# 머신 러닝·딥러닝에 필요한 수학 기초 with 파이썬

## Differentiations

### 미분의 개념

조준우
metamath@gmail.com

# 미분

- **미분의 중요성** :
  - 머신러닝을 최적화 과정이라 볼 때 다양한 최적화 알고리즘의 핵심 개념
  - 선형 회귀에서도 중요한 개념

- **알아보는 것**
  - 미분의 개념
  - 극한의 개념

# Sympy

- **Sympy** is a Python library for symbolic mathematics. It aims to become a full-featured computer algebra system (CAS) while keeping the code as simple as possible in order to be comprehensible and easily extensible. SymPy is written entirely in Python.
  [http://www.sympy.org](http://www.sympy.org)

- **용도**
  - **극한과 미분 예제의 확인을 위해 사용**
  - 실제 계산 보다는 개념과 코드 구현을 중심으로 리뷰
  - 계산은 sympy에게……

- **사용법**

```
x = sympy.Symbol('x')
sympy.limit((x**2-1)/(x-1), x, 1)
sympy.diff(sympy.E**(2*x), x)
```

jupyter Differentiation.ipynb

# 함수의 변화

- 함수의 입력이 변할 때 출력의 변화를 생각

$$\text{sensitivity} = \frac{\Delta \text{output}}{\Delta \text{input}}$$

출력의 변화

입력의 변화

입력의 단위 변화에 따른 출력 변화의 민감도

# 의문 1

- ? : 입력에 대한 출력의 변화가 왜 중요?

# 생활에서 느끼는 함수의 변화

- 주식투자를 할 때
  - 입력 : 투자 이후 시간
  - 출력 : 이익금
  - 단위시간당 이익이 얼마나 늘어나는가?
  - $f: t \rightarrow \$$

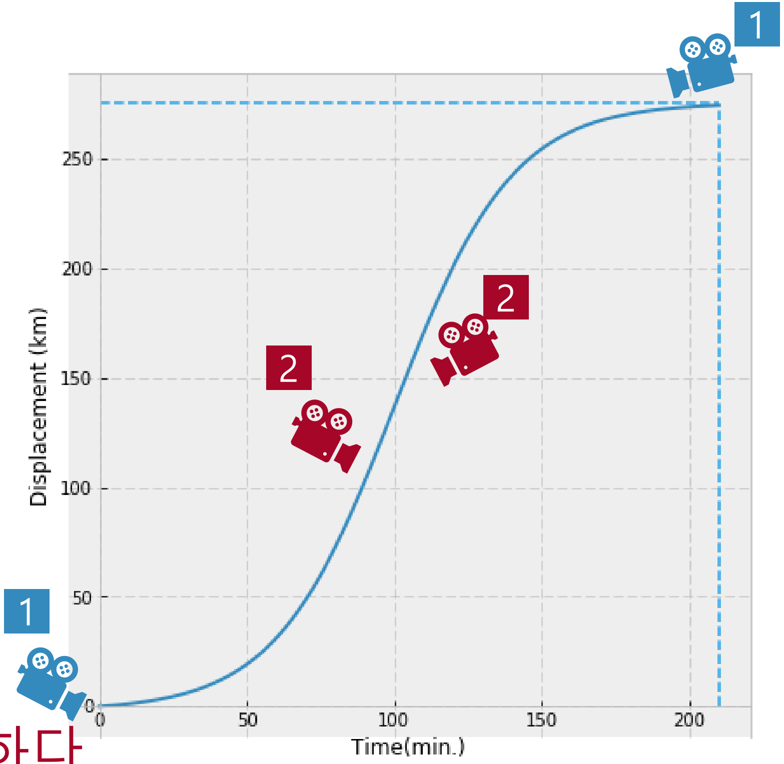
- 목욕을 할 때
  - 입력 : 수전의 회전각도
  - 출력 : 나오는 물의 온도
  - 수전을 얼마나 돌려야 물이 적당한 온도가 되는가?
  - $f: angle \rightarrow temp.$

# 의문 2

- ? : 얼마간 입력에 대한 출력의 변화를 관찰

# 구간 속도위반 단속

- 대구수성IC - 동서울TG
  - 1번 구간
    - 평균속도 : 78.35km/h
    - 단속 안됨
  - 2번 구간
    - 평균속도 : 203.67km/h
    - 단속
  - 함수의 변화를 보는 구간이 중요하다



# 평균변화율

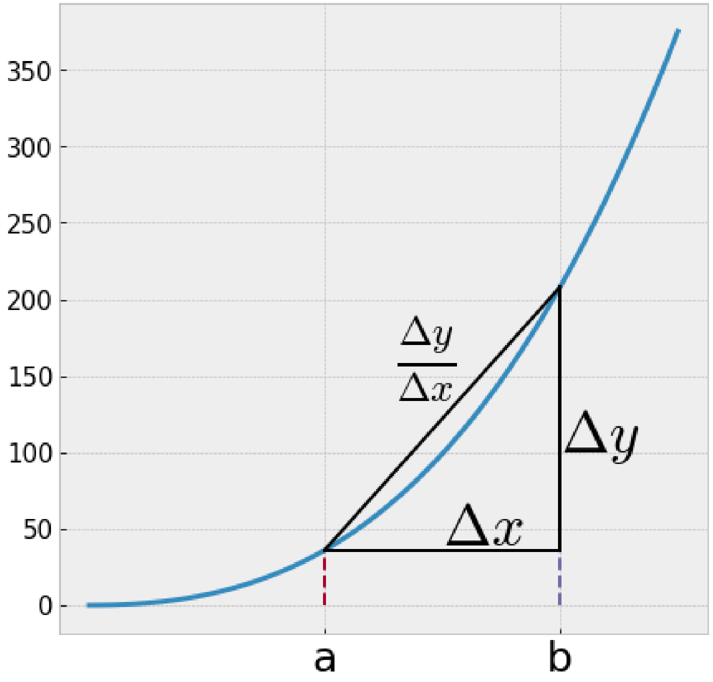
- **정의** : $x$의 증분 $\Delta x$에 대한 $y$의 증분 $\Delta y$ 의 비율

$$\frac{\Delta y}{\Delta x} = \frac{f(b) - f(a)}{b - a} = \frac{f(a + \Delta x) - f(a)}{\Delta x}$$

- 두 점 사이의 곡선의 형태는 평균변화율에 영향을 미치지 않음.

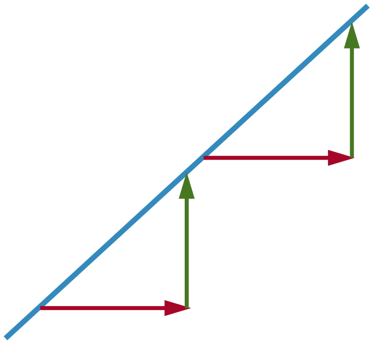
- 좀 더 자세한 정보를 위해 두 점 사이의 간격을 좁힐 필요가 있음.

- a-b의 간격을 충분히 가깝게 하여 한 순간으로 만들면 가장 자세한 정보를 얻을 수 있을 것이다.

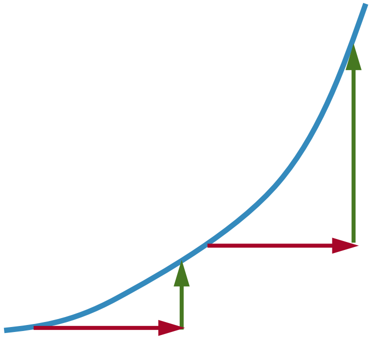


# 충분히 작은 구간?

- **직선인 경우** : 어느 지점이나 경사가 똑같다.
  - 구간을 좁힌 정보나 구간을 넓힌 평균 정보나 정보가 똑같다.
- **곡선인 경우** : 모든 지점에서 경사의 가파른 정도가 다 다르다.
  - 모든 지점에서 언덕의 가파른 정도를 알고 싶다.



- 가파른 정도를 계산하기 위해 우선 간격을 0.01 정도로 정해보자.



- 0.01이면 충분히 작은 간격이어서 순간의 변화를 계산해 줄 만하다.

$$df(x) = \frac{f(x + 0.01) - f(x)}{0.01}$$

# 0.01은 충분한가?

- $f(x) = x^2$ :
  - 위 함수의 순간 변화율은 해석적으로 $2x$라는 것이 알려져 있다고 가정

- $x = 3$에서 순간변화율은?

$$\frac{d}{dx}f(3) = 2(3) = 6$$

$$df(3) = \frac{f(3+0.01) - f(3)}{0.01} = \frac{3.01^2 - 3^2}{0.01} = 6.01$$

jupyter Differentiation.ipynb

- 문제 : 0.01 방법은 순간변화율이 아니다.

# 의문 3

- ? : 0.01도 부족한데 얼마나 더 작게 해야? 정말 움직이지 않는 한 '순간'에 대한 변화를 생각하는 것은 모순이 아닐까?

# 극한Limits

- **극한** : 함수에서 정의할 수 있는 함수 값과는 다른 또 다른 값
- 함수 $f(x)$에서 $x$의 값이 $a$와 다른 값을 가지면서 $a$에 가까이 갈 때 $f(x)$의 값이 일정한 값 $L$에 가까워지면

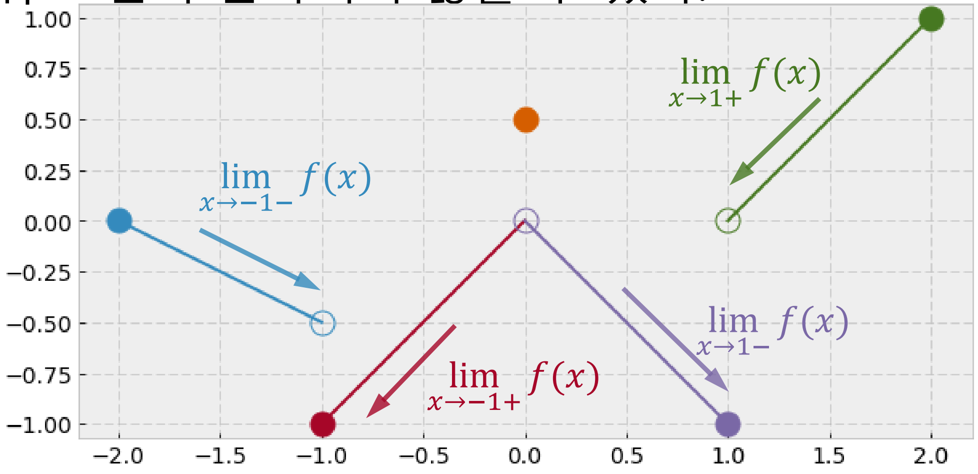
극한 기호 : 리미트 $\lim_{x \to a} f(x) = L$

- 이때 $L$을 $x \to a$ 일 때 함수 $f(x)$ 의 극한 값 또는 극한
- 특수한 경우 함수 값과 극한 값은 같고 이를 연속이라 함.
- 하지만 일반적으로는 $\lim_{x \to a} f(x) \neq f(a)$
- 중요한 것은 $x$가 결코 $a$가 되지 않는다는 점, $x = a$가 된다면 그 값은 함수 값

# 극한Limits

- **극한의 존재성** : $x \to a$일 때 함수 $f(x)$가 $L$로 수렴하면 $x \to a+$일 때 우극한과 $x \to a-$일 때 좌극한이 모두 존재하고 그 값이 $L$로 같다. 또한 그 역도 성립

- 경우에 따라 위 조건이 만족되지 않을 수 있다.



# 분수함수의 극한

- 분수함수의 경우 함수 값이 정의 되지 않을 수 있지만 극한은 존재할 수 있다.

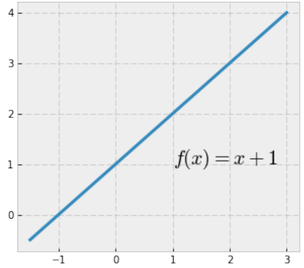
$$f(x) = \frac{x^2 - 1}{x - 1}$$

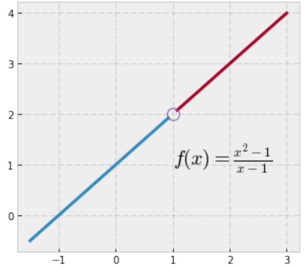
- $f(1)$은 정의조차 되지 않음

- 이 경우 $x = 1$ 근처에서 어떤 일이 일어날까?

$$\lim_{x \to 1} \frac{x^2 - 1}{x - 1} = 2 \neq f(1)$$

```
x = sympy.Symbol('x')
sympy.limit((x**2-1)/(x-1), x, 1)
2
```



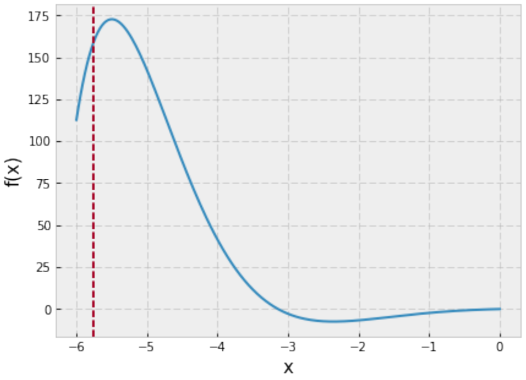

| | f(x) | x |
|---|---|---|
| 0 | 1.900000 | 0.900000 |
| 1 | 2.100000 | 1.100000 |
| 2 | 1.990000 | 0.990000 |
| 3 | 2.010000 | 1.010000 |
| 4 | 1.999000 | 0.999000 |
| 5 | 2.001000 | 1.001000 |
| 6 | 1.999999 | 0.999999 |
| 7 | 2.000001 | 1.000001 |

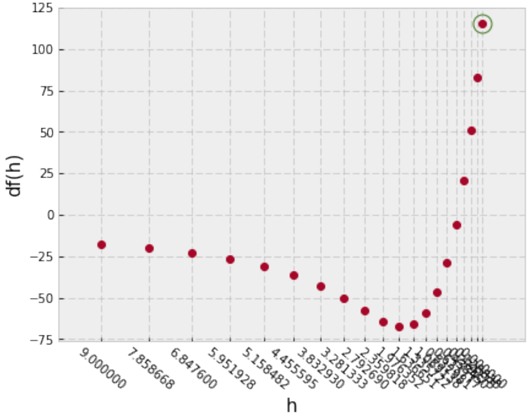
# 0.01을 넘어...

- 매우 작은 변화 구간 0.01을 극한의 개념으로 대체

$$\frac{f(x+0.01)-f(x)}{0.01}, \frac{f(x+0.001)-f(x)}{0.001}, \frac{f(x+0.0001)-f(x)}{0.0001}, \ldots$$

- 빨간 글자를 변수로 두면 $h$의 함수 $\quad df(h) = \dfrac{f(x+h)-f(x)}{h}$



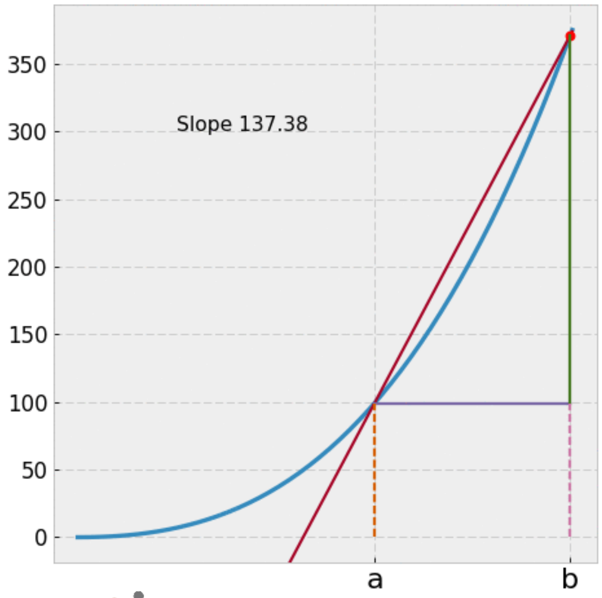
# 순간변화율The Derivative at a Point

- **정의** : 평균변화율의 분모를 순간에 이를 정도로 작게 만들어

$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h} = \lim_{x \to a} \frac{f(x) - f(a)}{x - a}$$

- 극한이 존재하면 이 극한값을 순간변화율 또는 미분계수라 함

- 순간변화율은 그 위치에서 접선의 기울기로 해석 가능




jupyter

Differentiation.ipynb

# 도(導)함수The Derivative as a Function

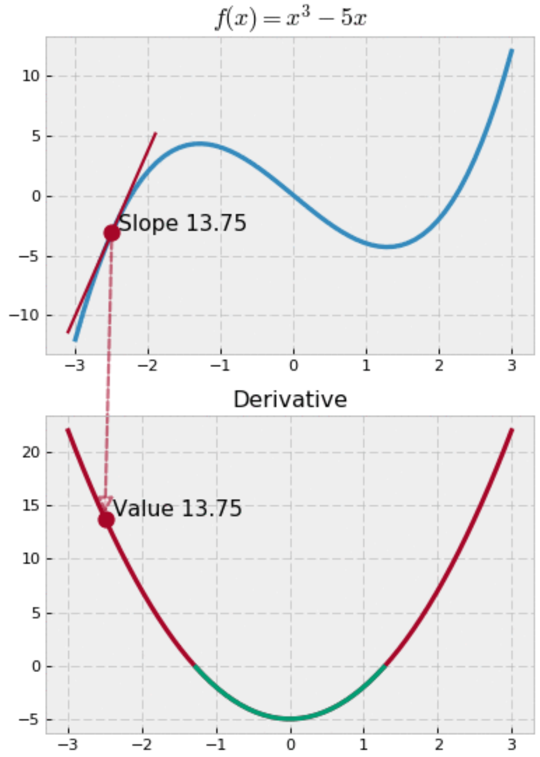
- **정의** : 미분계수를 함수 값으로 가지는 함수

- 미분 가능한 함수 $y = f(x)$의 도함수

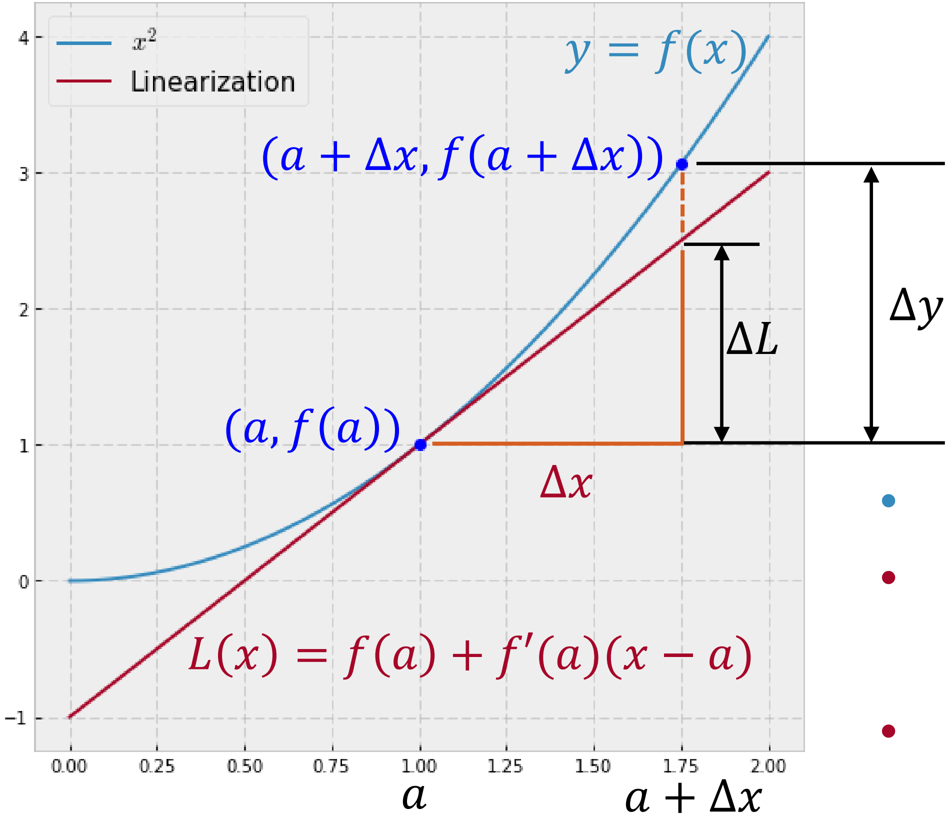
$$f'(x) = \frac{df(x)}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

- 그림에서 위 함수의 미분계수(slope)를 $x$에 대한 함수 값으로 하여 그림을 다시 그리면 아래 그림

- 아래 그림이 위 그림의 도함수



# 선형 근사와 함숫값 추정




$\Delta y = f(a + \Delta x) - f(a)$

- $y$ : 계산 어려움
- $L$ : 계산 쉬움

- $\Delta x$가 적당히 작다면 $\Delta y \approx \Delta L$

# 용어, 기호Notations

- $y = f(x)$ 일 때
  - **Leibniz's notation** : $\frac{dy}{dx} = \frac{df}{dx} = \frac{d}{dx}f(x)$
  - **Lagrange's notation** : $f'(x)$, $y'$
  - **Newton's notation** : $\dot{y}$, $\ddot{y}$
  - **Euler's notation** : $D_x y$, $D_x f(x)$
- **미분하다differentiation ≠ derivative도함수,미분계수**
  - The process of calculating a derivative is called differentiation.
  - If $f'$ exists at a particular $x$, we say that $f$ is differentiable (has a derivative) at $x$. If $f'$ exists at every point in the domain of $f$, we call $f$ differentiable.

# 다항함수의 도함수

- $y = f(x) = x^n$ 일 때 $\dfrac{d\,x^n}{dx} = nx^{n-1}$

$$a^3 - b^3 = (a-b)(a^2 + ab + b^2)$$

$$a^3 - b^3 = (a-b)(a^2b^0 + a^1b^1 + a^0b^2)$$

$$\frac{d}{dx}f(x) = \lim_{h\to 0}\frac{f(x+h)-f(x)}{h} = \lim_{h\to 0}\frac{(x+h)^n - x^n}{h}$$

$$= \lim_{h\to 0}\frac{\{(x+h)-x\}\{(x+h)^{n-1}x^0 + (x+h)^{n-2}x^1 + \cdots + (x+h)^1x^{n-2} + (x+h)^0x^{n-1}\}}{h}$$

$$= \lim_{h\to 0}(x+h)^{n-1}x^0 + (x+h)^{n-2}x^1 + \cdots + (x+h)^1x^{n-2} + (x+h)^0x^{n-1}$$

$$= \underbrace{x^{n-1} + x^{n-1} + \cdots + x^{n-1}}_{n}$$

$$= nx^{n-1}$$

# 다항함수의 도함수 예

- $y = f(x) = x^n$ 일 때 $\dfrac{d\,x^n}{dx} = nx^{n-1}$

- 예제 $y = x^2$
  - 정의에 의해

$$\lim_{h \to 0} \frac{(x+h)^2 - x^2}{h} = \lim_{h \to 0} \frac{x^2 + 2hx + h^2 - x^2}{h} = \lim_{h \to 0} \frac{h(2x+h)}{h} = 2x$$

  - 공식으로

$$\frac{dy}{dx} = 2x^{2-1} = 2x$$

```
x = sympy.Symbol('x')
sympy.diff(x**2, x)
```

$2x$

# 지수, 로그함수의 도함수

- 지수함수
  - $y = f(x) = \mathrm{e}^x$ 일 때 $\frac{d}{dx}\mathrm{e}^x = \mathrm{e}^x$
  - $y = f(x) = a^x$ 일 때 $\frac{d}{dx}a^x = a^x \ln a$
- 로그함수
  - $y = \ln(x)$ 일 때 $\frac{d}{dx}\ln(x) = \frac{1}{x}$
  - $y = \log_a(x)$ 일 때 $\frac{d}{dx}\log_a(x) = \frac{1}{x\ln a}$

# 지수함수의 도함수

- $\dfrac{d}{dx}\mathrm{e}^x = \mathrm{e}^x$

$$\frac{d}{dx}e^x = \lim_{h\to 0}\frac{e^{x+h}-e^x}{h} = \lim_{h\to 0}\frac{e^x e^h - e^x}{h}$$

$$= e^x \lim_{h\to 0}\frac{e^h-1}{h} = e^x$$

- $\dfrac{d}{dx}a^x = a^x \ln a$

$$\frac{d}{dx}a^x = \lim_{h\to 0}\frac{a^{x+h}-a^x}{h} = \lim_{h\to 0}\frac{a^x a^h - a^x}{h}$$

$$= a^x \lim_{h\to 0}\frac{a^h-1}{h}$$

$$= a^x \ln a$$

jupyter Differentiation.ipynb

# 로그함수의 도함수

- $\frac{d}{dx}\ln(x) = \frac{1}{x}$

$$\frac{d}{dx}\ln(x) = \lim_{h\to 0}\frac{\ln(x+h)-\ln(x)}{h} = \lim_{h\to 0}\frac{\ln\left(\frac{x+h}{x}\right)}{h} = \lim_{h\to 0}\frac{\ln\left(1+\frac{h}{x}\right)}{h}$$

$$= \lim_{h\to 0}\ln\left(1+\frac{h}{x}\right)^{1/h} = \lim_{z\to 0}\ln(1+z)^{\frac{1}{z}\frac{1}{x}} \qquad \because z = \frac{h}{x}$$

$$= \frac{1}{x}\lim_{z\to 0}\ln(1+z)^{\frac{1}{z}} = \frac{1}{x}\ln(e) = \frac{1}{x}$$

- $\frac{d}{dx}\log_a(x) = \frac{1}{x\ln a}$

$$\frac{d}{dx}\log_a x = \frac{d}{dx}\frac{\ln x}{\ln a} = \frac{1}{\ln a}\frac{d}{dx}\ln x = \frac{1}{x\ln a}$$

$$\log_a b = \frac{\log_c b}{\log_c a}$$

# 정리

| 분류 | 함수 | 도함수 |
|---|---|---|
| 다항함수 | $y = x^n$ | $\frac{d}{dx}x^n = nx^{n-1}$ |
| 지수함수 | $y = \mathrm{e}^x$ | $\frac{d}{dx}\mathrm{e}^x = \mathrm{e}^x$ |
| | $y = a^x$ | $\frac{d}{dx}a^x = a^x \mathrm{ln}a$ |
| 로그함수 | $y = \ln(x)$ | $\frac{d}{dx}\ln(x) = \frac{1}{x}$ |
| | $y = \log_a(x)$ | $\frac{d}{dx}\log_a(x) = \frac{1}{x\mathrm{ln}a}$ |

# 몇 가지 미분법

- **상수 미분**
  - $\frac{dc}{dx} = 0$
- **덧셈, 뺄셈의 미분**
  - $y = f(x) \pm g(x)$ $\quad y' = f'(x) \pm g'(x)$
- **곱셈의 미분법**
  - $y = f(x)g(x)$ $\quad y' = f'(x)g(x) + f(x)g'(x)$
- **나눗셈의 미분법**
  - $y = \frac{f(x)}{g(x)}$ $\quad y' = \frac{f'(x)g(x) - f(x)g'(x)}{\{g(x)\}^2}$

# 몇 가지 미분법:곱셈

- **곱셈의 미분법**
  - $y = f(x)g(x)$ $\quad y' = f'(x)g(x) + f(x)g'(x)$

$$\lim_{h\to 0}\frac{f(x+h)-f(x)}{h}$$

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

$$= \lim_{h\to 0}\frac{g(x+h)\big(f(x+h)-f(x)\big)+f(x)g(x+h)-f(x)g(x)}{h}$$

$$= \lim_{h\to 0}\frac{g(x+h)\big(f(x+h)-f(x)\big)+f(x)(g(x+h)-g(x))}{h}$$

$$= \lim_{h\to 0}\frac{g(x+h)(f(x+h)-f(x))}{h} + \lim_{h\to 0}\frac{f(x)(g(x+h)-g(x))}{h}$$

$$= f'(x)g(x) + f(x)g'(x)$$

# 몇 가지 미분법:나눗셈

- **나눗셈의 미분법**
  - $\boldsymbol{y = \frac{f(x)}{g(x)}}$ $\quad$ $\boldsymbol{y' = \frac{f'(x)g(x) - f(x)g'(x)}{\{g(x)\}^2}}$

$$\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) = \lim_{h\to 0}\frac{\frac{f(x+h)}{g(x+h)} - \frac{f(x)}{g(x)}}{h} = \lim_{h\to 0}\frac{\frac{f(x+h)g(x) - f(x)g(x+h)}{g(x+h)g(x)}}{h}$$

$$= \lim_{h\to 0}\frac{f(x+h)g(x) - f(x)g(x+h)}{h}\frac{1}{g(x+h)g(x)}$$

$$= \lim_{h\to 0}\frac{g(x)\big(f(x+h) - f(x)\big) + g(x)f(x) - f(x)g(x+h)}{h}\frac{1}{g(x+h)g(x)}$$

$$= \lim_{h\to 0}\frac{g(x)\big(f(x+h) - f(x)\big) - f(x)(g(x+h) - g(x))}{h}\frac{1}{g(x+h)g(x)}$$

$$= (g(x)f'(x) - f(x)g'(x))\frac{1}{g^2(x)}$$

# 합성함수 미분법

- $y = (2x-4)^3$을 미분하려면?! 전개하자!

$$\frac{d}{dx}(2x-4)^3 = \frac{d}{dx}8x^3 - 48x^2 + 96x - 64 = 24x^2 - 96x + 96$$

- $y = (2x-4)^{100}$ 라면?
- 두 함수 $y = f(u)$, u $= g(x)$가 미분 가능할 때, 합성함수 $y = f(g(x))$의 $x$에 대한 미분

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

- $u = g(x) = 2x-4, \;\; y = f(u) = u^3$

$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx} = 3(2x-4)^2 \cdot 2$$

```python
x = sympy.Symbol('x')
sympy.diff((2*x-4)**3, x)
```

$6(2x-4)^2$

# 겹겹이 합성된 함수

- 매우 많은 함수가 다음처럼 겹겹이 합성 되어있다 하더라도

$$f(x) = g\left(h\left(i\left(j\left(k\left(l(m(n(\ldots o(x))))\right)\right)\right)\right)\right)$$

- 각 함수의 미분계수의 곱 = $f(x)$의 미분계수

$$\frac{d}{dx}f(x) = \frac{dg(h)}{dh}\frac{dh(i)}{di}\frac{di(j)}{dj}\frac{dj(k)}{dk}\frac{dk(l)}{dl}\frac{dl(m)}{dm}\frac{dm(n)}{dn}\cdots\frac{do(x)}{dx}$$

- 신경망 역전파 알고리즘의 기본 아이디어

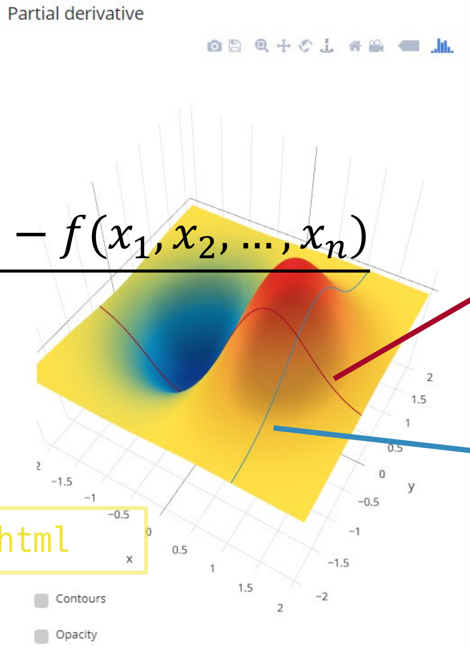
# 편도함수Partial Derivatives

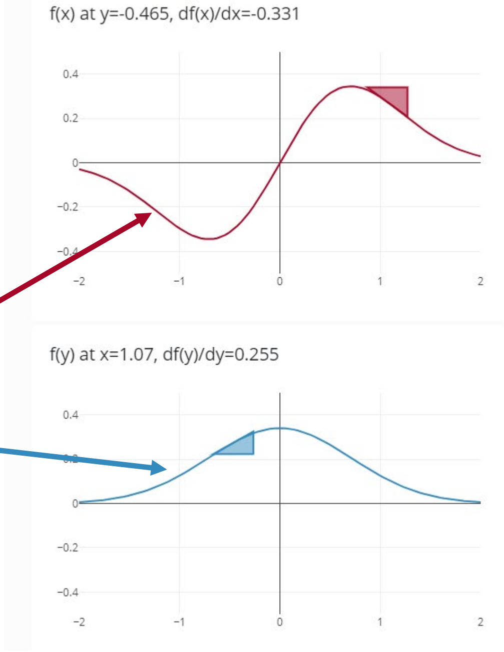
- $y = f(x_1, x_2, \dots, x_k, \dots, x_n)$ 인 다변수 함수일 때 하나의 변수에 대해서만 미분한 도함수

$$\frac{\partial f}{\partial x_k}(x_1, x_2, \dots, x_k, \dots, x_n)$$

$$= \lim_{h \to 0} \frac{f(x_1, x_2, \dots, x_k + h, \dots, x_n) - f(x_1, x_2, \dots, x_n)}{h}$$

편미분 기호 : 라운드




JS metamath1.github.io/noviceml/partial.html

# 편도함수Partial Derivatives

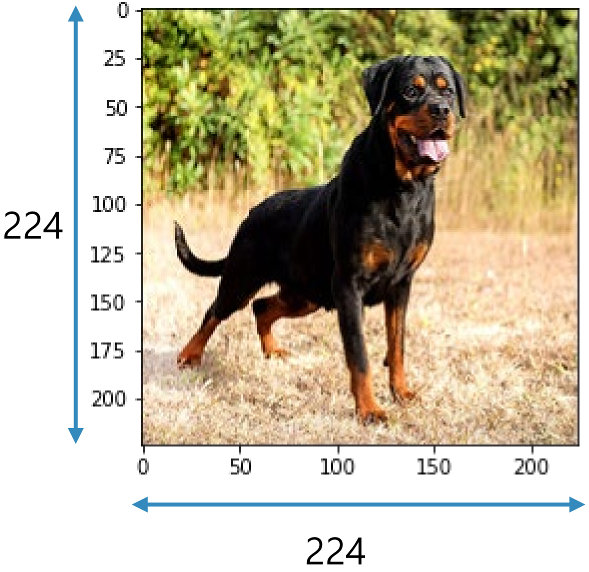
- 편미분 예
- (4,-2)에서 $f(x,y) = x^2 + 3xy + y - 3$에 대해 $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}$

$$\frac{\partial f}{\partial x} = 2x + 3y\Big|_{x=4,y=-2} = 2 \qquad \frac{\partial f}{\partial y} = 3x + 1\Big|_{x=4,y=-2} = 13$$

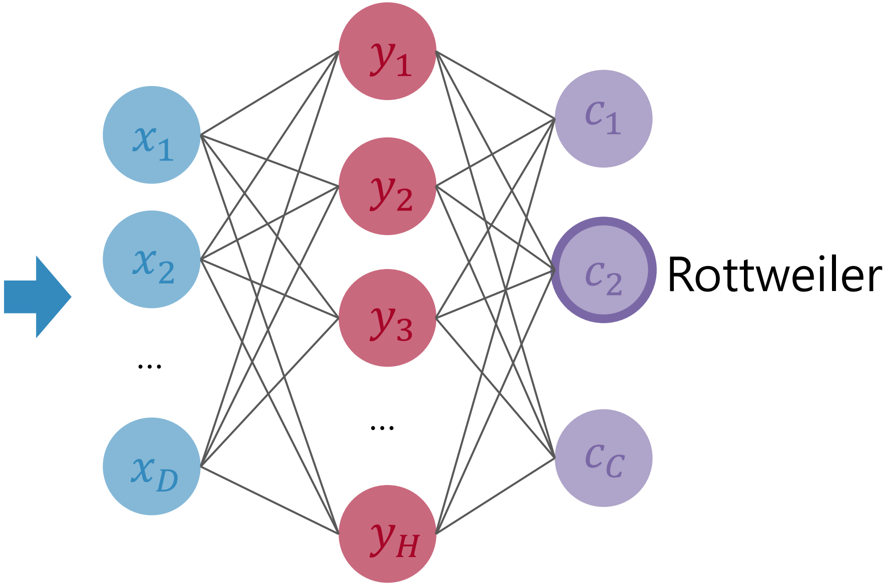
```
x = sympy.Symbol('x')
y = sympy.Symbol('y')
sympy.diff(x**2 + 3*x*y + y - 3, x)
sympy.diff(x**2 + 3*x*y + y - 3, y)
```

$2x + 3y, \ 3x + 1$

# Network Visualization



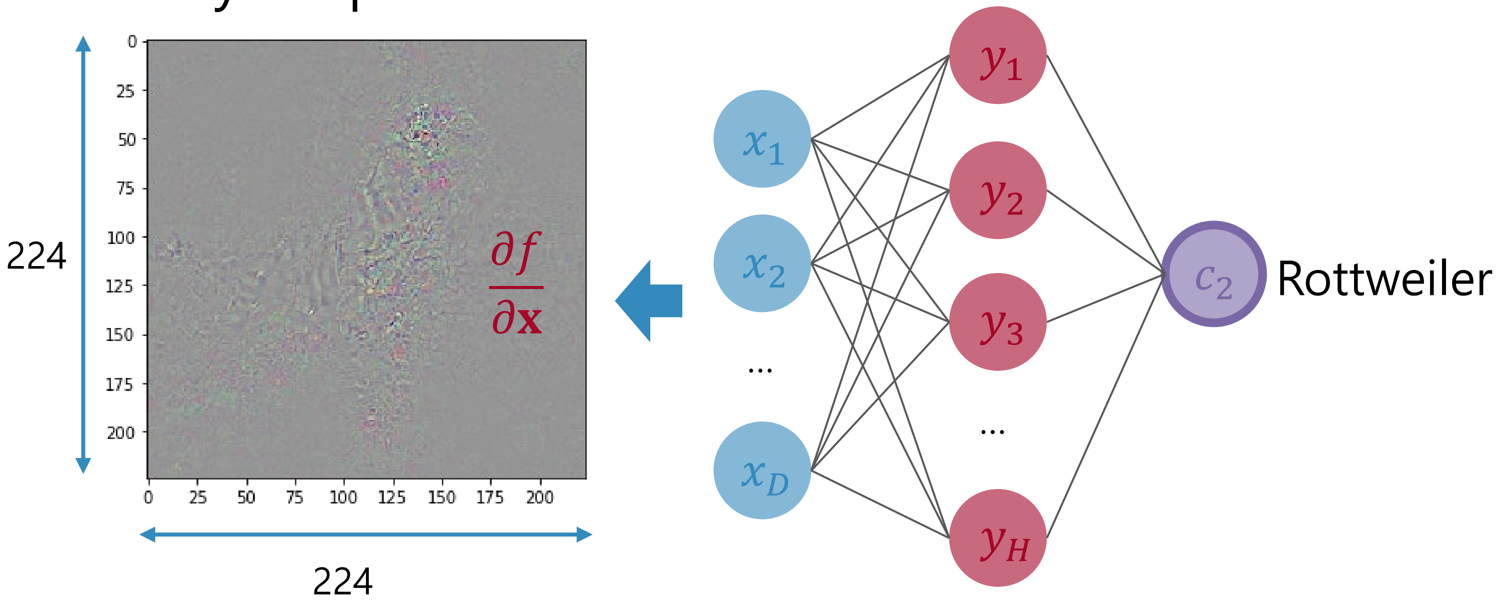

$\mathbf{x} \in \mathbb{R}^{224\times224}$




$f: \mathbb{R}^{224\times224} \rightarrow \mathbb{R}^C$

# Network Visualization

- Saliency Maps



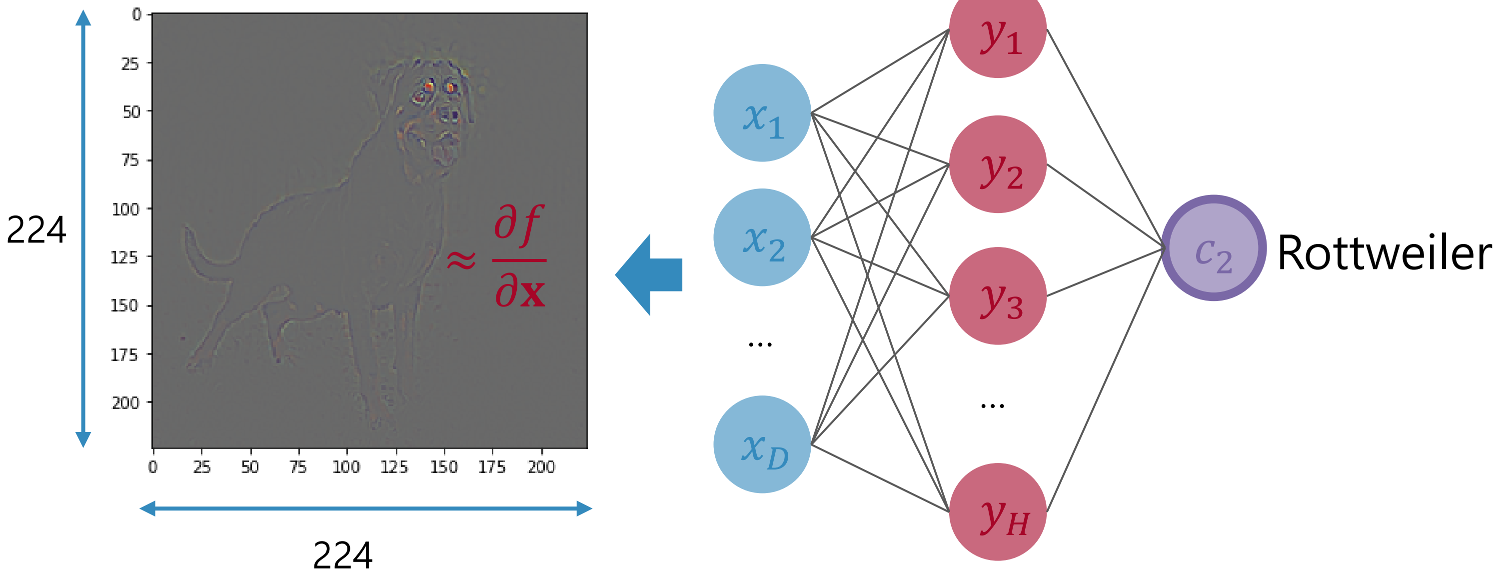

$\mathbf{x} \in \mathbb{R}^{224\times224}$

$f: \mathbb{R}^{224\times224} \rightarrow \mathbb{R}$

"Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps", Karen Simonyan et al, ICLR Workshop 2014.

# Network Visualization

- Guided Back Propagation




$\mathbf{x} \in \mathbb{R}^{224\times224}$

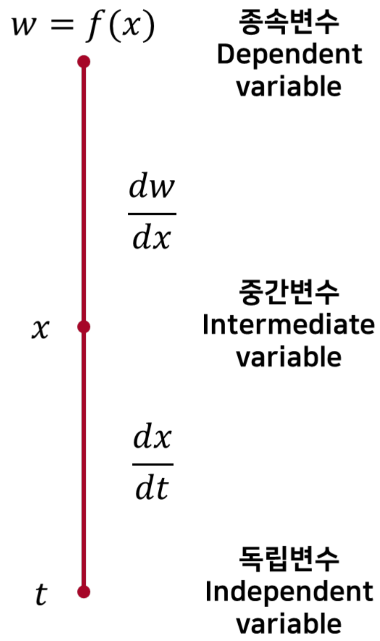
$f: \mathbb{R}^{224\times224} \rightarrow \mathbb{R}$

"Striving for Simplicity: The All Convolutional Net", J. T. Springenberg et. al., ICLR 2015

# 연쇄법칙Chain rule

- 일변수의 경우$x = g(t)$ 일 때 $w = f(x)$의 $t$에 대한 변화

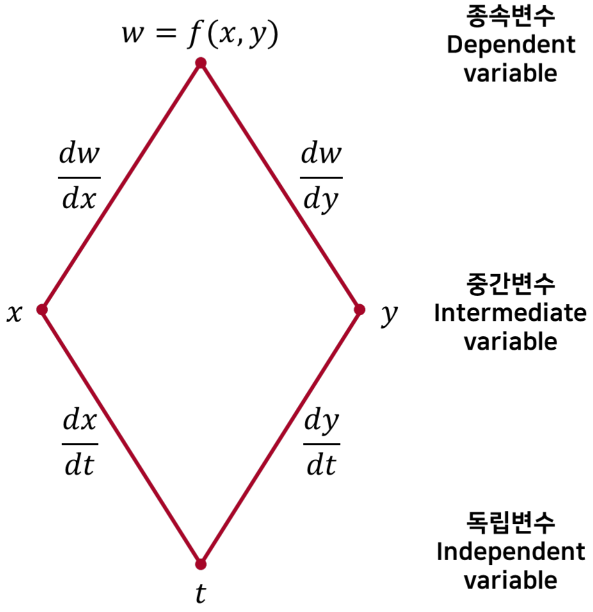
$$\frac{dw}{dt} = \frac{dw}{dx}\frac{dx}{dt}$$



# 연쇄법칙Chain rule

- $y = f(t) \quad x = g(t)$ 일 때 $w = f(x, y)$의 $t$ 에 대한 변화

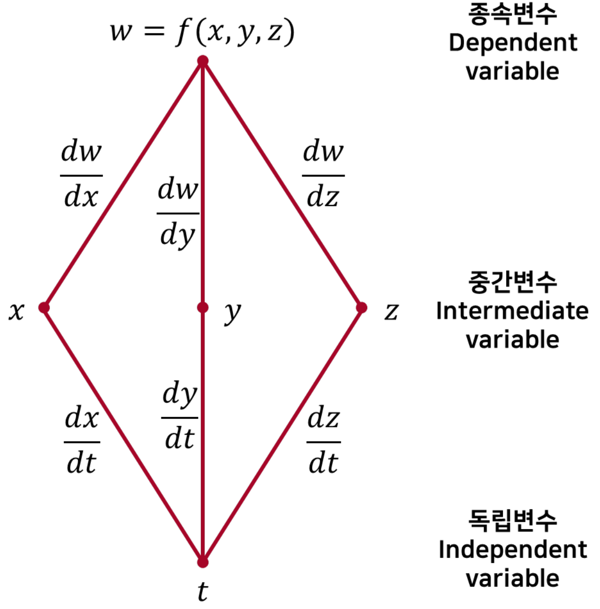
$$\frac{dw}{dt} = \frac{\partial w}{\partial x}\frac{dx}{dt} + \frac{\partial w}{\partial y}\frac{dy}{dt}$$



# 연쇄법칙Chain rule

- $z = f_1(t),\ \ y = f_2(t),\ \ x = f_3(t)$ 일 때 $w = f(x, y, z)$의 $t$ 에 대한 변화

$$\frac{dw}{dt} = \frac{\partial w}{\partial x}\frac{dx}{dt} + \frac{\partial w}{\partial y}\frac{dy}{dt} + \frac{\partial w}{\partial z}\frac{dz}{dt}$$



# 로지스틱 함수의 미분

- 로지스틱 함수를 미분하기 위해 분수함수의 미분과 합성함수 미분을 사용
  - $\sigma(z) = \frac{1}{1+\exp(-z)}$

$$\frac{d}{dz}\left(\frac{1}{1+\exp(-z)}\right) = \frac{0\cdot(1+\exp(-z)) - 1\cdot(1+\exp(-z))'}{(1+\exp(-z))^2}$$

$$y' = \frac{f'(x)g(x) - f(x)g'(x)}{\{g(x)\}^2}$$

$$= \frac{-1\cdot\exp(-z)(-1)}{(1+\exp(-z))^2} = \frac{\exp(-z)}{(1+\exp(-z))^2}$$

$$\frac{d\,\exp(-z)}{dz} = \frac{d\,\exp(u)}{du}\frac{du}{dz} \qquad u = -z$$

$$= \frac{1+\exp(-z)-1}{(1+\exp(-z))^2}$$

$$= \frac{1}{1+\exp(-z)} \times \frac{1+\exp(-z)-1}{1+\exp(-z)}$$

$$= \frac{1}{1+\exp(-z)}\left(\frac{1+\exp(-z)}{1+\exp(-z)} - \frac{1}{1+\exp(-z)}\right) = \sigma(z)(1-\sigma(z))$$

# 소프트맥스 함수의 미분

- 벡터함수의 편미분 → 인덱스별로 나눠서
- 소프트맥스 함수의 i번째 성분을 j번째 독립변수로 미분

$$\frac{\partial}{\partial z_j} s_i(\mathbf{z}) = \frac{\partial}{\partial z_j} \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}}$$

- i=j인 경우: 나눗셈의 미분법 적용
  – 분자 미분

$$\frac{\partial}{\partial z_j} e^{z_i} = e^{z_j}$$

  – 분모 미분

$$\frac{\partial}{\partial z_j} \sum_{k=1}^{K} e^{z_k} = e^{z_j}$$

# 소프트맥스 함수의 미분

- 분자, 분모 미분 결과를 나눗셈의 미분법에 대입

$$
\begin{aligned}
\frac{\partial}{\partial z_j} s_i(\mathbf{z}) &= \frac{\partial}{\partial z_j} \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}} \\
&= \frac{e^{z_j}\left(\sum_{k=1}^{K} e^{z_k}\right) - e^{z_j} \cdot e^{z_j}}{\left(\sum_{k=1}^{K} e^{z_k}\right)^2} \\
&= \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} - \frac{e^{z_j} \cdot e^{z_j}}{\left(\sum_{k=1}^{K} e^{z_k}\right)^2} \\
&= s_j(\mathbf{z}) - (s_j(\mathbf{z}))^2 \\
&= s_j(\mathbf{z})(1 - s_j(\mathbf{z}))
\end{aligned}
$$

분자 미분: $\frac{\partial}{\partial z_j} e^{z_i} = e^{z_j}$

분모 미분: $\frac{\partial}{\partial z_j} \sum_{k=1}^{K} e^{z_k} = e^{z_j}$

# 소프트맥스 함수의 미분

- i≠j인 경우
  - 분자 미분

$$\frac{\partial}{\partial z_j} e^{z_i} = 0$$

  - 분모 미분

$$\frac{\partial}{\partial z_j} \sum_{k=1}^{K} e^{z_k} = e^{z_j}$$

$$\frac{\partial}{\partial z_j} s_i(\mathbf{z}) = \frac{\partial}{\partial z_j} \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}}$$

$$= \frac{0 \cdot \left(\sum_{k=1}^{K} e^{z_k}\right) - e^{z_i} \cdot e^{z_j}}{\left(\sum_{k=1}^{K} e^{z_k}\right)^2}$$

$$= \frac{-e^{z_i} \cdot e^{z_j}}{\left(\sum_{k=1}^{K} e^{z_k}\right)^2}$$

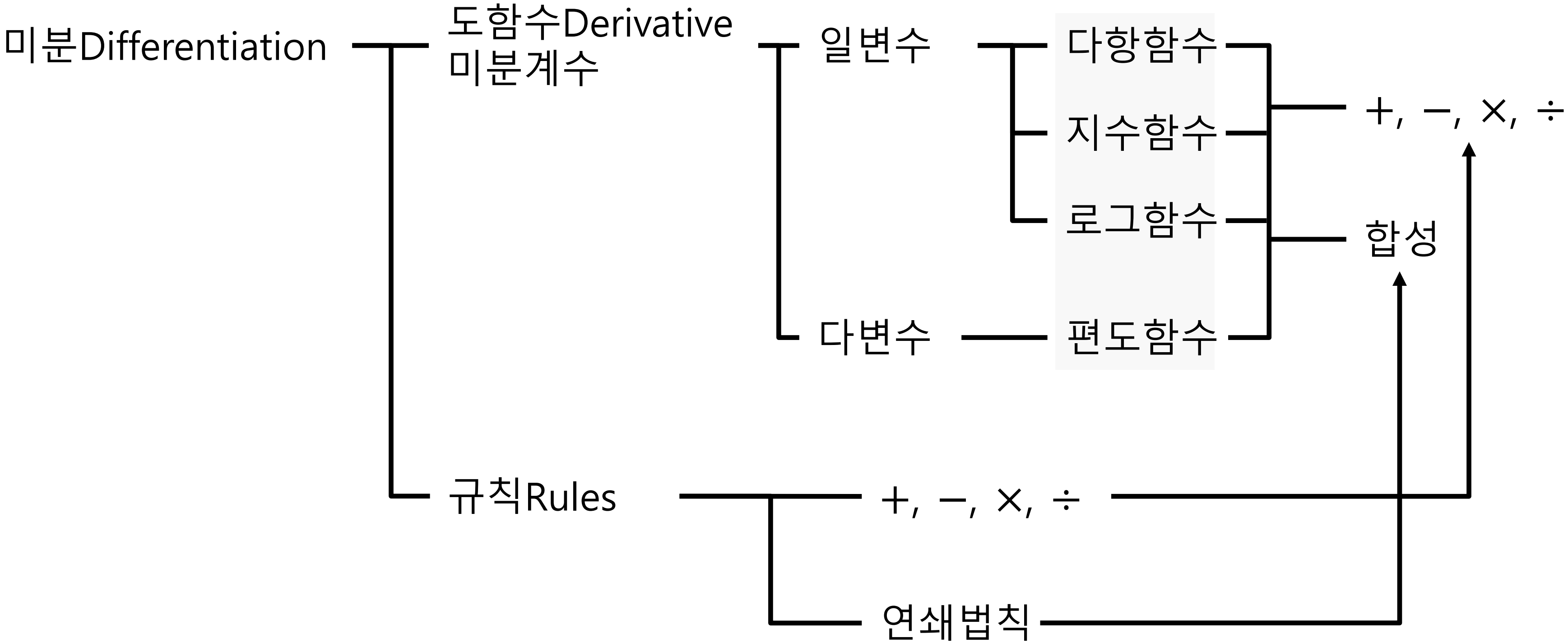
$$= -s_i(\mathbf{z}) s_j(\mathbf{z})$$

# 소프트맥스 함수의 미분

$$\frac{\partial}{\partial z_j} s_i(\mathbf{z}) = \begin{cases} s_j(\mathbf{z})(1 - s_j(\mathbf{z})) & i = j \\ -s_i(\mathbf{z}) s_j(\mathbf{z}) & i \neq j \end{cases}$$

$$\frac{\partial}{\partial \mathbf{z}} \mathbf{s}(\mathbf{z}) = \begin{bmatrix} s_1(\mathbf{z})(1 - s_1(\mathbf{z})) & -s_1(\mathbf{z}) s_2(\mathbf{z}) & \cdots & -s_1(\mathbf{z}) s_K(\mathbf{z}) \\ -s_2(\mathbf{z}) s_1(\mathbf{z}) & s_2(\mathbf{z})(1 - s_2(\mathbf{z})) & \cdots & -s_2(\mathbf{z}) s_K(\mathbf{z}) \\ \vdots & \vdots & \ddots & \vdots \\ -s_K(\mathbf{z}) s_1(\mathbf{z}) & -s_K(\mathbf{z}) s_2(\mathbf{z}) & \cdots & s_K(\mathbf{z})(1 - s_K(\mathbf{z})) \end{bmatrix}$$

# 총정리


미분Differentiation
도함수Derivative
미분계수
일변수
다항함수
지수함수
로그함수
다변수
편도함수
+, −, ×, ÷
합성
규칙Rules
+, −, ×, ÷
연쇄법칙