

Machine Learning

Mathematics in Machine Learning

Dept. SW and Communication Engineering

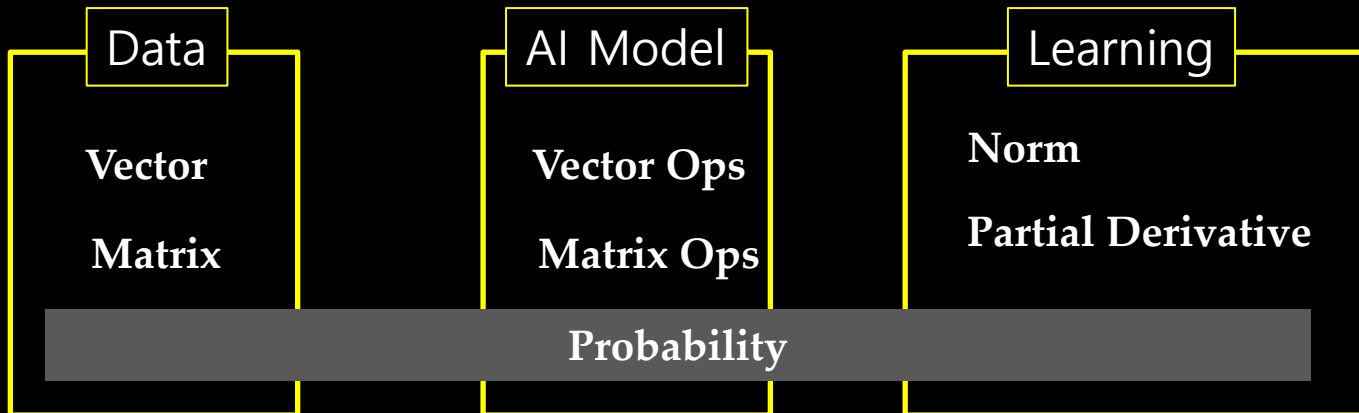
Prof. Giseop Noh (kafa46@hongik.ac.kr)

Contents

■ Study Goals

- Linear Algebra
 - Vector, Matrix, Tensor
- Derivative
- Probability

■ Mathematics in Machine Learning



Linear Algebra - Vector

Vector

■ In physics,

- A vector is a quantity that has both magnitude and direction and is represented by an arrow.
- Length of arrow: magnitude of the vector
- Direction of arrow: direction of the vector

■ Concept of Vector in ML

- Collection of values (data)
- A vector has an order

■ Representation of Vector

- Denoted in bold lowercase letters.

$$\mathbf{x} = (x_1, x_2, x_3)$$

- Can be represented as a column- or row vector with Matrix

Perspectives

Physics

Arrow

운동, 자기장, 전기장, ...



크기, 방향이 같다면
모두 같은 벡터
좌표공간에서 자유롭게
이동하는 것도 가능

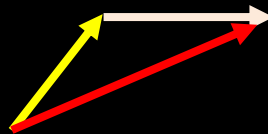
Mathematics

Arrow, List(Tuple)

In any notation,

can be usable

(add, multiplication, ...)



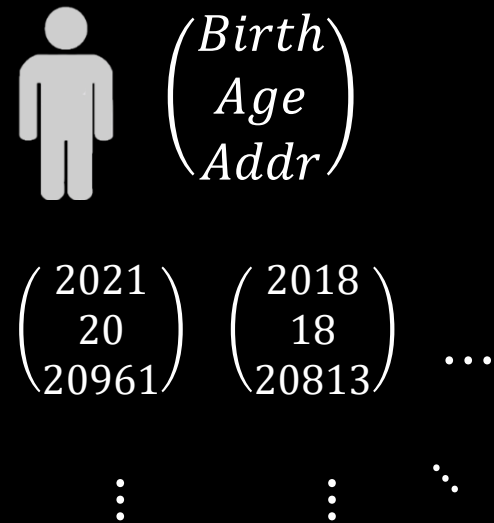
$$\begin{pmatrix} 1 \\ 2 \\ -4 \end{pmatrix} + \begin{pmatrix} 2 \\ 3 \\ 5 \end{pmatrix} = \begin{pmatrix} 3 \\ 5 \\ -1 \end{pmatrix}$$

$$2 \begin{pmatrix} 1 \\ 2 \\ -4 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \\ -8 \end{pmatrix}$$

Computer Science

List, Tuple

Features, Dataset, ...



Transpose & Vector Interpretation

■ Transpose of \mathbf{x} is \mathbf{x}^T

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \quad \mathbf{x}^T = (x_1, x_2, x_3)^T$$

■ Example of Vector Interpretation

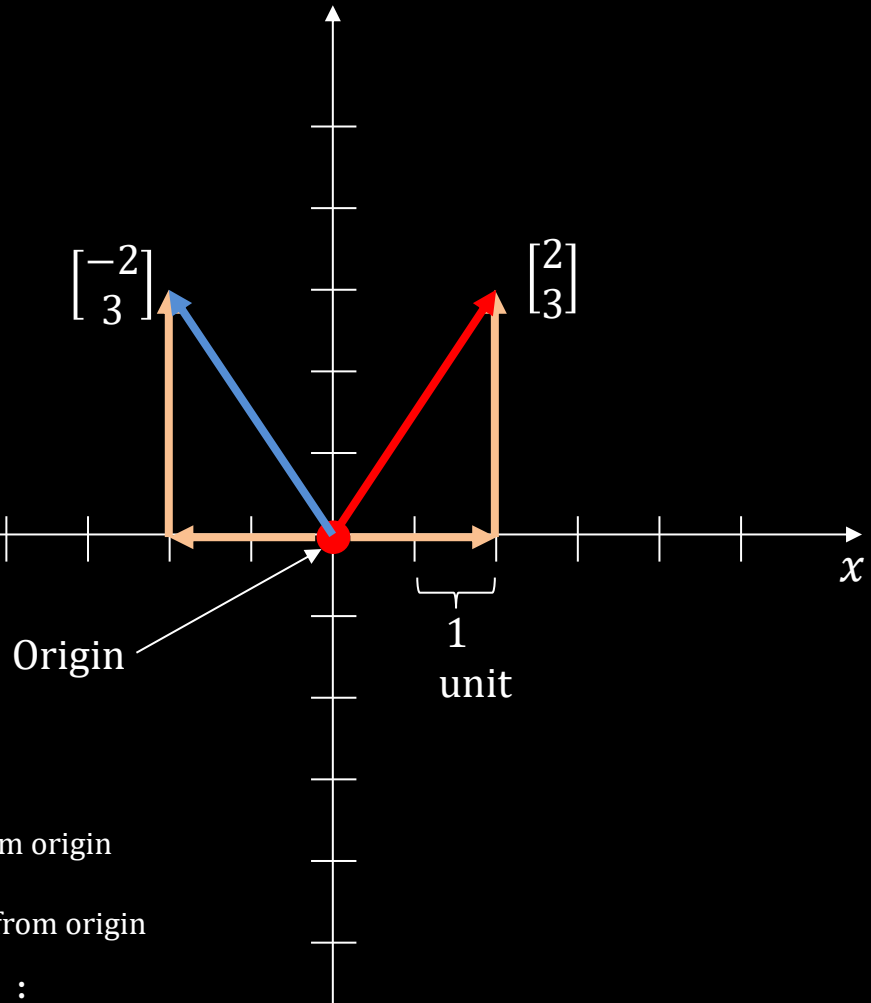
$$\mathbf{a} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

$$\mathbf{b} = \begin{bmatrix} -2 \\ 3 \end{bmatrix}$$

First axis: How far from origin

Second axis: How far from origin

⋮



n-Dimensional Space and n-Vector

■ *n*-vector

- A vector with n components

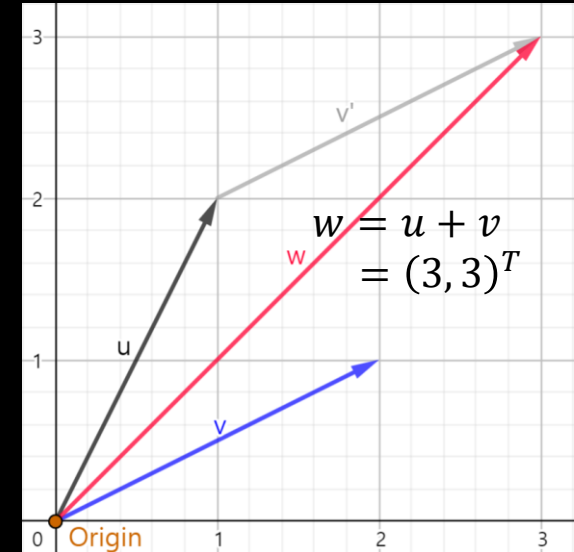
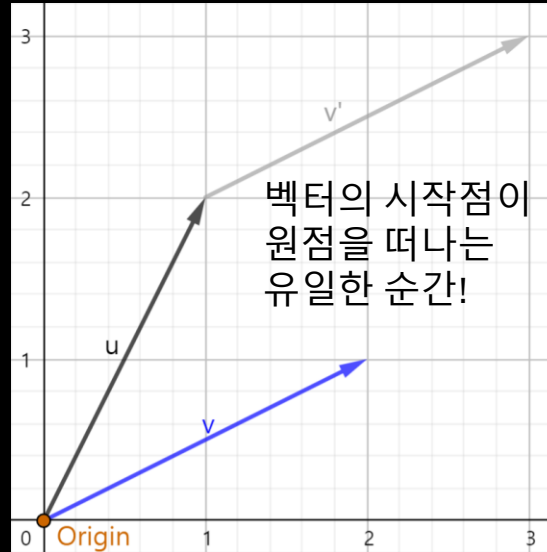
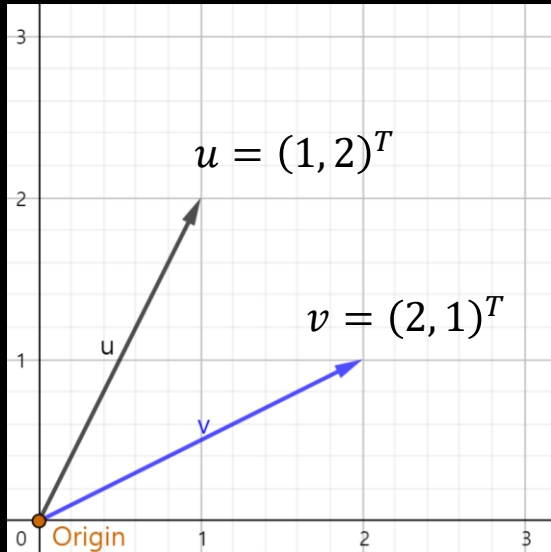
$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

■ *n*-vector exists in *n*-dimensional space

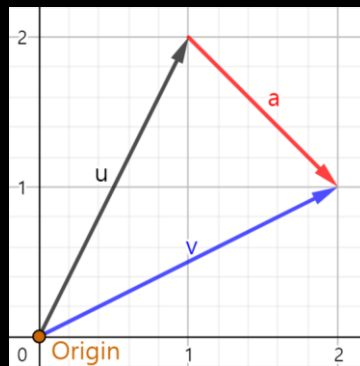
$$x \in \mathbb{R}^n$$

Vector Operations - Add

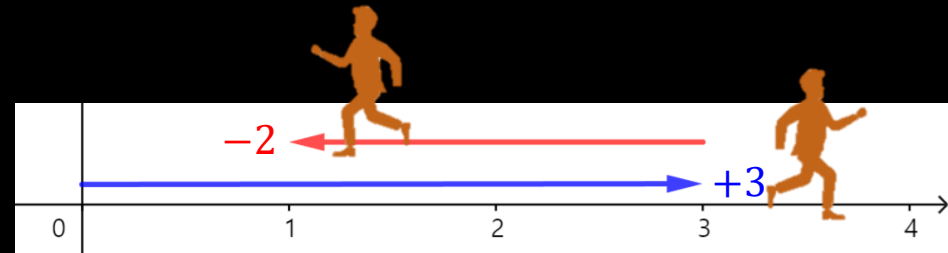
■ Add



●●● Why not this way???



$3 + (-2) = 1$ ← What this means?



Think about carefully length & direction and adding

Vector Operations - Scaling (multiplication)

Scaling

Multiplying a vector by a scalar
(Scalar Multiplication,
스칼라배, 실수배)

$$k \in \mathbb{R}, \text{ and } u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

$$ku = \underbrace{k}_{\text{value of vector scaling}} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix} = \begin{pmatrix} ku_1 \\ ku_2 \\ \vdots \\ ku_n \end{pmatrix}$$

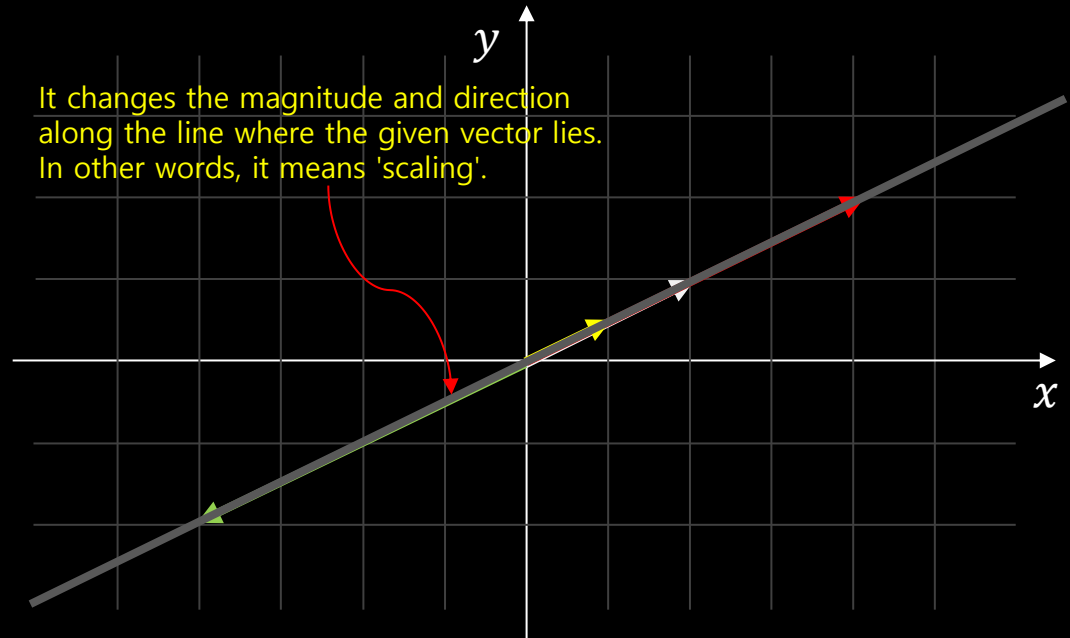
value of vector scaling

Scale + er/or

공식 표기: Scalar

"Scaler" 또는 "Scalar" 라고 부름

한국말로 "스칼라"



[Toy example]

$$u = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

$$0.5u = 0.5 \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0.5 \end{pmatrix}$$

$$2u = 2 \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \end{pmatrix}$$

$$-2u = -2 \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} -4 \\ -2 \end{pmatrix}$$

Vectors and Data Preprocessing

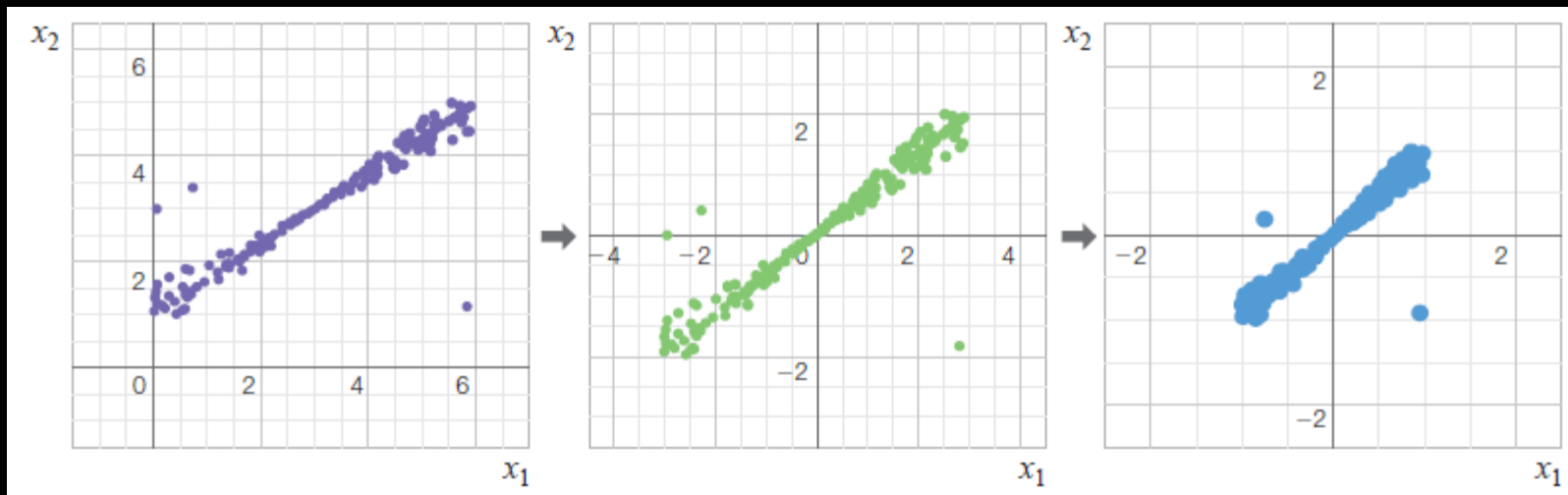
- Interpreting a vector as a point in space
- ML dataset to be considered as a single point in a dimensional space corresponding to the number of observed features
- By treating vectors as points in space,
 - data can be visualized in various ways
 - helping to intuitively understand its characteristics
 - Additionally, a data preprocessing process is applied to improve the performance of the model.

Data Preprocessing

- Improves the performance of a model using a training dataset.

- **Example: Data preprocessing through Standardization**

- Zero-Centering: Shifting the dataset's mean to 0
- STD Adjustment: Adjusting the standard deviation of the data to 1



Original dataset

Zero centering

STD adjustment

Vector Norm

■ Definition of Norm

In mathematics, a norm is a function from a real or complex vector space to the non-negative real numbers that behaves in certain ways like the distance from the origin.

(source: [online wiki](#)) 한글로 '노름' 으로 읽음 ([정보통신기술용어해설집](#))

■ Vector Norm

- Definition $\|v\| = \sqrt{v_1^2 + v_2^2 + \dots + v_n^2}$, where $v = (v_1, v_2, \dots, v_n)^T$
- Toy Example



$$v = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$$

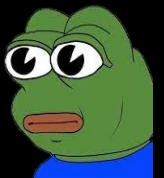
$$\|v\| = \sqrt{2^2 + 3^2} = \sqrt{13}$$

2-dim real number space \rightarrow distance from origin

However n -dimension case

Distance is ambiguous.

We use the term 'Norm'



Formal Definition of Norm

■ Norm must satisfy following conditions

- Scaling: $f(\alpha x) = |\alpha| f(x)$
- Triangel Inequality: $f(x + y) \leq f(x) + f(y)$
- Positive Function: $f(x) \geq 0$

■ General Representation of L_p Norm

$$\|X\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{\frac{1}{p}}, \text{ where } X \in \mathbb{R}^n$$

- $p = 1, 2, \infty$ are most frequently used.
- When $p = \infty$, The largest absolute value among the components of vector X

Vector Product (Scalar Product)

How to multiply a vector and a vector

Result of product: yields only scalar

The magnitude that changes when the size of one vector is applied to another vector (scalar)

Representation (notation)

$$u = (u_1, u_2, \dots, u_n)^T$$

$$v = (v_1, v_2, \dots, v_n)^T$$

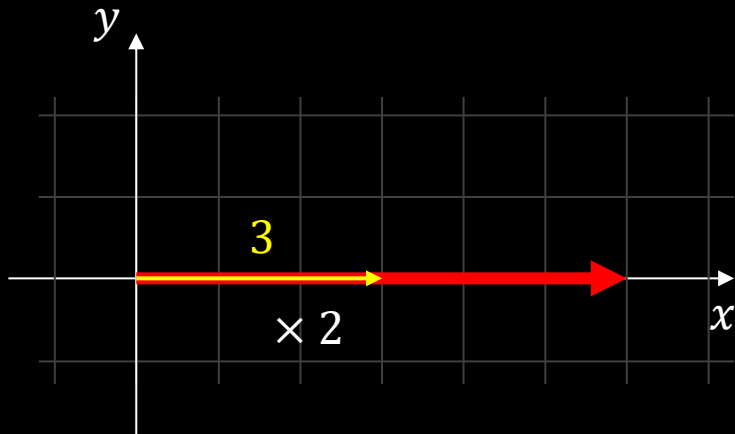
$$u \cdot v = u_1 v_1 + u_2 v_2 + \dots + u_n v_n$$

Intuitive understanding → Refer to next slide

Vector Product (Scalar Product) - Intuitive Understanding

Meaning of multiplying a number to real number

$$3 \times 2 = 6$$



We have a vector u .

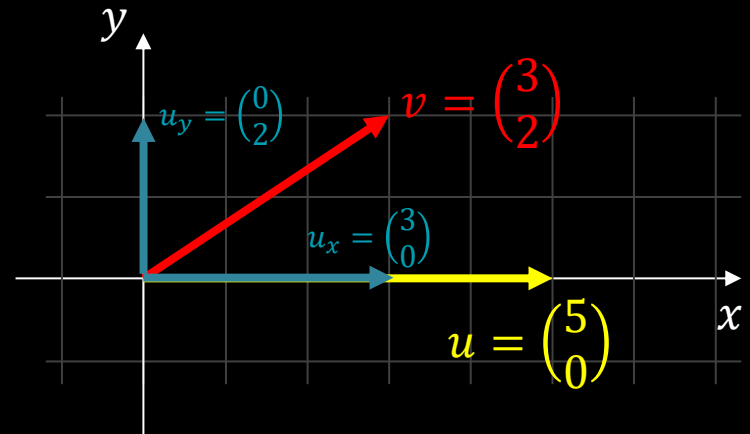
u 에 벡터 v 의 크기가 곱해졌을때

How does the size of vector u change?

$$u \cdot v = \|u\| \|v\| \cos \theta$$

Meaning of multiplying vector & vector

$$u \cdot v = \begin{pmatrix} 5 \\ 0 \end{pmatrix} \cdot \begin{pmatrix} 3 \\ 2 \end{pmatrix} = ???$$



1. Size of v ? $\|v\| = \sqrt{3^2 + 2^2} = \sqrt{13}$

Since v has directions
Not all $\sqrt{13}$ affect to u

벡터 u 와 방향이 일치한 크기만큼만 곱해주자!
 $u_x = \begin{pmatrix} 3 \\ 0 \end{pmatrix}$

$$\|u_x\| = \sqrt{3^2 + 0^2} = 3$$

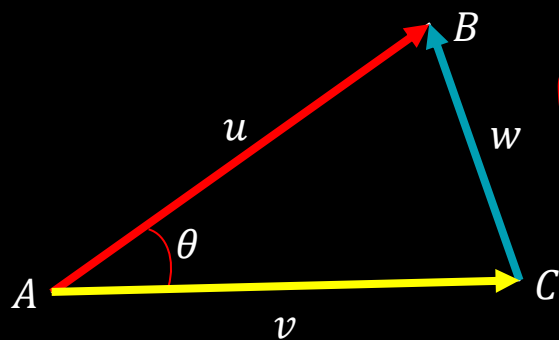
Geometric Interpretation

$$u \cdot v = \|u\| \|v\| \cos \theta$$

Why ???

$$u \cdot v = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n$$

Proof → The second Cosine Law



$$u = v + w$$

$$w = u - v$$

$$\|w\|^2 = \|u\|^2 + \|v\|^2 - 2\|u\|\|v\|\cos \theta$$

$$\|u - v\|^2 = \|u\|^2 + \|v\|^2 - 2u \cdot v$$

$$2u \cdot v = \|u\|^2 + \|v\|^2 - \|u - v\|^2$$

$$2u \cdot v = \sqrt{u_1^2 + \cdots + u_n^2}^2 + \sqrt{v_1^2 + \cdots + v_n^2}^2 - \sqrt{(u_1 - v_1)^2 + \cdots + (u_n - v_n)^2}^2$$

$$2u \cdot v = \cancel{u_1^2} + \cdots + \cancel{u_n^2} + \cancel{v_1^2} + \cdots + \cancel{v_n^2} - \left((\cancel{u_1^2} - 2u_1 v_1 + \cancel{v_1^2}) + \cdots + (\cancel{u_n^2} - 2u_n v_n + \cancel{v_n^2}) \right)$$

$$\cancel{2u \cdot v} = \cancel{2u_1 v_1} + \cdots + \cancel{2u_n v_n} \quad \Rightarrow \quad u \cdot v = u_1 v_1 + \cdots + u_n v_n$$

Vector Product (Scalar Product)

$$u \cdot v = \|u\| \|v\| \cos \theta$$

$$\cos \theta = \frac{u \cdot v}{\|u\| \|v\|}$$

Cosine Similarity(코사인 유사도)
Quantifies the similarity between two vectors
using the value of $\cos(\theta)$ between them.

- $u \cdot v$ can be expressed as $\langle u, v \rangle$
- The dot product satisfies the commutative property

$$u \cdot v = u^T v = v^T u = v \cdot u$$

$u \cdot v$ 에서 u 와 v 는 벡터를 의미하며 이 때 "."는 내적 연산을 의미

$\sum_{i=1}^n u_i v_i$ 에서 u_i 와 v_i 는 u 와 v 의 i 번째 성분을 의미

이 때 $u_i v_i$ 는 u_i 와 v_i 의 곱셈 연산을 한 값

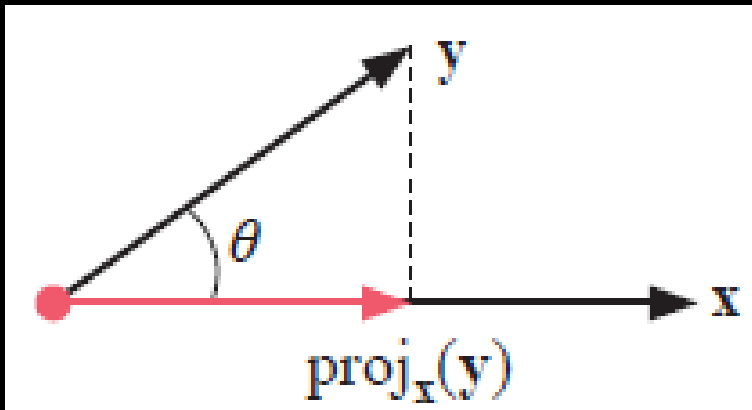
Orthogonal Vector

■ Orthogonal

- Two vectors that form a 90-degree angle with each other are said to be "orthogonal" to each other.
- If vector u and v are orthogonal, where $u \cdot v = 0$

■ Vector Projection

- Can be happen in 2D space or over dimensions



$$\text{Projection}_x(y) = \frac{x}{\|x\|} \cos(\theta)$$

Hyperplanes (초평면)

■ In a d -dimensional vector space

- A hyper-plane has $(d - 1)$ -dimensions
- A hyper-plane divides a higher-dimensional space into two separate regions

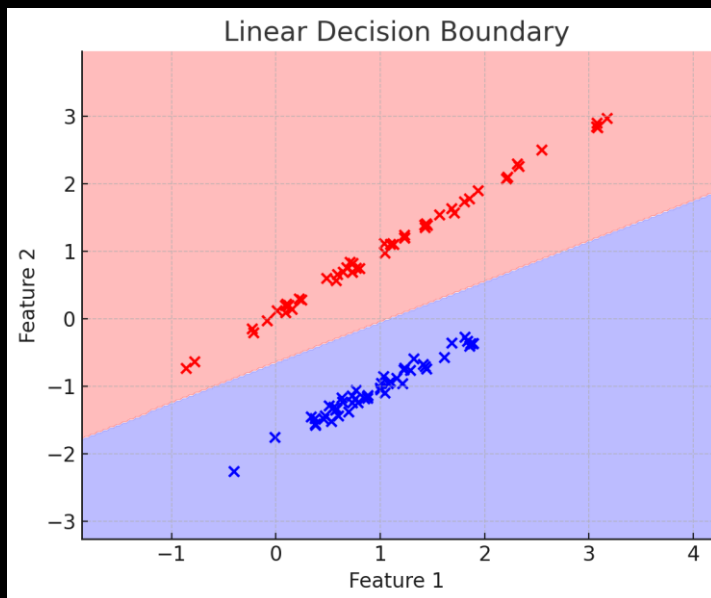
■ What 'Hyper' means?

- beyond, higher-dimensional, or extended from
- In 2D space, $ax + by + c = 0 \rightarrow$ A **line** be drawn on a 2D plane
- In 3D space, $ax + by + cz + d = 0 \rightarrow$ A **plane** can be drawn in a 3D space
- In 4D space, $ax_1 + bx_2 + cx_3 + dx_4 + e = 0 \rightarrow$ A **space** can be drawn in a 4D space
- In 5D space, $ax_1 + bx_2 + cx_3 + dx_4 + ex_5 + f = 0$
 - \rightarrow A **???** can be drawn in a 5D space
- In higher space, we need some generalized concept.
 - In math, it's called as "**hyperplane**"

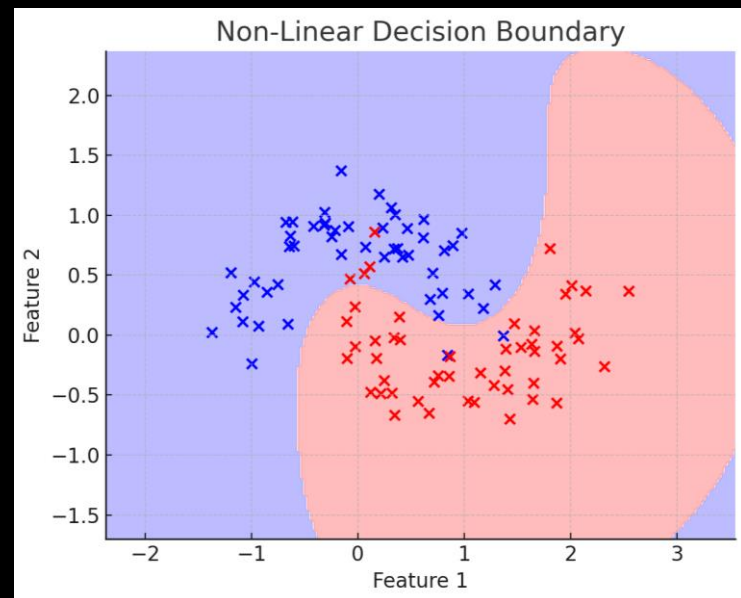
Decision Boundary in ML Classification

■ Hyperplane is a decision boundary in ML

- In 2D space
 - A linear classifier might draw a straight line to separate them.
 - A non-linear classifier might draw a more complex boundary.



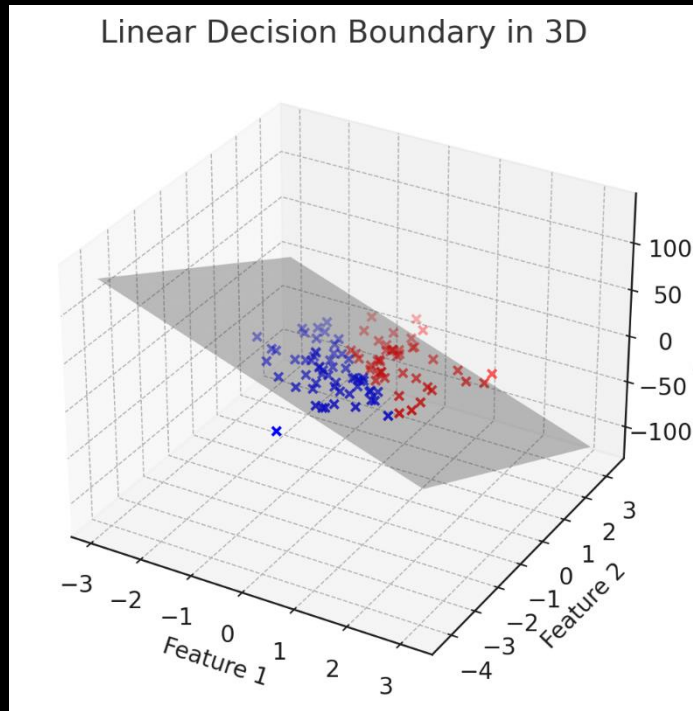
Linear boundary in 2D vector space



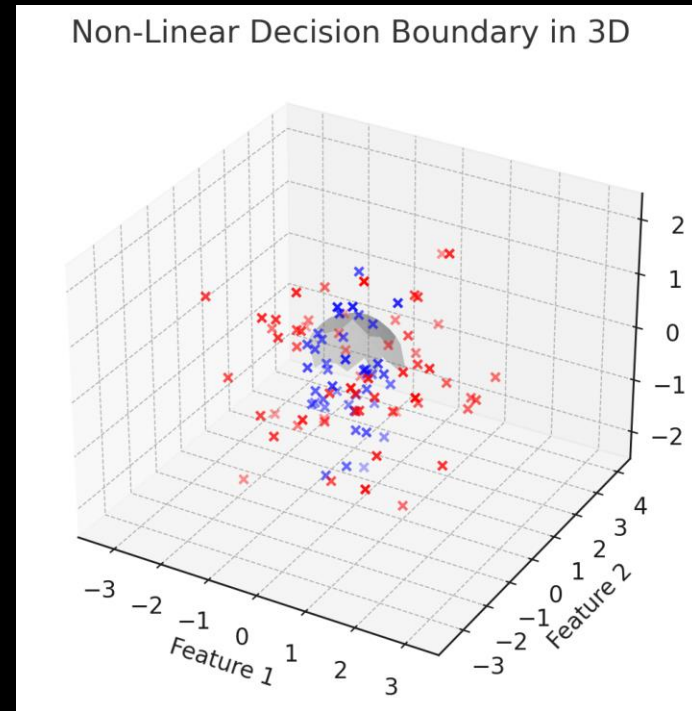
Non-linear boundary in 2D vector space

ML Classification example in 3D vector space

■ In 3D vector space,



Linear boundary in 3D vector space



Non-linear boundary in 3D vector space

Generalization of Decision Boundary in ML

■ A decision boundary is typically represented as a function of input features.

- In a linear classifier, the boundary is a **hyperplane** defined by:

$$W \cdot x + b = 0$$

where,

W is the weigh vector,

x is the input feature vecor,

b is the bias term.

If $W \cdot x + b > 0$, the data is classified as Class 1.

If $W \cdot x + b < 0$, the data is classified as Class 2.

Linear Algebra - Matrix

Matrix (행렬)

■ Matrix?

- Engineering
 - An arrangement of numbers or polynomials in a rectangular shape.
- For general audiences
 - A collection of numbers arranged in a rectangular form.
- The Origin of Meaning (Matrix)
 - Latin word mater (mother)
 - In English womb, matrix, foundation, array.
 - The basis for growth and development...
 - The "mother" of mathematics



Definition of Matrix

■ A matrix $m \times n$ over a Ring R

Note. Ring R : satisfy: the commutative property of addition, The associative property of multiplication, The multiplicative identity

- For each row $i \in \{1, 2, \dots, m\}$ and for each column $j \in \{1, 2, \dots, n\}$,
- A function maps each ordered pair (i, j) to an element $A_{ij} \in R$

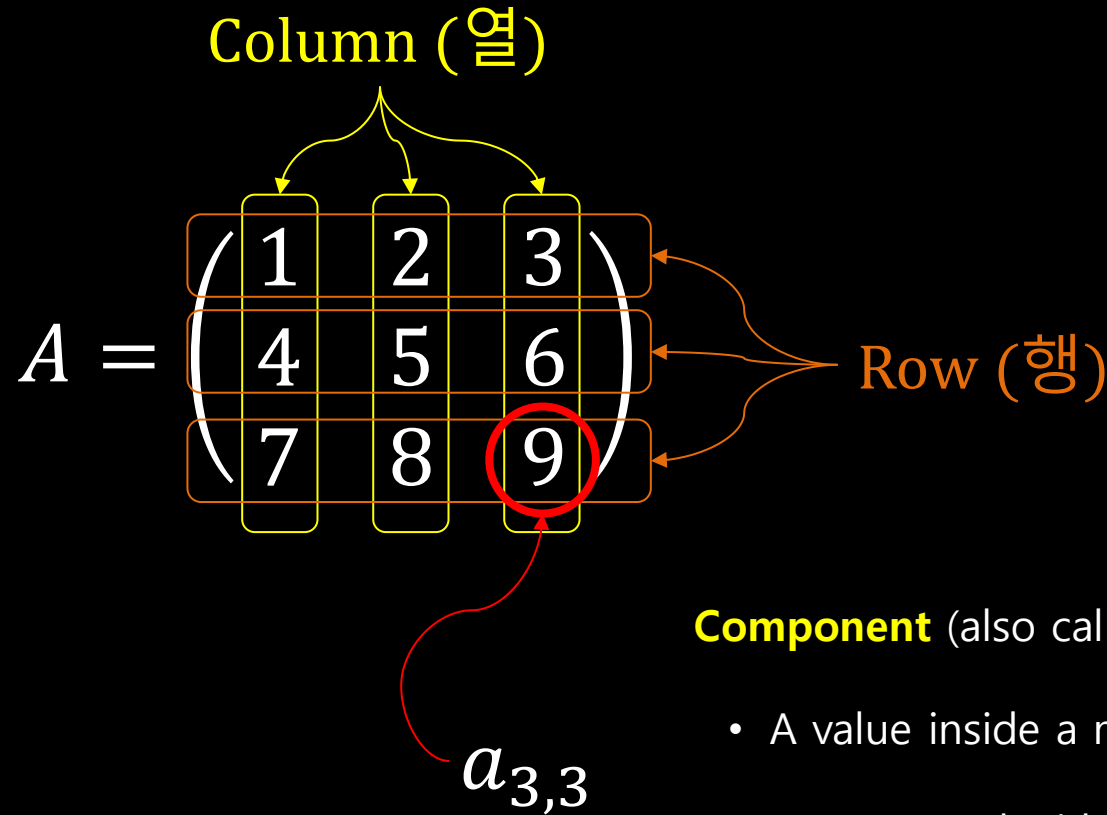
$$A = (a_{ij})_{i,j}$$

- According to a rigorous definition, a function that maps ordered pairs.

■ Representation

$$A = (a_{ij})_{i,j} \quad \text{or} \quad (A)_{i \times j} \quad \text{or} \quad A_{i,j} \quad \text{or} \quad A_{i \times j}$$

Components in Matrix



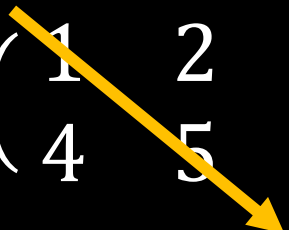
Component (also called term, element, or entry)

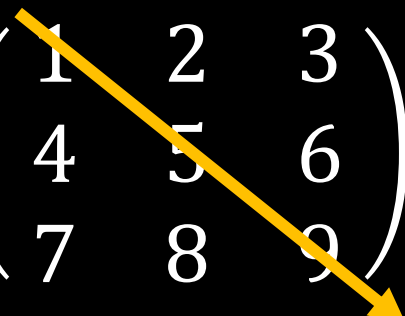
- A value inside a matrix
- Represented with a lowercase letter
- and indexed using subscripts

Diagonal

Diagonal (주대각선)


행렬의 왼쪽 위에서 오른쪽 아래를 가로지르는 선

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


$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$


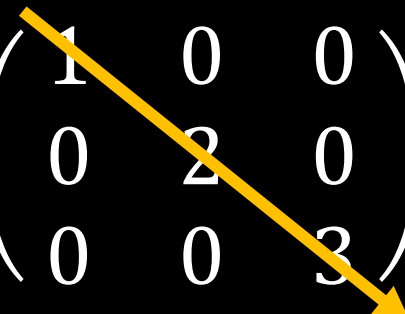
Diagonal Entry (대각성분)

주대각선 위의 성분

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


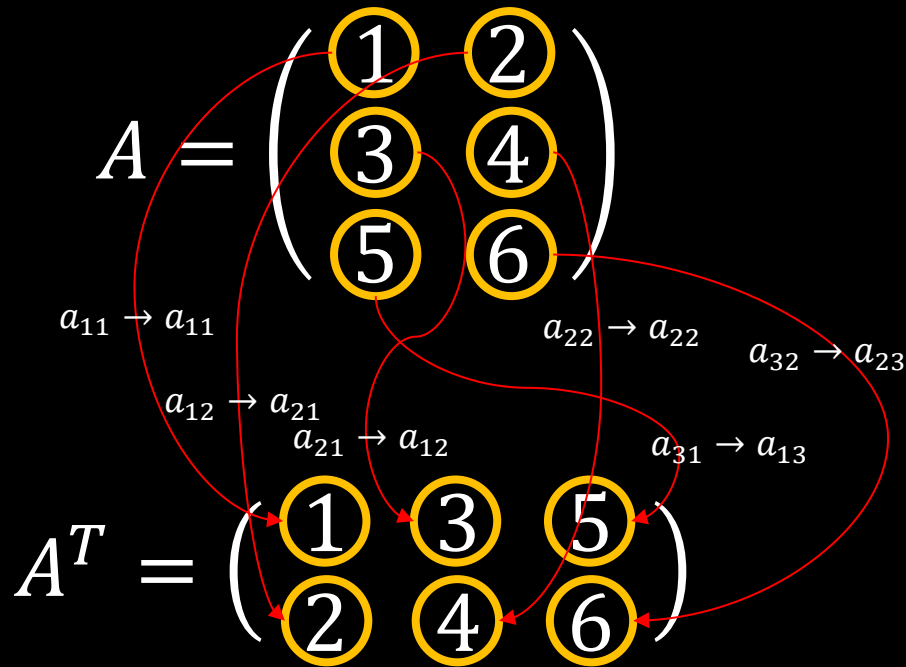
Diagonal Matrix (대각행렬)

대각성분이 아닌 모든 원소가 0 인 정사각행렬

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$


Transpose Matrix

Transpose Matrix (전치 행렬): 행렬 이름 윗 첨자로 T 표시
(a_{ij})에 대하여 (a_{ji}) \rightarrow 위치 인덱스를 바꾸어서 만든 행렬



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

주 대각선을 기준으로
대칭 이동

Types of Matrix

Zero (Null) Matrix (영행렬), 0 으로 표기
행렬의 모든 원소가 0 인 행렬

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Symmetric Matrix (대칭행렬)
 $A = A^T$ 인 행렬

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \quad A^T = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$$

Square Matrix (정사각행렬)
행과 열의 개수가 같은 행렬

Identity Matrix (단위행렬), I_x 로 표기
모든 대각성분이 1, 나머지는 0 인 정사각행렬
행렬에서 항등원 역할
(곱하기의 1, 더하기의 0과 같은 역할)

$$I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$AI = A \quad IA = A$$

Matrix Operation - Add/Sub/Scalar

■ Matrix Addition / Subtraction

$$A \pm B = (a_{ij} \pm b_{ij})$$

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \quad A + B = \begin{pmatrix} 1+5 & 2+6 \\ 3+7 & 4+8 \end{pmatrix} = \begin{pmatrix} 6 & 8 \\ 10 & 12 \end{pmatrix}$$

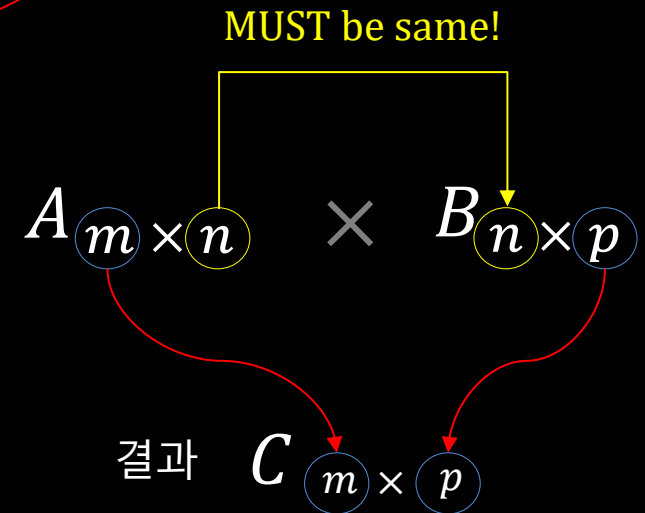
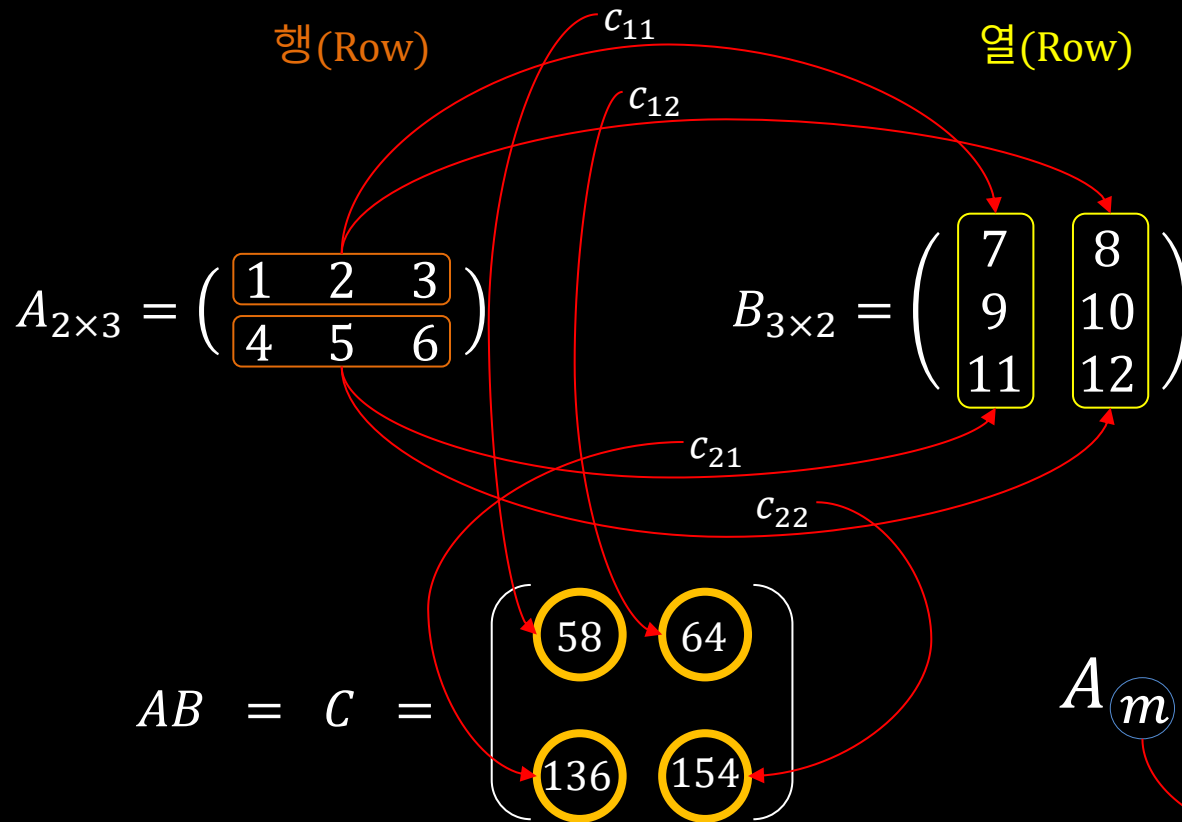
$$B = \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} \quad A - B = \begin{pmatrix} 1-5 & 2-6 \\ 3-7 & 4-8 \end{pmatrix} = \begin{pmatrix} -4 & -4 \\ -4 & -4 \end{pmatrix}$$

■ Scalar Multiplication (also called "scalar product" ^^.)

$cA = (ca_{ij})$, where c is a constant number (NOT a matrix)

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \quad 3A = 3 \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 3 & 6 \\ 9 & 12 \end{pmatrix}$$

Matrix Operation - Multiplication (Intuitive)



Matrix Operation - Multiplication (Math Notation)

$$A_{m \times n}$$

$$B_{n \times p}$$

$$AB = (c_{ik}) \text{ such that } c_{ik} = \sum_{j=1}^n a_{ij} \times b_{jk}$$

$$A_{2 \times 3} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$$

$$B_{3 \times 2} = \begin{pmatrix} 7 & 8 \\ 9 & 10 \\ 11 & 12 \end{pmatrix}$$

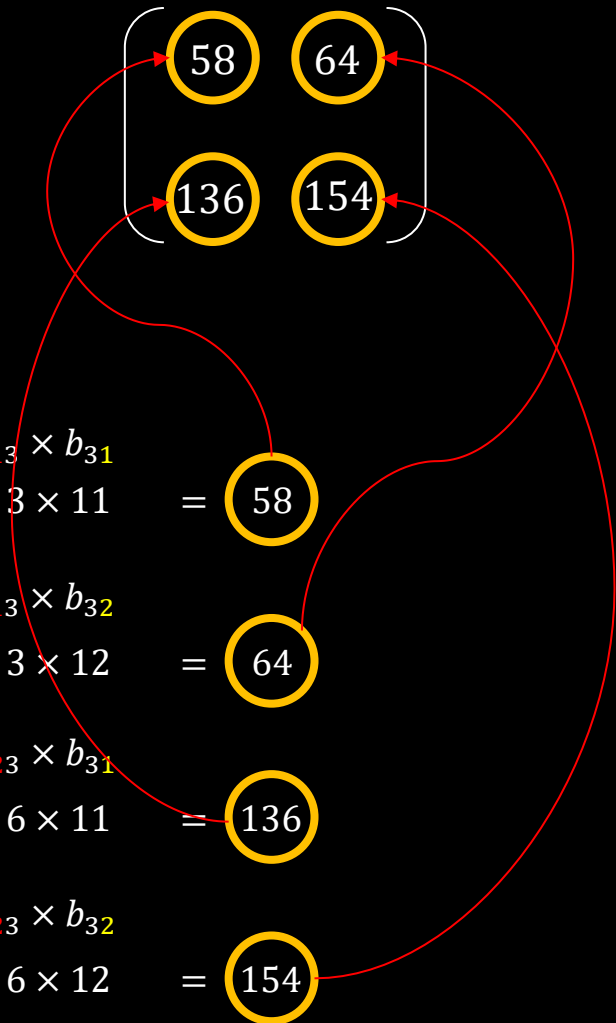
$$AB = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \begin{pmatrix} 7 & 8 \\ 9 & 10 \\ 11 & 12 \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}$$

$$c_{11} = \sum_{j=1}^3 a_{1j} \times b_{j1} = a_{11} \times b_{11} + a_{12} \times b_{21} + a_{13} \times b_{31} = 1 \times 7 + 2 \times 9 + 3 \times 11 = 58$$

$$c_{12} = \sum_{j=1}^3 a_{1j} \times b_{j2} = a_{11} \times b_{12} + a_{12} \times b_{22} + a_{13} \times b_{32} = 1 \times 8 + 2 \times 10 + 3 \times 12 = 64$$

$$c_{21} = \sum_{j=1}^3 a_{2j} \times b_{j1} = a_{21} \times b_{11} + a_{22} \times b_{21} + a_{23} \times b_{31} = 4 \times 7 + 5 \times 9 + 6 \times 11 = 136$$

$$c_{22} = \sum_{j=1}^3 a_{2j} \times b_{j2} = a_{21} \times b_{12} + a_{22} \times b_{22} + a_{23} \times b_{32} = 4 \times 8 + 5 \times 10 + 6 \times 12 = 154$$



Matrix Operation - Properties

■ Properties in Matrix Operations

$$A + B = B + A$$

$$A(B + C) = AB + AC$$

$$(A + B)^T = A^T + B^T$$

$$A(BC) = (AB)C$$

$$(A^T)^T = A$$

$$(AB)^T = B^T A^T$$

Types of Matrix Operation

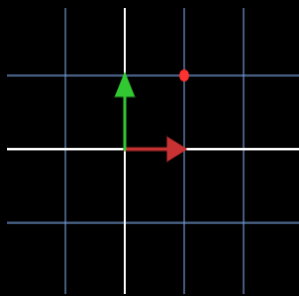
Finally, we want to know this

- 행렬에 어떤 행렬을 곱한다는 의미....

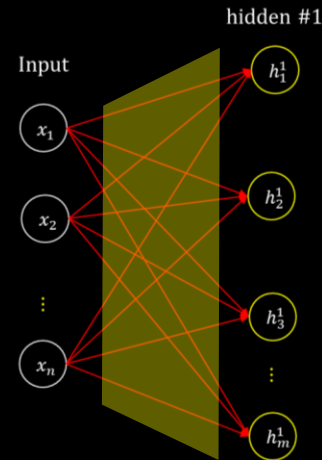
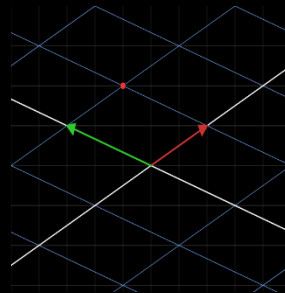
$$\begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 20 \\ 12 \end{pmatrix}$$

- 어떤 행렬에 역행렬을 곱한다는 의미

$$\begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 20 \\ 12 \end{pmatrix}$$



A
선형 변환
→
←
선형 변환
 A^{-1}



$$h^1 = \begin{pmatrix} w_{11} & w_{21} & \cdots & w_{n1} \\ w_{12} & w_{22} & \cdots & w_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1m} & w_{2m} & \cdots & w_{nm} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix}$$

$$h^1 = W_1^T X + b_1$$

First, check how to compute!

- Linear Equation using Inverse Matrix
- Determinant
- Inverse Matrix

Yet, still hard to understand πππ

Inverse Matrix on System of Linear Equations

In System of Linear Equations $AX = B$

If Inverse matrix A^{-1} of A exists,

$$X = A^{-1}B$$

$$\begin{pmatrix} 1 & 2 \\ 1 & -3 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 6 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} x \\ y \end{pmatrix} = \boxed{\begin{pmatrix} 1 & 2 \\ 1 & -3 \end{pmatrix}^{-1}} \begin{pmatrix} 1 \\ 6 \end{pmatrix}$$

If A^{-1} exists??

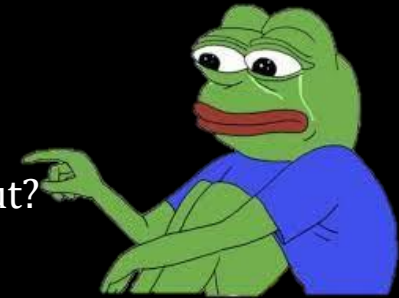
YES!

A unique solution exists.

NO!

No solution or infinitely many solutions.

How to check it out?



We can check it by using
determinant (행렬식) ^^

Determinant - Definition & Notation

■ Determinant

- 한국어로 '행렬식' 이라고 부름
- A function maps Square Matrix into a scalar

In fact, determinants have even deeper meanings.

We will learn them step by step. ^^

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \xrightarrow{\text{determinant}} \text{Real value}$$

$$f(\text{Square Matrix}) = \mathbb{R}$$

Representation in *Linear Algebra* ^^

$$\det A \quad \text{or} \quad |A|$$

Determinant - Operations



Computing Determinant !!

Don't worry...
The computer does the calculations.
Just focus on understanding
the concept. ^^

$$0 \times 0 \rightarrow \det(A) = 0$$

$$1 \times 1 \rightarrow \det(a) = a \quad \leftarrow \text{Only one value exists}$$

$$2 \times 2 \rightarrow \det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = a_{11}a_{22} - a_{12}a_{21}$$

$$3 \times 3 \rightarrow \det \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = a_{11}M_{11} - a_{12}M_{12} + a_{13}M_{13}$$

$$= a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{22} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

$$4 \times 4 \rightarrow \det \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} = a_{11}M_{11} - a_{12}M_{12} + a_{13}M_{13} - a_{14}M_{14}$$

⋮

⋮

⋮

The matrix obtained
by removing the i -th
row and j -th column
from the original
matrix.

M_{ij}

Computing the Inverse Matrix

'Cofactor (여인수)' 라고 부름
원래 행렬에서
 i 행과 j 열을 제외한 행렬

$$A^{-1} = \frac{1}{\det A} \begin{pmatrix} C_{11} & C_{21} & \cdots \\ C_{12} & C_{22} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}, \text{ where } C_{ij} = (-1)^{i+j} \times M_{ij}$$

$$AX = I$$

We call it as 'Adjoint Matrix'

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & a_{nn} \end{pmatrix} \frac{1}{\det A} \begin{pmatrix} C_{11} & C_{21} & \cdots & C_{n1} \\ C_{12} & C_{22} & \cdots & C_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ C_{1n} & C_{2n} & \cdots & C_{nn} \end{pmatrix}$$

$$= \frac{1}{\det A} \begin{pmatrix} \det A & 0 & \cdots & 0 \\ 0 & \det A & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \det A \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} = I$$

The inverse matrix also satisfies the commutative property. ^^

$$AA^{-1} = A^{-1}A = I$$

Toy Example

Solving Linear System using Inverse Matrix

Toy Example

$$A^{-1} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

$$AX = B \quad \begin{pmatrix} 1 & 2 \\ 1 & -3 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 6 \\ 1 \end{pmatrix}$$

영영사전 9

결정인자

determinant +

1. 명사 formal a thing that controls or influences what happens
2. 명사 formal often + of

If inverse matrix exist: $ad - bc \neq 0$
(Non-zero determinant exist)

Multiply A^{-1} on the right side of both sides..

$$A^{-1}AX = A^{-1}B \quad \frac{1}{1 \cdot -3 - 2 \cdot 1} \begin{pmatrix} -3 & -2 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 1 & -3 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{1 \cdot -3 - 2 \cdot 1} \begin{pmatrix} -3 & -2 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 6 \\ 1 \end{pmatrix}$$

Just simply re-arrange equations

$$IX = A^{-1}B \quad \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = -\frac{1}{5} \begin{pmatrix} -3 & -2 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 6 \\ 1 \end{pmatrix}$$

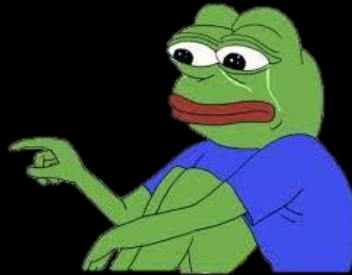
$$X = A^{-1}B \quad \begin{pmatrix} x \\ y \end{pmatrix} = -\frac{1}{5} \begin{pmatrix} -18 - 2 \\ -6 + 1 \end{pmatrix} = -\frac{1}{5} \begin{pmatrix} -20 \\ -5 \end{pmatrix} = \begin{pmatrix} 4 \\ 1 \end{pmatrix}$$

$$x = 4, y = 1$$

Do I have to know all of this????????????????

■ What!!!

- Hey prof! Too complex π
- Do we must know this?
 - Yes!!
- However, we dont need to compute in hand ^^
 - We will use Python package "numpy"



However, you must understand
how it works!!!

```
import numpy as np

matrix = np.array([[2, 5], [1, 3]]) # 2x2 행렬 생성
matrix_det = np.linalg.det(matrix) # determinant 계산
print(matrix_det)                  # 결과 출력
>>> 1.0

matrix_inverse = np.linalg.inv(matrix) # inverse matrix(역행렬) 계산
print(matrix_inverse)
>>> [[ 3. -5.]
      [-1.  2.]
```


Dependent vs. Independent

■ Linear Combination (선형 조합)

- For vectors b_1 and b_2 ,
- and real numbers a_1 and a_2 , the expression
- $a_1b_1 + a_2b_2$ is called a **linear combination** of b_1 and b_2

■ Linearly Dependent

- A set of vectors v_1, \dots, v_k is linearly dependent
 - if there exist non-zero scalars a_1, \dots, a_k such that

$$a_1v_1 + \dots + a_kv_k = 0$$

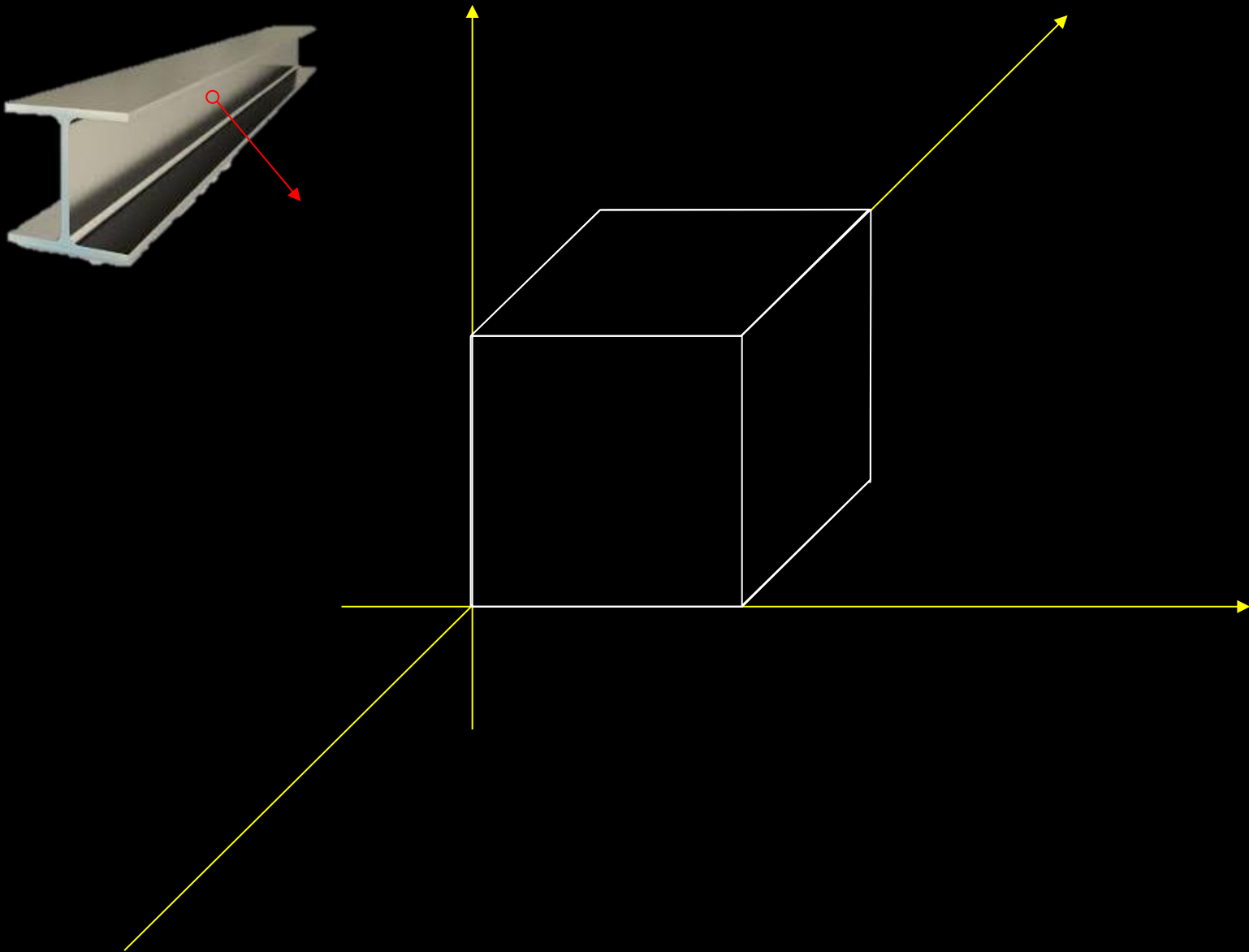
■ Linearly Independent

- If the equation

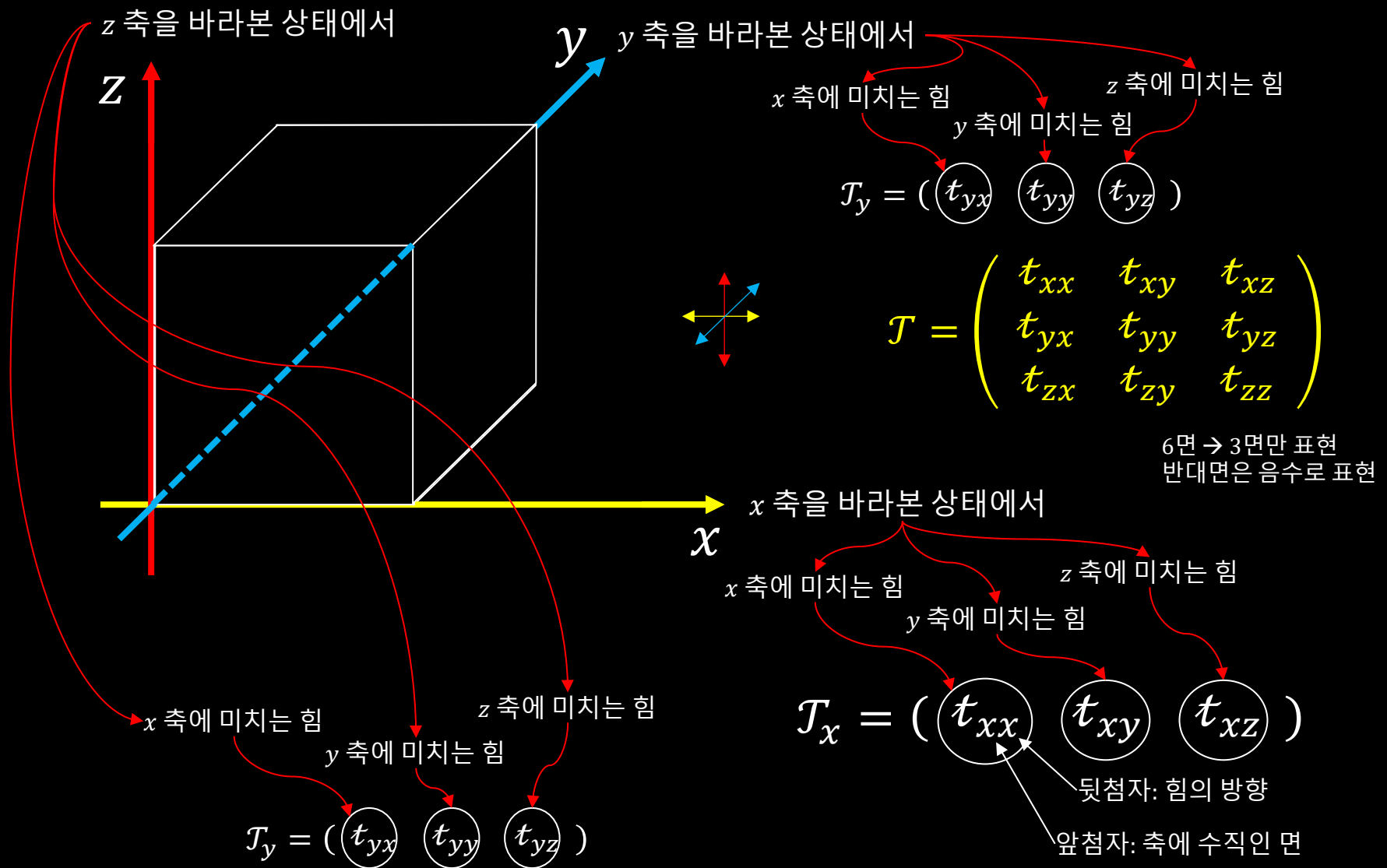
$$\sum_{i=1}^n a_i v_i = 0$$

holds only when $a_1 = \dots = a_n = 0$

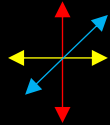
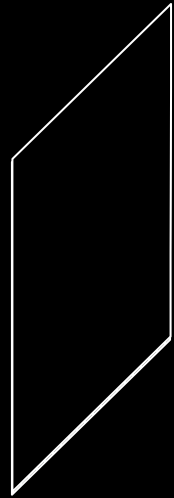
Tensor in Pysics



Tensor in Pysics



Tensor in Pysics - Extension 조금 더 확장하기



Rank: \mathcal{T} 의 원소들이 갖는 Basis 수

Rank 2

의미는 다르지만,
어쨌든 Matrix와 같은 모양 ^^

$$\mathcal{T} = \begin{pmatrix} t_{xx} & t_{xy} & t_{xz} \\ t_{yx} & t_{yy} & t_{yz} \\ t_{zx} & t_{zy} & t_{zz} \end{pmatrix}$$

\mathcal{T}_{ijk}

i : row index

$$\begin{pmatrix} t_{xxx} & t_{xyx} & t_{xzx} \\ t_{yxx} & t_{yyx} & t_{yzx} \\ t_{zxx} & t_{zyx} & t_{zzx} \end{pmatrix} \begin{pmatrix} t_{xxz} & t_{xyz} & t_{xzz} \\ t_{yyz} & t_{yzz} & t_{zzz} \end{pmatrix}$$

j : column index

k : object index

$$\begin{pmatrix} t_{xxz} & t_{xyz} & t_{xzz} \\ t_{yyz} & t_{yzz} & t_{zzz} \end{pmatrix}$$

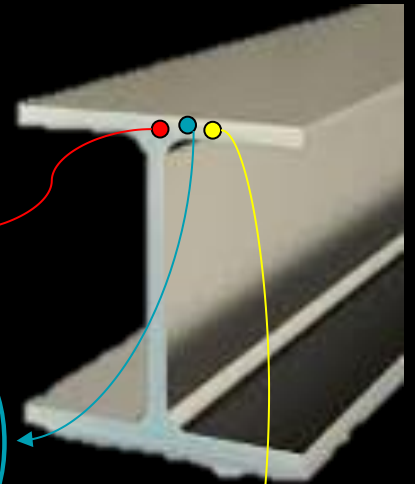
$$\begin{pmatrix} t_{xxy} & t_{xyy} & t_{xzy} \\ t_{yxy} & t_{yyx} & t_{yzy} \\ t_{zxy} & t_{zyx} & t_{zzy} \end{pmatrix}$$

의미는 다르지만, 어쨌든! Matrix를
여러 개 표현할 수 있는 모양 ^^



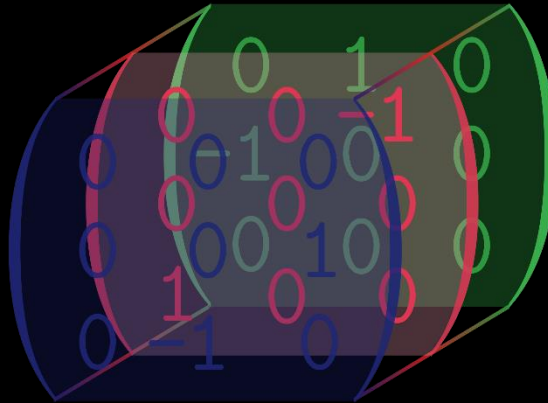
우리가 찾던 것!!!

$$\begin{pmatrix} t_{xxx} & t_{xyx} & t_{xzx} \\ t_{yxx} & t_{yyx} & t_{yzx} \\ t_{zxx} & t_{zyx} & t_{zzx} \end{pmatrix}$$



Let's use Tensor in ML

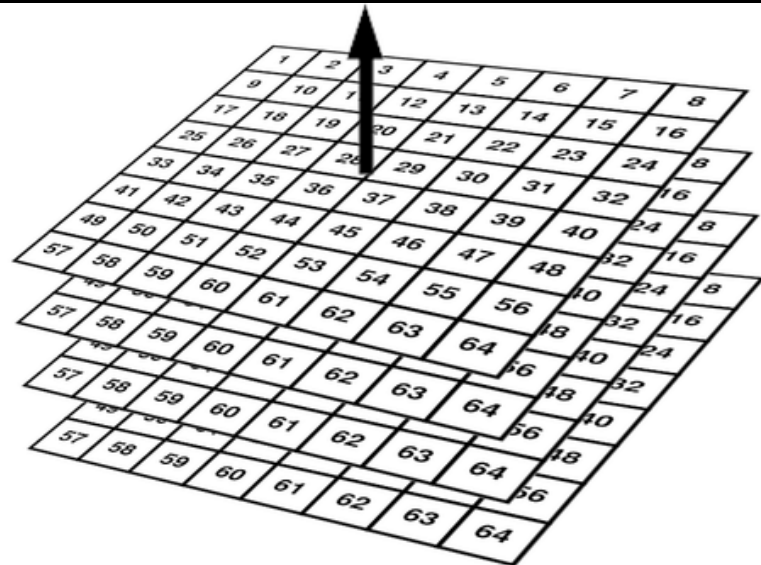
$$T_{ijk} =$$



Oh
Yes!

이제는 matrix 보다 더 편리한
자료구조로 Tensor를 쓰면 되겠구나!!

1	2	3	4	5	6	7	8
9	10	11	12	13	14	15	16
17	18	19	20	21	22	23	24
25	26	27	28	29	30	31	32
33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48
49	50	51	52	53	54	55	56
57	58	59	60	61	62	63	64



Deep learning Tensor

사실 Tensor를 정확히 이해하는 것은 매우 어렵습니다.

간단하게 생각하면 선형대수 공부하면서 다루었던
Scalar, Vector, Matrix를 포함하는 포괄적인 데이터
표현이라고 생각할 수 있습니다.

우리는 딥러닝 학습/추론을 위한 다차원 데이터 자료구조가 필요할 뿐입니다.




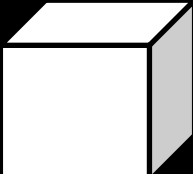
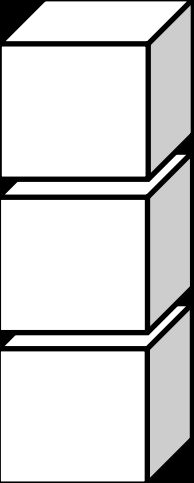
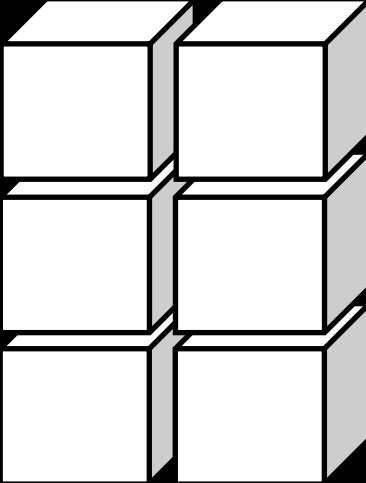
우리는 그냥 행렬을 여러 개 담아 놓은 배열이라고 생각하겠습니다.

물리적 특성은 일단 무시합니다.

Python에서 제공하는 List, Tuple, Set, Dictionary와 같은
자료구조 중 하나라고 생각합니다.

Tensor Simple Understanding → Computer Science Version!

원소(entry)를 표현하기 위해 필요한 기저(basis)의 수

Rank (차원)	0	1	2	3	4	5	...
Name (명칭)	scalar	vector	matrix	3D tensor	4D tensor	5D tensor	...
Visualization (시각화)							...
We know these				Need more data structure			

Tensor Example in Dataset

■ Tensors

- A tensor is an n -dimensional array of scalars.
 - Vector: 1D tensor, $\mathbf{v} \in \mathbb{R}^n$
 - Matrix: 2D tensor, $\mathbf{A} \in \mathbb{R}^{m \times n}$
 - 4D tensor: $\mathbf{T} \in \mathbb{R}^{n_1 \times n_2 \times n_3 \times n_4}$

■ An RGB image

- 3D array, making it a 3D tensor.
- The three axes correspond to width, height, and channels
 - e.g., $224 \times 224 \times 3$
- The channel axis corresponds to the color channels
 - red, green, and blue

Tensor Operation in Deeplearning

In fact, a Tensor is a multi-dimensional array, allowing numerous operations.

It varies depending on which dimension or axis is used as the reference ^^.

Ultimately, it must be broken down to enable Matrix, Vector, and Scalar operations.

Let's explore methods for decomposing a Tensor ^^.

Move to next slide~~

Tensor Decomposition

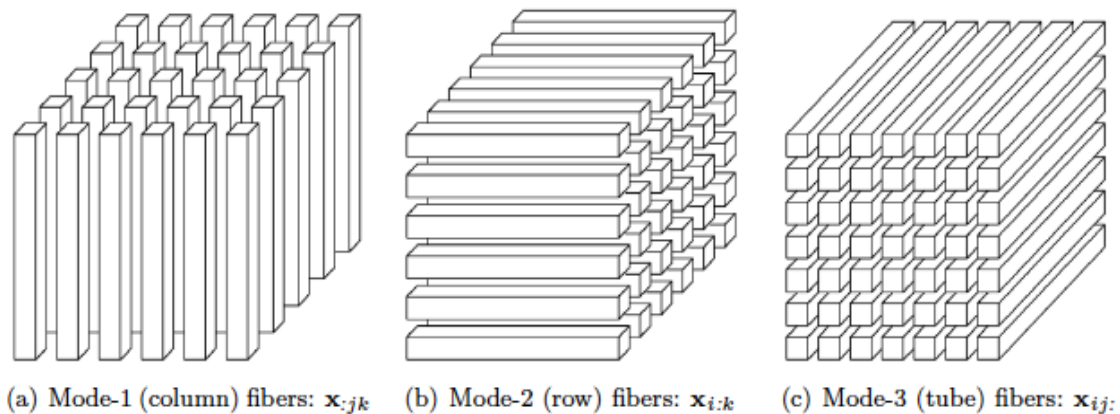


Fig. 2.1 *Fibers of a 3rd-order tensor.*

Fiber decomposition

1개의 인덱스는 자유롭게
나머지 인덱스는 모두 고정
설정에 따라 다양한
Vector가 생성됨

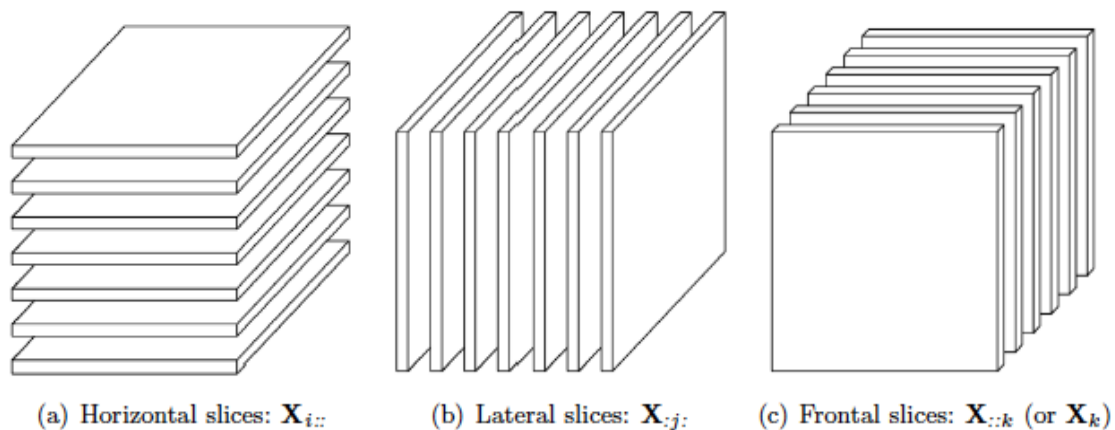


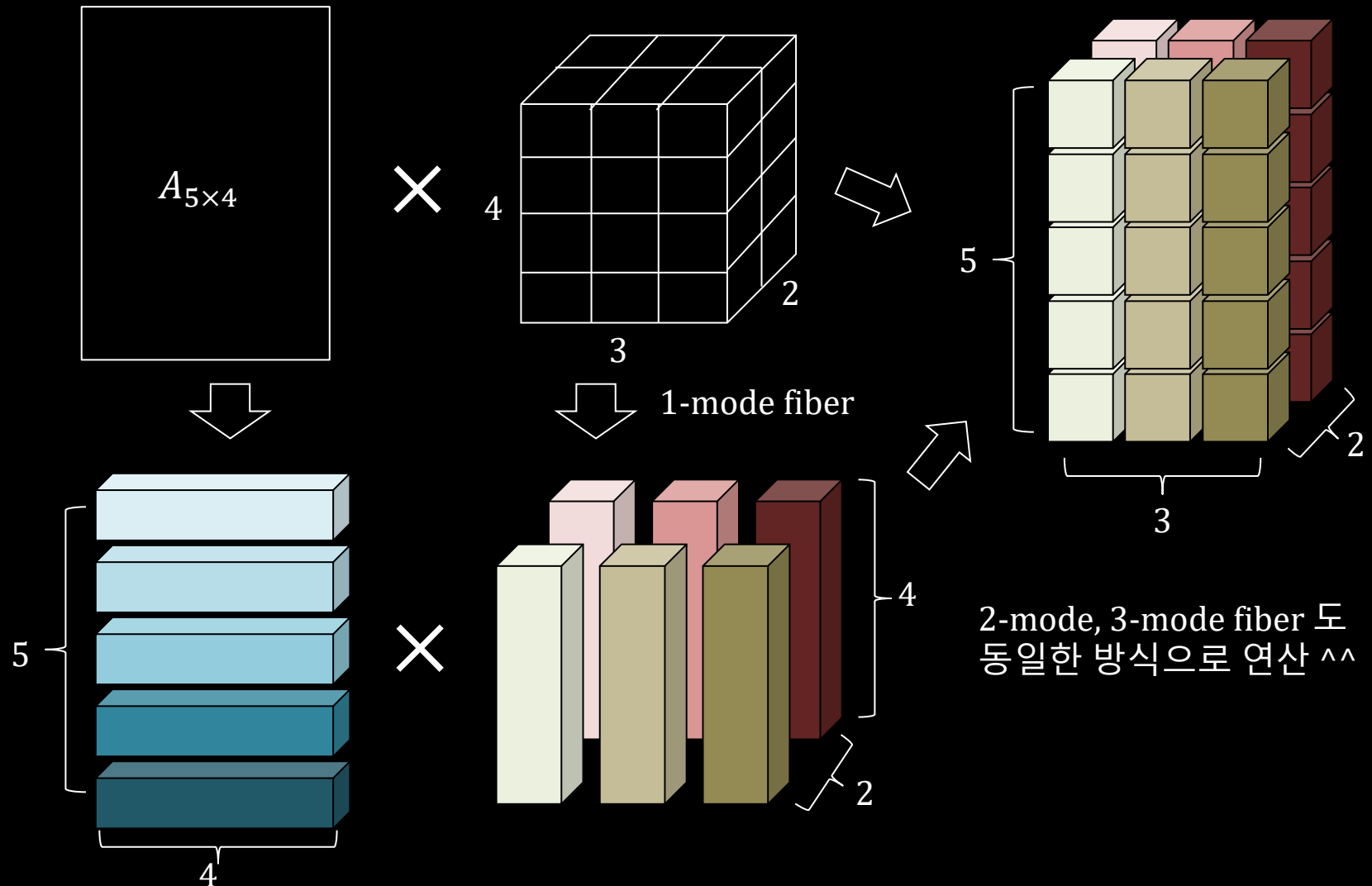
Fig. 2.2 *Slices of a 3rd-order tensor.*

Slice decomposition

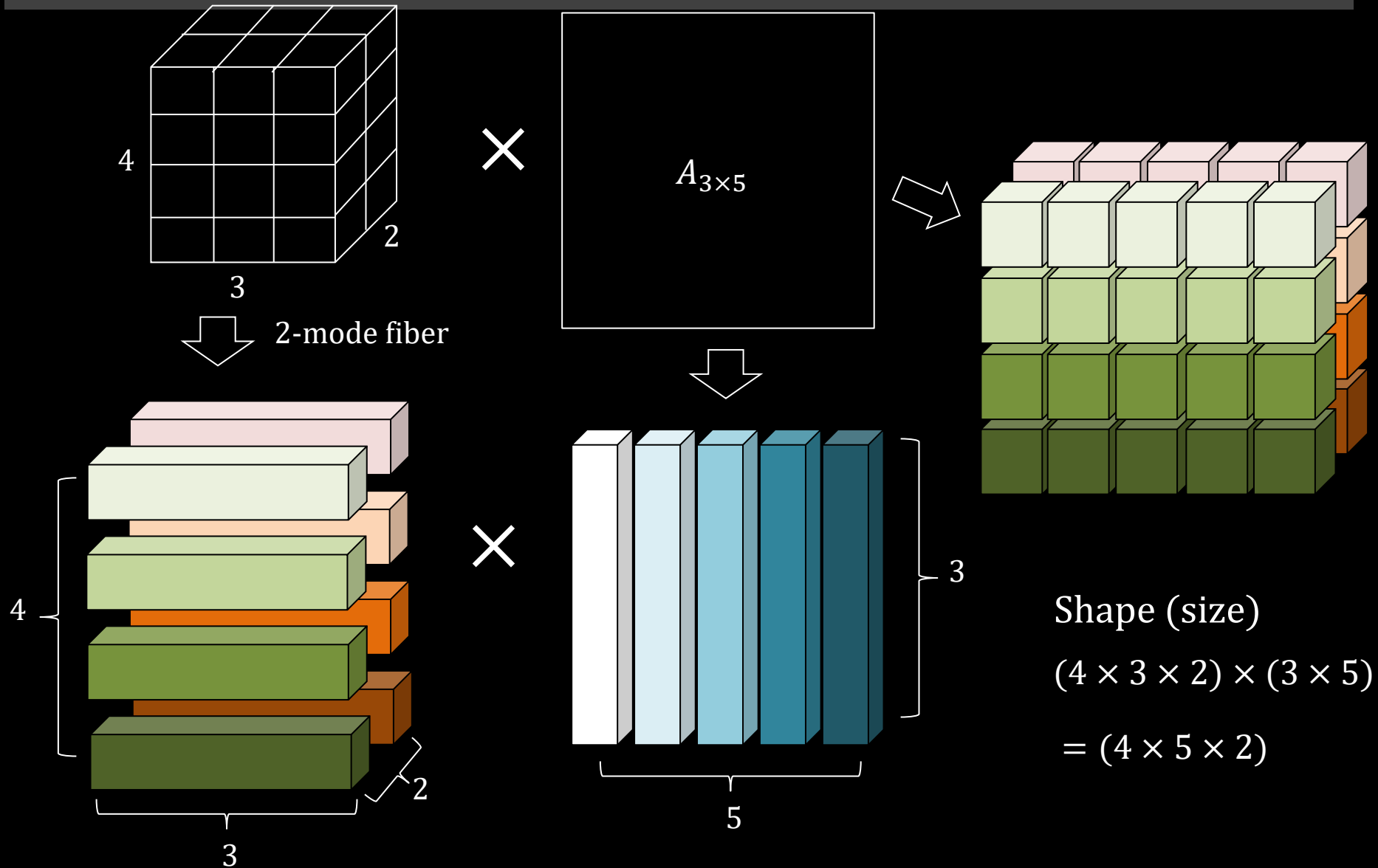
2개의 인덱스는 자유롭게
나머지 인덱스는 모두 고정
설정에 따라 다양한
Matrix가 생성됨

이밖에 다양한 방법도 가능

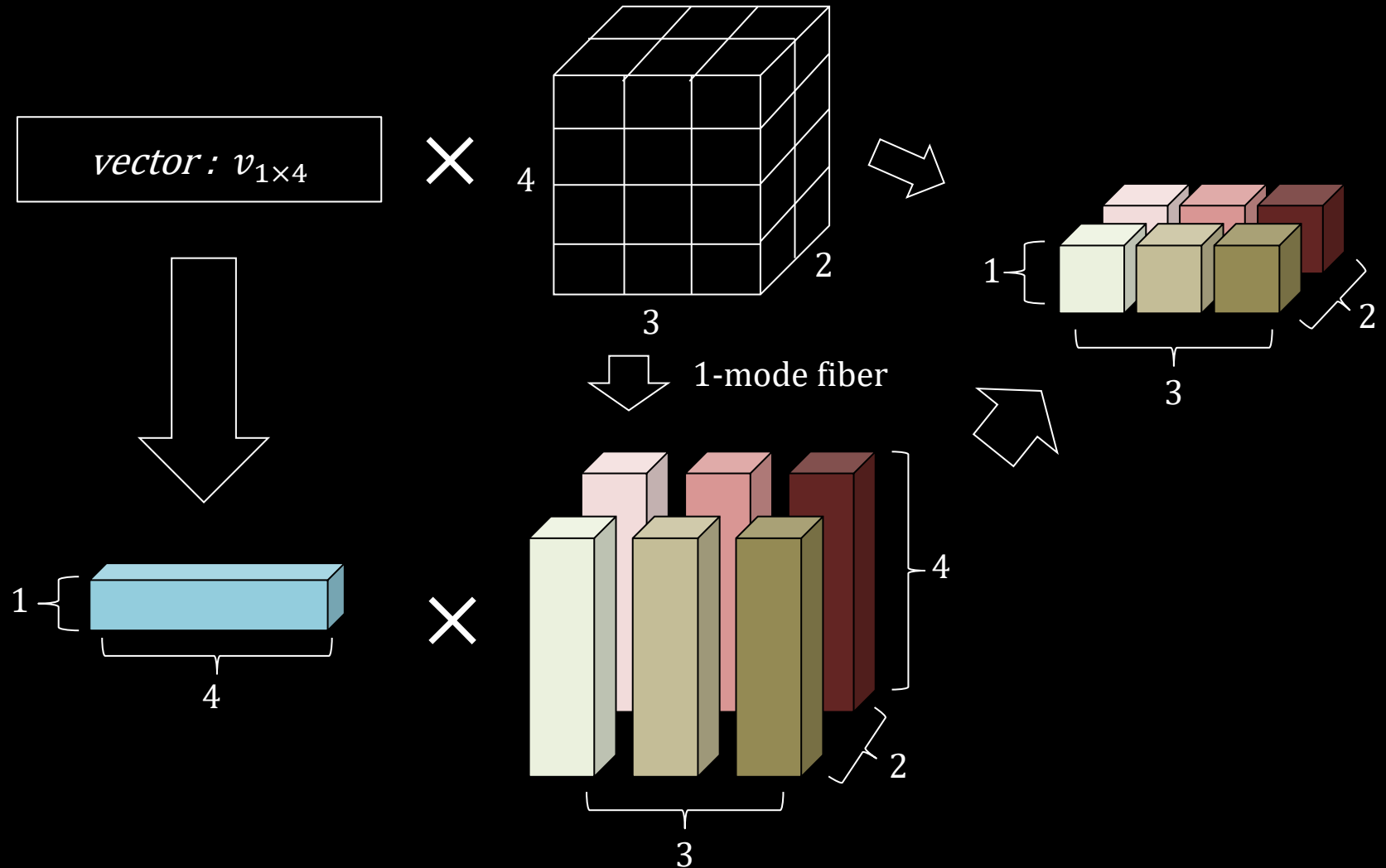
Operation in Deeplearning: $Matrix \times Tensor$ (1-mode fiber)



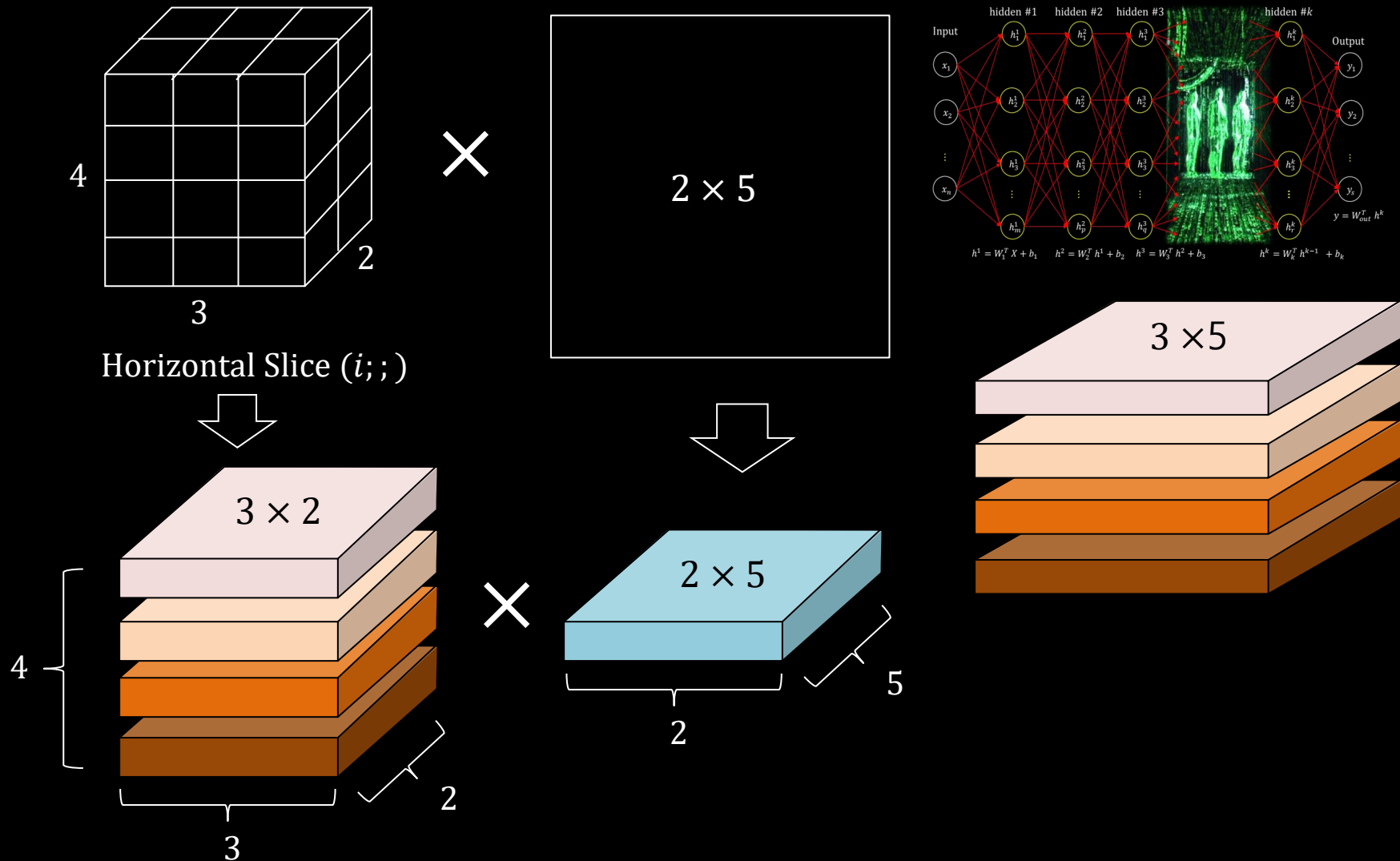
Operation in Deeplearning: *Tensor* (2-mode fiber) \times *Matrix*



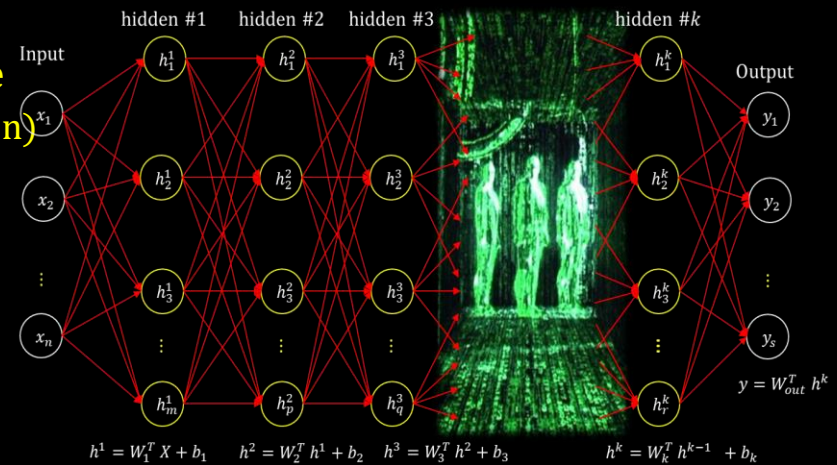
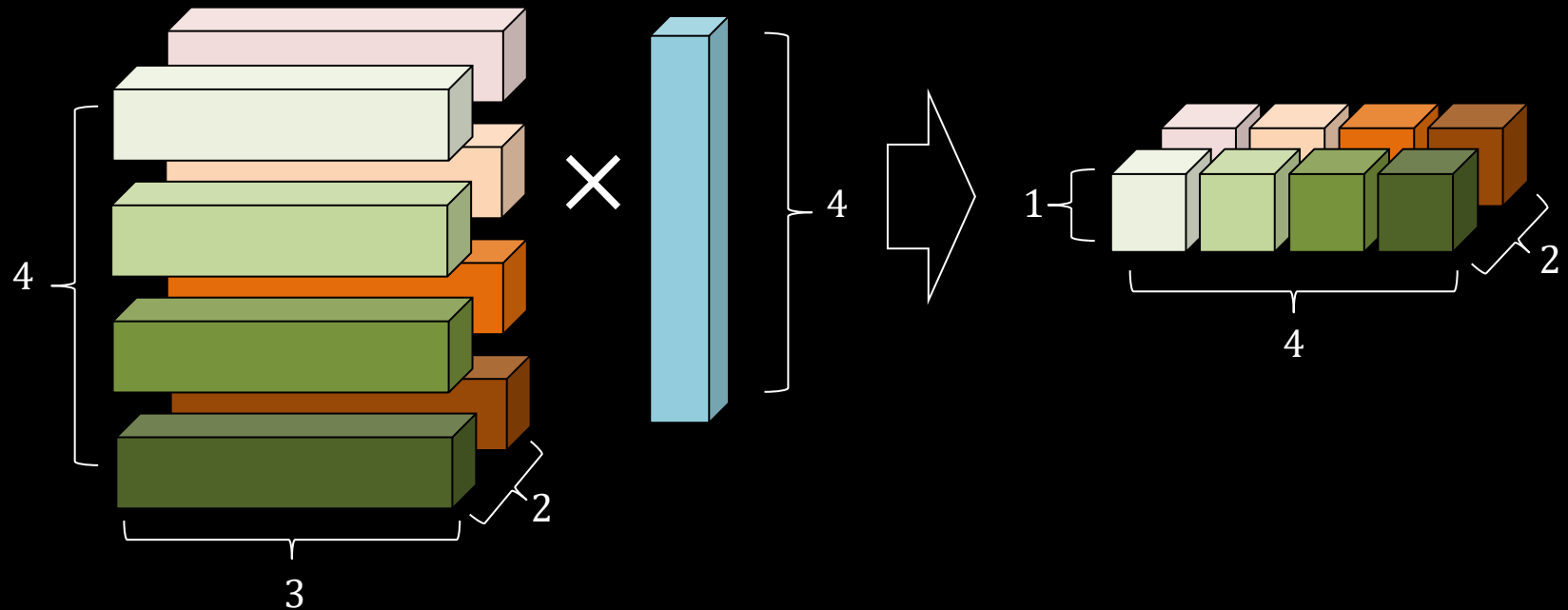
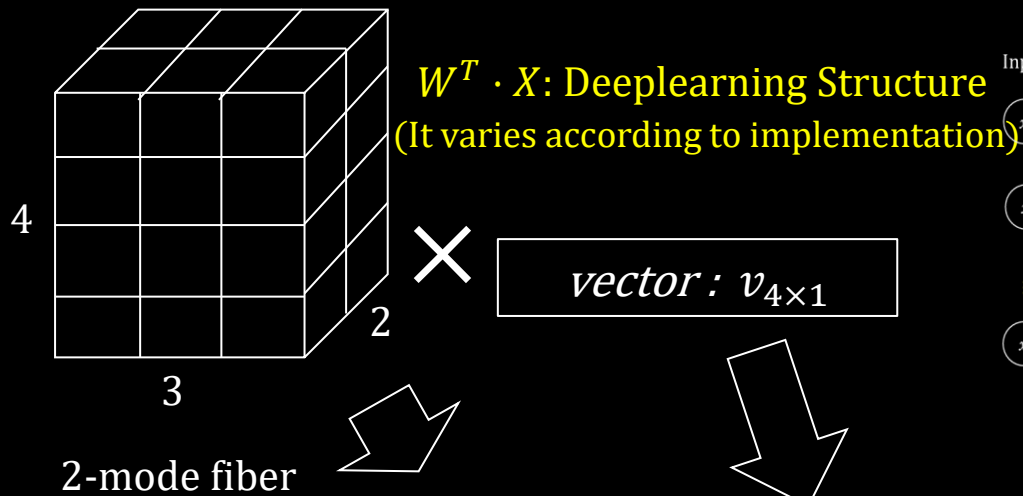
Operation in Deeplearning: $Vector \times Tensor$ (1-mode fiber)



Operation in Deeplearning: *Tensor (slice) \times Matrix*



Operation in Deeplearning: $Tensor(2\text{-mode fiber}) \times Vector$



Derivative

Derivative in ML

■ Definition of derivative

- The derivative of the function $f: \mathbb{R} \rightarrow \mathbb{R}$ is defined as follows.

$$f'(x) = \frac{d}{dx} f(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

- If the limit value of $f'(a)$ exists, then f is differentiable at a .
- If $f'(c)$ exists for all $c \in [a, b]$, then f is differentiable on this interval.
- The derivative $f'(x)$ can also be interpreted as the instantaneous rate of change of $f(x)$ with respect to x .
- The symbols $\frac{d}{dx}$, D , and D_x represent differentiation operators.
- If x is the independent variable and y is the dependent variable, given $y=f(x)$
 - then the following expressions are equivalent:

$$f'(x) = f' = \frac{dy}{dx} = \frac{df}{dx} = \frac{d}{dx} f(x) = Df(x) = D_x f(x)$$

Frequently used Differential Equations

상수값에 대한 미분	$\frac{d}{dx}c = 0$
선형 함수에 대한 미분	$\frac{d}{dx}(ax) = a$
거듭제곱 ^{Power} 에 대한 미분	$\frac{d}{dx}x^n = nx^{n-1} (n \text{은 양의 정수})$
지수 함수에 대한 미분	$\frac{d}{dx}e^x = e^x$
로그에 대한 미분	$\frac{d}{dx}\log(x) = \frac{1}{x}$
덧셈 규칙에 대한 미분	$\frac{d}{dx}(g(x) + h(x)) = \frac{d}{dx}g(x) + \frac{d}{dx}h(x)$
곱셈 규칙에 대한 미분	$\frac{d}{dx}(g(x) h(x)) = g(x) \left(\frac{d}{dx}h(x) \right) + \left(\frac{d}{dx}g(x) \right) h(x)$
체인룰 ^{Chain rule}	$\frac{d}{dx}g(h(x)) = \frac{d}{dh}g(h(x)) \frac{d}{dx}h(x)$

Higher Order Derivatives

■ The second derivative quantifies

the rate of change of the rate of change of $f(x)$.

■ For example, in physics,

- If a function represents an object's displacement
- The first derivative represents velocity, which is the rate of change of position.
- The second derivative represents acceleration, which is the rate of change of velocity.

■ The n -th derivative of $f(x)$ is expressed as follows:

$$f^{(n)}(x) = \frac{d^n f}{dx^n} = \left(\frac{d}{dx} \right)^n f(x)$$

Partial Derivatives

■ A function where multiple variables make up the domain ($f: \mathbb{R}^n \rightarrow \mathbb{R}$)

- A function $y = f(\mathbf{x}) = f(x_1, x_2, \dots, x_n)$ with n variables is called a **multivariable function**.
 - Input: n -dimensional vector: $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_n]^T$
 - Output: is a scalar y .
- The **partial derivative** of y with respect to the i -th parameter x_i is given by:

$$\frac{\partial y}{\partial x_i} = \lim_{h \rightarrow 0} \frac{f(x_1, x_2, \dots, x_i + h, \dots, x_n) - f(x_1, x_2, \dots, x_i, \dots, x_n)}{h}$$

참고: ∂y 는 partial y 라고 읽는다.

- To compute $\frac{\partial y}{\partial x_i}$, treat $x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n$ as constants and differentiate y only respect to x_i

- All same representation in partial derivative: $\frac{\partial y}{\partial x_i} = \frac{\partial f}{\partial x_i} = \frac{\partial}{\partial x_i} f(\mathbf{x}) = f_{x_i} = D_i f$

Gradient

■ Gradient

- The **gradient** of a multivariable function $f(x)$ with respect to an n -dimensional input vector

$$\mathbf{x} = [x_1 \ x_2 \ \dots \ x_n]^T$$

is defined as a vector composed of partial derivatives, as follows:

$$\nabla f(x) = \left[\frac{\partial f(X)}{\partial x_1} \quad \frac{\partial f(X)}{\partial x_2} \quad \dots \quad \frac{\partial f(X)}{\partial x_n} \right]^T$$

Gradient는 그리스 문자 ∇ 로 표기하고 "nabla (나블라)"로 발음한다.

$\nabla f(x)$ is referred to as "the gradient of f with respect to the vector X ."

Optimization in ML

- **Objectives such as minimizing the difference**

- between predicted and actual values

(or maximizing classification accuracy)

- **The function is called the **Objective Function**.**

- **Through the optimization process,**

- ML model searches for parameter values that achieve the desired objective function.

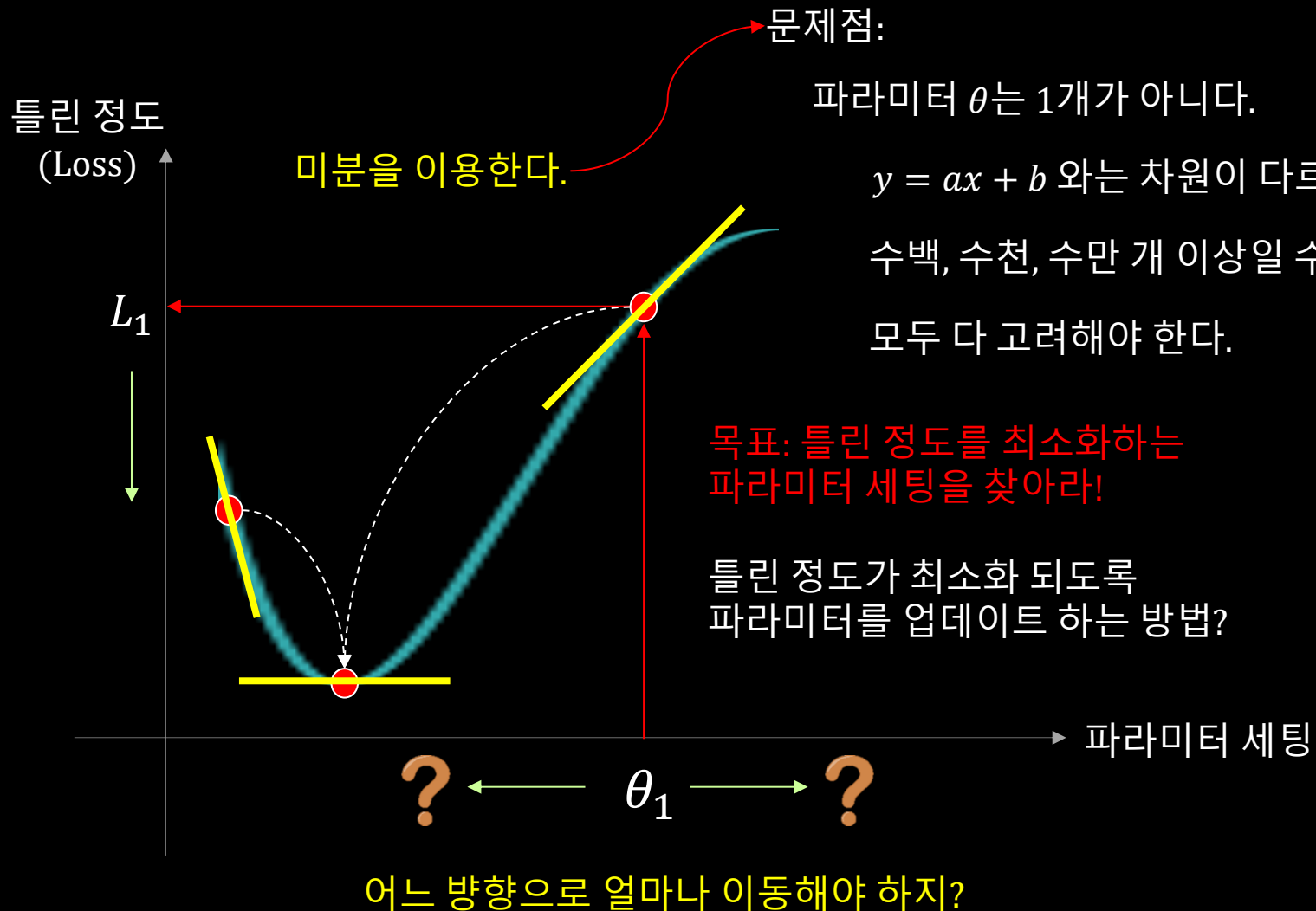
- **In minimization problems,**

- the objective function is referred to as **Loss Function**, **Cost Function**, or **Error Function**.

- **Goal of ML**

- find the model parameters
 - that achieve the **optimal objective function value**
 - based on the given training data.

Recap: 미분!!! 딥러닝 어디에 사용하는가?



Global Minimum vs. Local Minimum

■ Finding the parameter values that achieve the optimal objective function value → **extremely challenging**.

- Global Minimum:

- The lowest value of the objective function $f(x)$ across the entire range of x .

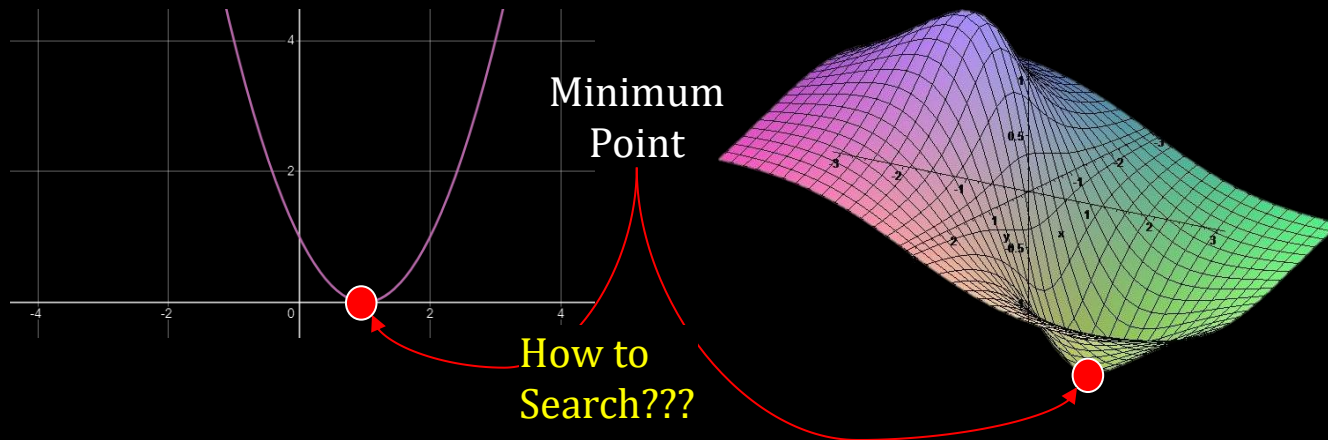
- Local Minimum:

- A point where the objective function $f(x)$ has a smaller value than at other nearby points within a small range of x .

Minimum Points in ML

■ Machine learning objective functions

- Often have multiple local minima.
- During the optimization process, gradient-based methods are used to find the minimum value of the loss function.
- However, once the model parameters reach a local minimum, it becomes difficult to explore other parameter values to find a potentially better global minimum.



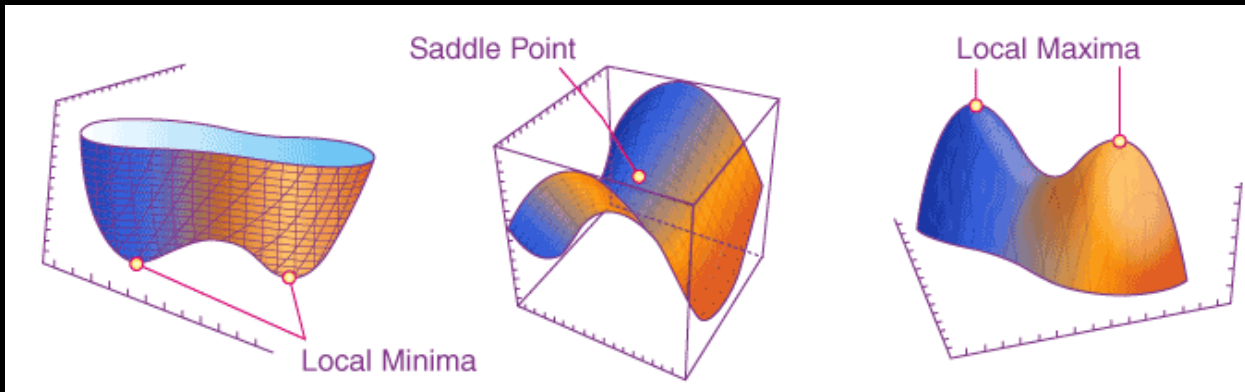
Critical Point (임계점)

- A critical point of a differentiable function $f(x)$ of one variable is a point where its derivative is zero:

$$\frac{d}{dx}f(x) = 0$$

- Types of Critical Points

- **Minimum:** A point where the derivative changes from **negative** to **positive**.
- **Maximum:** A point where the derivative changes from **positive** to **negative**.
- **Saddle Point**
 - Appears as a **maximum** in one direction but a **minimum** in another direction.

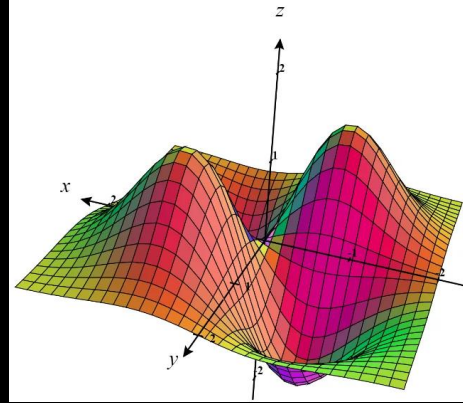


Example

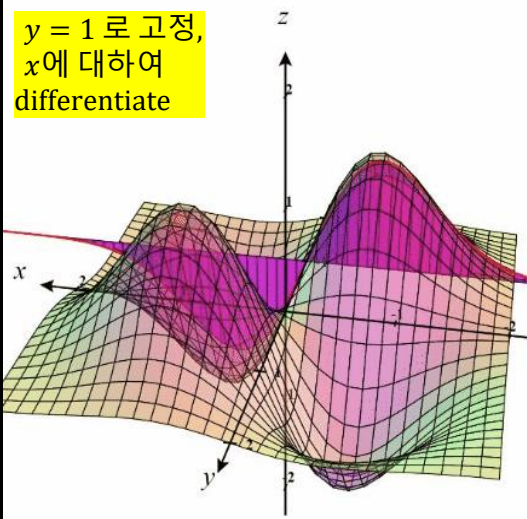
$$f(x, y) = \frac{7xy}{e^{x^2+y^2}}$$

, where $f: \mathbb{R}^2 \rightarrow \mathbb{R}$

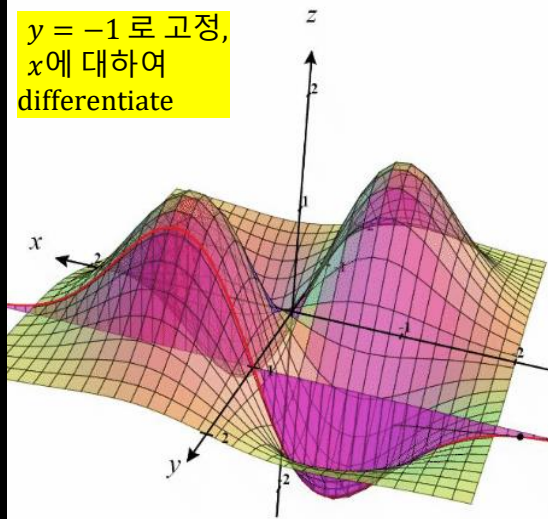
Let $z = f(x, y)$



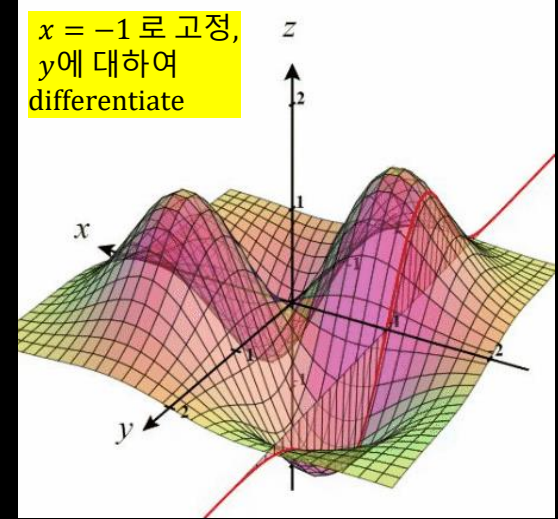
$y = 1$ 로 고정,
 x 에 대하여
differentiate



$y = -1$ 로 고정,
 x 에 대하여
differentiate



$x = -1$ 로 고정,
 y 에 대하여
differentiate



모든 y 에 대한 변화는?

모든 x 에 대한 변화는?

Probability

Random Variable

- A random variable is a function that maps a value from the sample space (domain) to a real number (codomain)

- Example:

When rolling a fair die, if we define the outcome as a random variable X , then:

- Sample space: $S = \{ 1, 2, 3, 4, 5, 6 \}$
- The event where the die lands on 5 is written as:
 - $\{ X = 5 \}$ or simply $X = 5$
- The probability of this event is expressed as:
 - $P(\{X = 5\})$ or simply $P(X = 5)$

Probability Distribution

- A probability distribution is a function that assigns probabilities to all possible values X that a random variable X can take:

$$P(X = x)$$

For simplicity, it can also be written as $P(X)$ or $P(x)$.

- The notation

$$X \sim P_X(x)$$

indicates that the random variable X follows
the probability distribution $P_X(x)$.

Types of Random Variables

■ Discrete Random Variables

- A random variable is discrete if its possible values are finite or countably infinite.
- Examples:
 - The possible outcomes when flipping a coin once.
 - The number of times a die is rolled until a "2" appears.

■ Continuous Random Variables

- A random variable is **continuous** if its possible values are **uncountable** and can take an **infinite number of values**.
- Examples:
 - Measuring a person's **height** without rounding.
 - Measuring a person's **weight** with unlimited decimal precision.

Axioms of Probability

■ **An axiom is a statement that is accepted as true without proof.**

■ **Notations**

- **S: sample space** (the set of all possible outcomes of an experiment)
- $P(A)$: the probability of an event A occurring

■ **The probability function $P(\cdot)$ must satisfy the following axioms:**

- Non-Negativity
 - $P(A) \geq 0$, for all events $A \subseteq S$ (Probabilities are always non-negative real numbers.)
- Normalization
 - $P(S)=1$ (The probability of the entire sample space is always 1.)
- Additivity
 - $P(A_1 \cup A_2) = P(A_1) + P(A_2)$ for any mutually exclusive events A_1 and A_2 meaning
 - $A_1 \cap A_2 = \emptyset$

Probability Mass Function (PMF) & Probability Density Function (PDF)

■ Probability Mass Function (PMF)

- A function that represents the probability of **discrete** random variables.
- Example: If X represents the sum of two dice rolls, then:

$$P(X = x), x \in \{2, 3, \dots, 12\}$$

■ Probability Density Function (PDF)

- A function that represents the probability of **continuous** random variables.

[Example]

- The probability that the height of a man in his 20s falls between **168 cm and 175 cm**.
- For continuous random variables, the probability that the variable falls within an interval $[a, b]$ is calculated as the integral of the PDF over that range.

$$P(X \in [a, b]) = \int_a^b P_X(x) dx$$

Multivariate Random Variable

- A multivariate random variable is a **list of multiple random variables**.
- When expressed in vector form, it is called a **random vector**.

$$X = [X_1, X_2, \dots, X_n]^T$$

Bayes' Theorem

- The multiplication rule for joint distributions is used.

The diagram illustrates Bayes' Theorem with the following components and labels:

- Likelihood (우도, 가능도)**: Points to the term $P(B|A)$ in the numerator.
- Prior probability (사전 확률)**: Points to the term $P(A)$ in the numerator.
- Posterior probability (사후 확률)**: Points to the term $P(A|B)$ on the left side of the equation.
- Evidence (증거) / Marginal likelihood**: Points to the term $P(B)$ in the denominator.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

What happens in Deep Learning?

■ 딥러닝에서 발생한 상황,

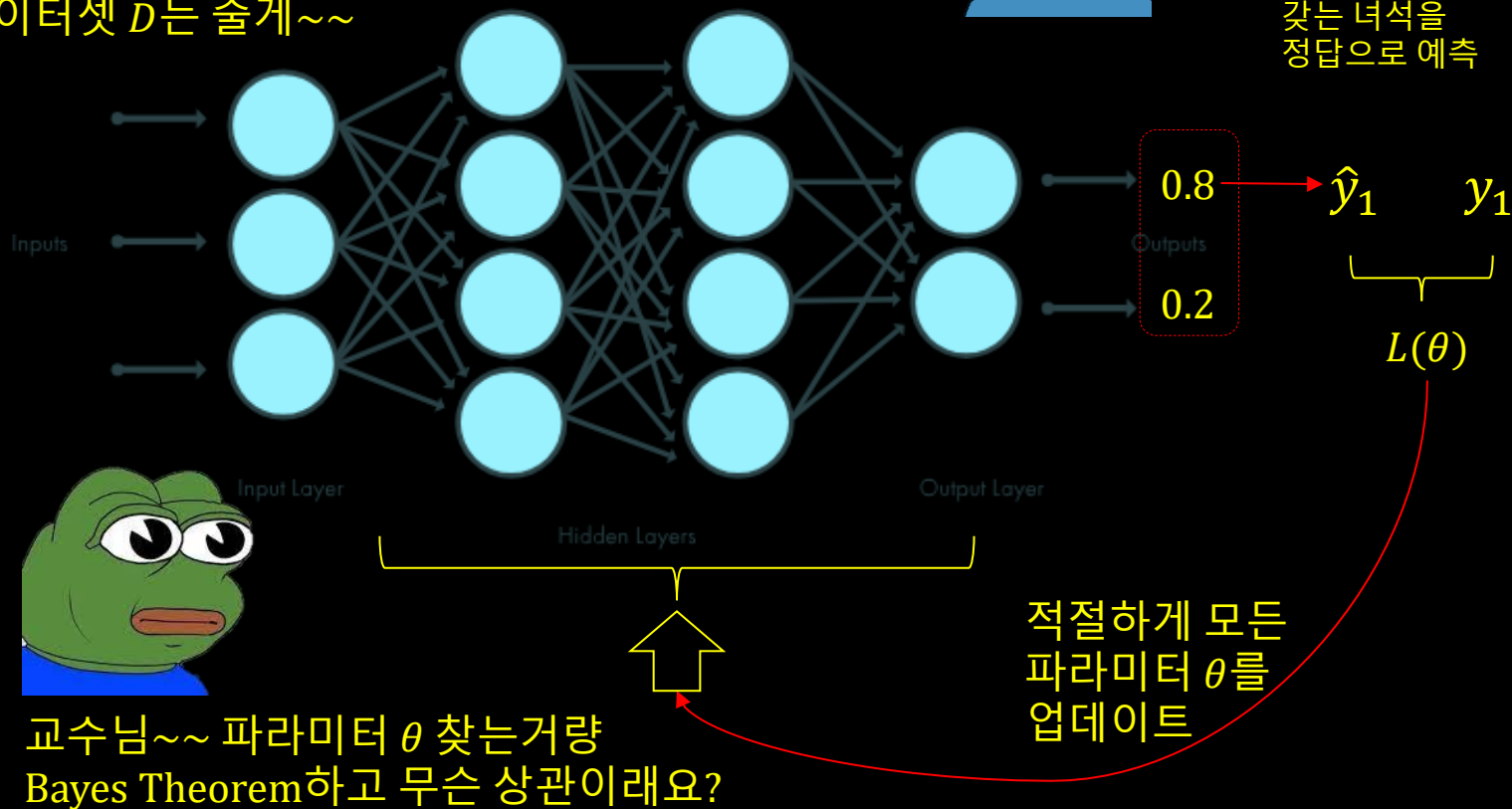
정답과 가장 가까운 출력을 생성하는
파라미터 집합 θ 를 찾아라!

아... 데이터셋 D 는 줄게~~

$y = ax + b$??
무작정 파라미터
 $\theta = \{a, b\}$ 찾으라고??

뭐래! 장난하냐?

가장 큰 확률값
갖는 녀석을
정답으로 예측



Where is Bayes Theorem in Deep Learning?

■ 앞에서 설명한 상황을 Bayes Theorem으로 생각하면?

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

$$\text{Dataset } D = \{ (x_i, y_i) \}_{i=1}^n$$

bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	...	grade	sqft_above	sqft_basement	price	type
3	1.00	1180	5650	1.0	0	0	...	7	1180	0	221900.0	아파트
3	2.25	2570	7242	2.0	0	0	...	7	2170	400	538000.0	빌라
2	1.00	770	10000	1.0	0	0	...	6	770	0	180000.0	전원주택
4	3.00	1960	5000	1.0	0	0	...	7	1050	910	604000.0	다가구
3	2.00	1680	8080	1.0	0	0	...	8	1680	0	510000.0	아파트

$$P(y|x) = P(\text{정답}|\text{데이터})$$

데이터가 주어졌을 때
정답 맞출 확률을 최대화

Dataset D 주어졌을 때 정답 맞출
확률을 최대화하는 파라미터 θ

$$P(\theta|D)$$

Bayes Theorem 적용

$$P(\theta|D) = \frac{P(D|\theta) \times P(\theta)}{P(D)}$$

Expected Value

- When a random variable X follows a probability distribution $P(X)$,
denoted as

$$X \sim P(X)$$

- For a Discrete Random Variable

$$E_{X \sim P(X)}[f(X)] = \sum_x P(x)f(x)$$

- For a Continuous Random Variable

$$E_{X \sim P(X)}[f(X)] = \int_{-\infty}^{\infty} P(x)f(x) dx$$

Variance

- **Given a random variable X that follows a probability distribution $P(X)$,**
 - the **variance** of a function $f(X)$ measures how much the values of $f(X)$ deviate from their expected value $E[f(X)]$.

- **Definition**

$$\text{Var}(f(X)) = E[(f(X) - E[f(X)])^2]$$

- **Alternative Formula**

$$\text{Var}(f(X)) = E[(f(X)^2] - (E[f(X)])^2$$



수고하셨습니다 ..^^..
Thank you!