size.

3.2 Multiplying a Vector by a Number

We can multiply every element of a vector by a number.

For example:

$$3 \begin{bmatrix} 2 \\ 4 \\ 1 \end{bmatrix} = \begin{bmatrix} 6 \\ 12 \\ 3 \end{bmatrix}.$$

The number 3 is called a **scalar**.

3.3 Dot Product

One of the most important operations in machine learning is the **dot product**.

Suppose:

$$\boldsymbol{x} = \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}, \qquad \boldsymbol{w} = \begin{bmatrix} 10 \\ 20 \\ 5 \end{bmatrix}.$$

The dot product is:

$$\boldsymbol{w} \cdot \boldsymbol{x} = (10)(2) + (20)(3) + (5)(4).$$

Therefore:

$$\boldsymbol{w} \cdot \boldsymbol{x} = 20 + 60 + 20 = 100.$$

Notice that the result is a **single number**, not a vector.

The dot product is extremely important because a simple neural-network neuron essentially performs:

$$\text{output} = \boldsymbol{w} \cdot \boldsymbol{x} + b.$$

Here:

- $\boldsymbol{x}$ = input values,
- $\boldsymbol{w}$ = weights,
- b = bias,
- output = result.

4 Matrices

A **matrix** is a rectangular arrangement of numbers.

For example:

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}.$$

This matrix has:

$$2 \text{ rows} \times 3 \text{ columns}.$$

We call it a 2×3 matrix.

4.1 Matrix Addition

Just like vectors, matrices can be added element by element.

For example:

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \qquad B = \begin{bmatrix} 10 & 20 \\ 30 & 40 \end{bmatrix}.$$

Then:

$$A + B = \begin{bmatrix} 11 & 22 \\ 33 & 44 \end{bmatrix}.$$

Again, the matrices must have compatible dimensions.

4.2 Matrix–Scalar Multiplication

We can multiply every element by a number:

$$3 \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ 9 & 12 \end{bmatrix}.$$

5 Matrix Multiplication

Matrix multiplication is more interesting.

Consider:

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

and

$$B = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}.$$

Then:

$$AB = \begin{bmatrix} 1(5) + 2(7) & 1(6) + 2(8) \\ 3(5) + 4(7) & 3(6) + 4(8) \end{bmatrix}.$$

Therefore:

$$AB = \begin{bmatrix} 19 & 22 \\ 43 & 50 \end{bmatrix}.$$

5.1 The Important Dimension Rule

Suppose:

$$A \text{ is } m \times n$$

and

$$B \text{ is } n \times p.$$

Then:

$$AB$$

is an

$$m \times p$$

matrix.

The "inside" dimensions must match:

$$\boxed{(m \times n)(n \times p) = (m \times p)}.$$

For example:

$$(3 \times 2)(2 \times 4)$$

is valid and produces a:

$$3 \times 4$$

matrix.

But:

$$(3 \times 2)(3 \times 4)$$

is not valid because the inside dimensions, 2 and 3, do not match.

6 Matrices and Machine Learning

Suppose we have three houses and two features:

$$X = \begin{bmatrix} 2000 & 3 \\ 1500 & 2 \\ 2500 & 4 \end{bmatrix}.$$

Each row represents one house.

Each column represents a feature:

	size	bedrooms
house 1	2000	3
house 2	1500	2
house 3	2500	4

Suppose our model has weights:

$$\boldsymbol{w} = \begin{bmatrix} 100 \\ 50000 \end{bmatrix}.$$

Then:

$$X\boldsymbol{w}$$

produces predictions for all three houses at once.

This is one reason matrix operations are so important in machine learning.

Instead of calculating every prediction separately, a computer can perform one large matrix operation.

7 Tensors

A tensor is a generalization of scalars, vectors, and matrices.

A useful way to think about them is:

Object	Number of dimensions
Scalar	0
Vector	1
Matrix	2
Tensor	3 or more

For example, a color image can be represented as a tensor.

A color image might have:

$$\text{height} \times \text{width} \times \text{color channels}.$$

For a 100×200 RGB image:

$$100 \times 200 \times 3.$$

The three channels are:

$$R = \text{red}, \qquad G = \text{green}, \qquad B = \text{blue}.$$

A collection of images adds another dimension:

$$\text{number of images} \times \text{height} \times \text{width} \times \text{channels}.$$

For example:

$$32 \times 100 \times 200 \times 3.$$

This could represent a batch of 32 images.

8 Calculus: Why Do We Need It?

Calculus allows us to understand **change**.

This is exactly what we need when training a machine learning model.

Suppose a model makes a prediction:

$$\hat{y} = f(x).$$

We want to know:

> How much does the output change when we change the input?

The derivative answers this question.

9 Derivatives

Consider:

$$f(x) = x^2.$$

Some values are:

x	1	2	3	4	5
$f(x)$	1	4	9	16	25

As x increases, $f(x)$ increases faster and faster.

The derivative tells us the **rate of change**.

For:

$$f(x) = x^2,$$

the derivative is:

$$\boxed{f'(x) = 2x}.$$

At $x = 3$:

$$f'(3) = 6.$$

This means that near $x = 3$, increasing x by a small amount causes $f(x)$ to increase by approximately six times that amount.

9.1 A Simple Derivative Rule

One of the most useful rules is:

$$\boxed{\frac{d}{dx}x^n = nx^{n-1}}$$

For example:

$$\frac{d}{dx}x^3 = 3x^2.$$

And:

$$\frac{d}{dx}x^2 = 2x.$$

For a linear function:

$$f(x) = 3x + 5,$$

we have:

$$f'(x) = 3.$$

The slope is constant.

10 The Slope Interpretation

For a straight line:

$$y = mx + b,$$

the number m is the slope.

For example:

$$y = 2x + 1.$$

The slope is:

$$2.$$

The derivative gives us the same idea for curves.

For:

$$y = x^2,$$

the slope changes depending on where we are:

$$y' = 2x.$$

At $x = 1$:

$$y' = 2.$$

At $x = 3$:

$$y' = 6.$$

At $x = 5$:

$$y' = 10.$$

Thus the derivative tells us how steep the curve is at a particular point.

11 The Chain Rule

The chain rule is one of the most important ideas in machine learning.

It tells us how to differentiate a function that is built from other functions.

Suppose:

$$y = (3x + 1)^2.$$

There are two steps:

$$x \longrightarrow 3x + 1 \longrightarrow (3x + 1)^2.$$

Let:

$$u = 3x + 1.$$

Then:

$$y = u^2.$$

We can differentiate each step:

$$\frac{dy}{du} = 2u$$

and

$$\frac{du}{dx} = 3.$$

The chain rule says:

$$\boxed{\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx}}$$

Therefore:

$$\frac{dy}{dx} = 2u(3).$$

Since:

$$u = 3x + 1,$$

we get:

$$\boxed{\frac{dy}{dx} = 6(3x+1)}.$$

The important idea is:

$$\boxed{\text{change in final output} = \text{change at each step multiplied together}}$$

This is exactly the idea used by **backpropagation** in neural networks.

12 A Machine Learning Example of the Chain Rule

Suppose a very simple model is:

$$x \longrightarrow z = wx \longrightarrow y = z^2.$$

Thus:

$$y = (wx)^2.$$

We want to know:

$$\frac{dy}{dw}.$$

There are two steps:

$$w \longrightarrow z = wx \longrightarrow y = z^2.$$

Using the chain rule:

$$\frac{dy}{dw} = \frac{dy}{dz}\frac{dz}{dw}.$$

We know:

$$\frac{dy}{dz} = 2z$$

and:

$$\frac{dz}{dw} = x.$$

Therefore:

$$\frac{dy}{dw} = 2zx.$$

Since:

$$z = wx,$$

we obtain:

$$\boxed{\frac{dy}{dw} = 2wx^2}.$$

The chain rule allows us to determine how changing a model parameter w changes the final output.

13 Gradients

A derivative is useful when there is one variable.

But machine learning models usually have many parameters.

Suppose:

$$f(x, y) = x^2 + y^2.$$

There are two variables:

$$x, \qquad y.$$

We can calculate a derivative with respect to each variable.

With respect to x:

$$\frac{\partial f}{\partial x} = 2x.$$

With respect to y:

$$\frac{\partial f}{\partial y} = 2y.$$

We combine these derivatives into a vector called the **gradient**:

$$\boxed{\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}}$$

Therefore:

$$\boxed{\nabla f = \begin{bmatrix} 2x \\ 2y \end{bmatrix}}$$

13.1 Example

Suppose:

$$x = 3, \qquad y = 4.$$

Then:

$$\nabla f = \begin{bmatrix} 2(3) \\ 2(4) \end{bmatrix} = \begin{bmatrix} 6 \\ 8 \end{bmatrix}.$$

The gradient tells us how the function changes when we move in different directions.

14 Gradient Descent

Gradient descent is one of the fundamental ideas behind machine learning.

Suppose we have a function representing our model's error:

$$L(w).$$

We want to find a value of w that makes the error as small as possible.

Imagine standing on a hill.

The gradient tells us which direction is "uphill."

Therefore, to go downhill, we move in the opposite direction.

This gives us the basic gradient-descent equation:

$$\boxed{w_{\text{new}} = w_{\text{old}} - \eta \frac{dL}{dw}}$$

where η is called the **learning rate**.

The learning rate controls how big a step we take.

14.1 Simple Example

Suppose:

$$L(w) = w^2.$$

Then:

$$\frac{dL}{dw} = 2w.$$

Suppose:

$$w = 5$$

and:

$$\eta = 0.1.$$

Then:

$$w_{\text{new}} = 5 - (0.1)(2)(5).$$

Therefore:

$$w_{\text{new}} = 5 - 1 = 4.$$

We moved from:

$$5 \longrightarrow 4.$$

The error decreased:

$$L(5) = 25$$

while:

$$L(4) = 16.$$

We are moving toward the minimum.

15 Gradient Descent with Multiple Parameters

A real machine learning model may have thousands or millions of parameters.

Instead of:

$$w,$$

we have a vector:

$$\boldsymbol{w} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \end{bmatrix}.$$

The derivative becomes a gradient:

$$\nabla L(\boldsymbol{w}).$$

The update becomes:

$$\boxed{\boldsymbol{w}_{\text{new}} = \boldsymbol{w}_{\text{old}} - \eta \nabla L(\boldsymbol{w})}$$

This simple equation captures a central idea of machine learning:

Measure the error → calculate its gradient → change the parameters to reduce the error.

16 Putting Everything Together

Many machine learning models can be understood as a sequence of mathematical operations.

For example:

$$\text{input} \longrightarrow \text{vectors} \longrightarrow \text{matrix multiplication} \longrightarrow \text{function} \longrightarrow \text{prediction} \longrightarrow \text{error}.$$

During training, we work backwards:

$$\text{error} \longrightarrow \text{derivatives} \longrightarrow \text{gradients} \longrightarrow \text{parameter updates}.$$

The major mathematical concepts therefore fit together:

$$\boxed{\begin{array}{c}\text{Vectors and matrices}\\ \downarrow \\ \text{Model computations}\\ \downarrow \\ \text{Prediction}\\ \downarrow \\ \text{Loss / Error}\\ \downarrow \\ \text{Derivatives}\\ \downarrow \\ \text{Chain Rule}\\ \downarrow \\ \text{Gradient}\\ \downarrow \\ \text{Gradient Descent}\\ \downarrow \\ \text{Better Model}\end{array}}$$

17 Key Ideas to Remember

1. A **vector** is an ordered list of numbers.
2. A **matrix** is a rectangular grid of numbers.
3. A **tensor** is a generalization of vectors and matrices to additional dimensions.
4. The **dot product** combines two vectors into one number:

$$\boldsymbol{a} \cdot \boldsymbol{b} = \sum_i a_i b_i.$$

5. Matrix multiplication combines rows of one matrix with columns of another.
6. A **derivative** measures how quickly something changes.
7. The **chain rule** tells us how changes propagate through multiple mathematical operations.
8. A **gradient** is a vector containing derivatives with respect to multiple variables.
9. **Gradient descent** uses the gradient to change model parameters in a direction that reduces error.

Quick Practice

1. Calculate: $\begin{bmatrix}1\\2\\3\end{bmatrix} + \begin{bmatrix}4\\5\\6\end{bmatrix}$.

2. Calculate the dot product:
$$\begin{bmatrix}2\\3\end{bmatrix} \cdot \begin{bmatrix}5\\4\end{bmatrix}.$$

3. Calculate:
$$\begin{bmatrix}1 & 2\\3 & 4\end{bmatrix} \begin{bmatrix}2\\5\end{bmatrix}.$$

4. Find the derivative:
$$f(x) = x^3 + 2x.$$

5. Find the gradient of:
$$f(x, y) = x^2 + 3y^2.$$

6. If:
$$L(w) = w^2,$$
and $w = 4$, calculate one gradient-descent step using $\eta = 0.1$.