Math for AI

A.K.A. 인공지능을 위해 꼭 필요한 것만 고른 수학

Seongjin Lee

July 11, 2020

Gyeongsang National University

Table of contents

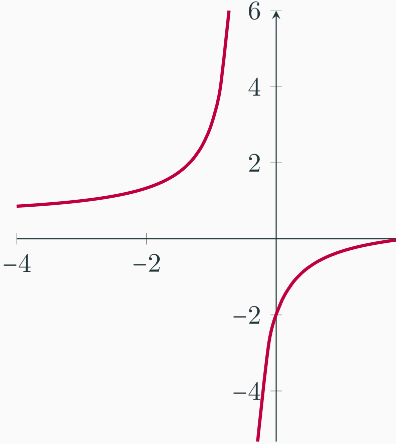
기초

변수와 상수

1. 변수(variable)는 비어있는 상자와 같아서 기본적으로 한 번의 한 가지 정보를 저장할 수 있음
2. 상수(constant)는 고정된 값으로 한 번 정해지면 변하지 않음

$$y = (x - 2)/(2x + 1)$$

학습할 때 가중치는 변수로서 계속 변하지만, 학습이 완료된 후 모델에서 사용될 때는 상수의 역할을 함



항 숫자나 문자, 또는 그 둘의 곲으로 표현 되는 식
예: $3, 3a, -4ab, \frac{x}{3}, a^2$

차수 각 항에 변수가 곱해진 횟수. 이 때 상수만 곱해진 항은 차수가 0 (예: 3이라는 항은 차수가 0)
$-4ab$는 a의 차수 1과 b의 차수의 합으로 2, a^2의 차수는 2

계수 각 항에서 문자(변수)를 제외한 부분
(예: 3이라는 항에서 계수는 3, $\frac{x}{3}$은 $\frac{1}{3} \times x$이므로 계수는 $\frac{1}{3}$)

단항식 1개의 항으로 이루어진 식

다항식 여러 항이 덧셈이나 뺄셈으로 연결된 식
$3a - 2b + 4a^2b + 6$
계수는 순서대로 $3, -2, 4, 6$,
차수는 순서대로 $1, 1, 3, 0$

최고차항 다항식에서 차수가 가장 높은 항. 최고차항이 다항식의 차수가 됨

x의 1차식 수식의 항 중에서 최고차항의 차수가 1인 식

절편 $x = 0$일 때 y의 값

기울기 $y = ax + b$에서 계수 a. x의 증가 속도를 나타냄

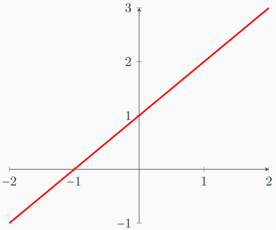


Figure 1: $a > 0$

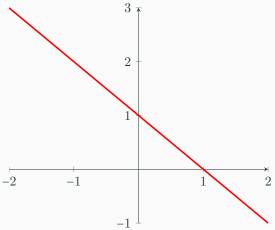


Figure 2: $a < 0$

x의 2차식 수식의 항 중 최고차항이 2인 식

$$y = ax^2 + bx + c$$

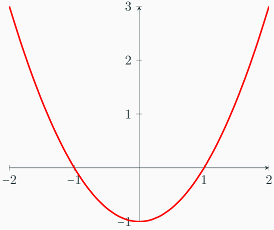


Figure 3: $a > 0$

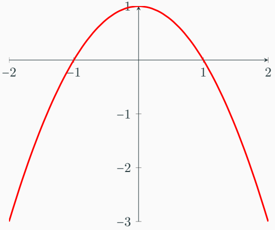


Figure 4: $a < 0$

함수의 개념

함수 어떤 입력값 x에 따라 하나의 출력값 y가 결정된다면 y는 x의 함수이다.

$$x \to f(x) = 2x \to y$$

컴퓨터공학에서의 함수
수학의 함수보다 더 확장된 개념으로 어떤 입력 값에 대해 참이나 거짓 같은 형태 또는 문자열 같은 형태도 출력으로 사용될 수 있음

제곱근

제곱근 제곱을 했을 때 어떤 수가 되는 값을 그 어떤 수에 대한 제곱근이라고 부름.
어떤 수 a에 대해 $a = b^2$을 만족하는 b가 있다면 이러한 b를 a의 제곱근이라고 함. 실수에서는 양수에 대한 제곱근이 반드시 두 개 존재함 $\pm\sqrt{a}$
$\sqrt{}$ 로 표현하고 루트라고 읽음.

$a > 0, b > 0, c > 0$일 때 다음의 식이 성립함

1. $\sqrt{a^2} = a$
2. $a \times \sqrt{b} = a\sqrt{b}$
3. $b\sqrt{a} + c\sqrt{a} = (b + c)\sqrt{a}$
4. $\sqrt{a} \times \sqrt{b} = \sqrt{ab}$
5. $\sqrt{a} \div \sqrt{c} = \frac{\sqrt{a}}{\sqrt{c}} = \sqrt{\frac{a}{c}}$
6. $\sqrt{a^2 \times b} = a\sqrt{b}$

거듭제곱과 거듭제곱근

거듭제곱 a를 p번 곱한 것을 a의 p 제곱 또는 a의 p승이라고 부르고 a^p라고 표기함.

a를 밑수(base) p를 지수(exponent)라고 함. 지수는 분수 또는 음수가 될 수 있음.

거듭제곱근 p제곱을 하면 a가 되는 수를 a의 p 제곱근이라고 부르고 $\sqrt[p]{a}$라고 표기함

$\sqrt[2]{a}$는 평방근이라고 부르고 2를 생략하기도 함

$a > 0, b > 0$이라고 가정

$$
\begin{aligned}
a^0 &= 1 \\
a^p\, a^q &= a^{p+q} \\
(a^p)^q &= a^{p+q} \\
(ab)^p &= a^p\, b^p \\
a^{-p} &= \frac{q}{a^p} \\
\sqrt[p]{a}\sqrt[p]{b} &= \sqrt[p]{ab} \\
\sqrt[p]{\sqrt[q]{a}} &= \sqrt[pq]{a} \\
\sqrt[p]{a} &= a^{\frac{1}{p}}
\end{aligned}
$$

지수함수와 로그함수 i

지수함수 $a > 0, a \neq 1$ 일 때 $y = a^x$와 같이 표현되는 함수

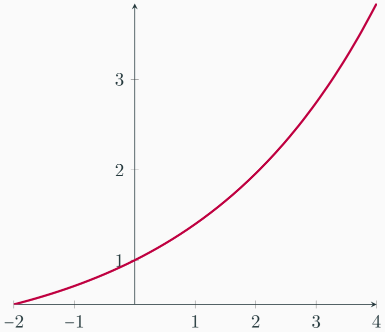


Figure 5: $a > 1$

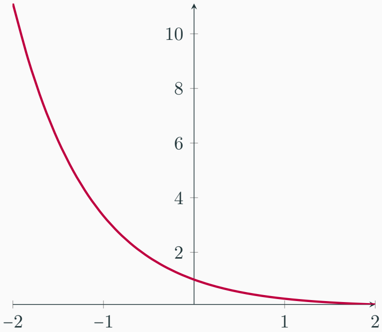


Figure 6: $0 < a < 1$

지수함수와 로그함수 ii

로그 어떤 x가 a^y이라고 표현될 때의 지수 y를 a를 밑으로 하는 x의 로그라고 하며, 기호 $\log$를 사용하여 $y = \log_a^x$와 같이 표현함. 이 때, x를 진수(antilogarithm)라고 하는데 $a > 0, a \neq 1, x > 0$이다.

$a > 0, a \neq 1, x, y > 0$일 때 다음이 성립함

$$
\begin{aligned}
\log_a a &= 1 \\
\log_a 1 &= 0 \\
\log_a xy &= \log_a x + \log_a y \\
\log_a \frac{x}{y} &= \log_a x - \log_a y \\
\log_a x^y &= y \log_a x \\
\log_a x &= \frac{\log_c x}{\log_c a}, c > 0, c \neq 1
\end{aligned}
$$

로그함수 진수를 변수로 사용하는 함수. $a > 0, a \neq 1, x > 0$일 때 다음과 같이 표현되는 함수를 로그함수라 함.

$$y = \log_a x$$

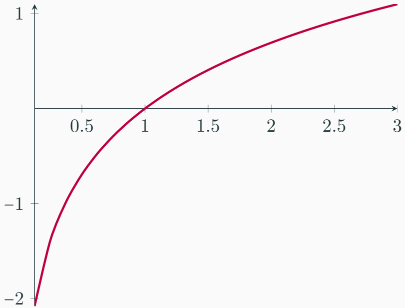


Figure 7: $a > 1$

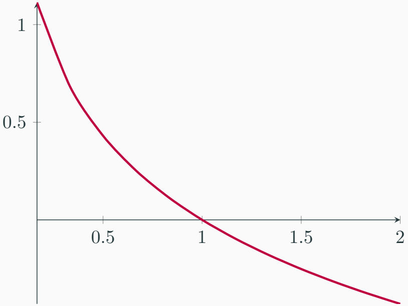


Figure 8: $0 < a < 1$

지수함수와 로그함수 iv

- 가능성을 나타내는 척도로 우도 또는 가능도 (likelihood)를 사용하며, 가능도를 나타내는 함수를 likelihood function 가능도 함수라고함
- likelihood 식은 확률 식과 같고 0과 1 사이의 값을 갖음
- likelihood를 계속 곱하다 보면 값이 계속 작아져서 다루기 어려워짐. 그래서 likelihood와 로그 함수를 같이 사용함
- 로그를 사용하면 $\log_a XY = \log_a X + \log_a Y$와 같이 곱셉을 덧셈으로 표현할 수 있어서 계산이 쉬워짐

자연로그의 밑, 네이피어 상수 e

$$e = \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n = 2.718281\ldots$$

- 자연상수 e를 사용하는 이유는 몇 가지 유용한 특징 때문임

$$\begin{aligned} \frac{d}{dx} e^x &= e^x \\ \frac{d}{dx} \ln x &= \frac{1}{x} \end{aligned}$$

시그모이드 함수 Sigmoid 는 다음과 같이 표현됨.

$$\varsigma_a(x) = \frac{1}{1 + \exp(-ax)}$$

이 때, a를 게인 (gain)이라 부르는데 특별히 $a = 1$일 때의 시그모이드 함수를 표준 시그모이드 함수라고 부름. x가 음의 무한대로 갈 때 0, 양의 무한대로 갈 때 1, 그리고 0일 때 $\varsigma_a(0) = \frac{1}{2}$가 됨

시그모이드 함수 ii

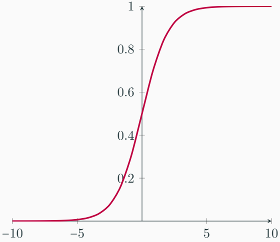


Figure 9: $a > 1$

다양한 활성화 함수

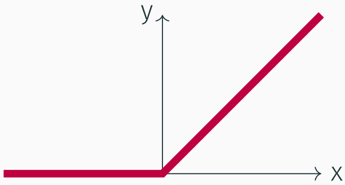


Figure 10: ReLU

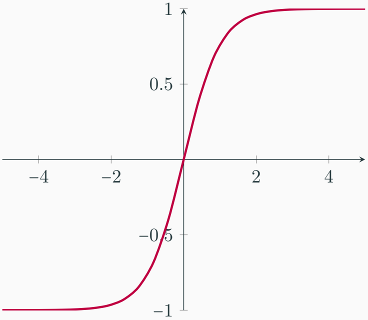


Figure 11: tanh 함수

삼각 함수 각의 크기에 따라 값이 달라지는 함수

도수법 원이 한 바퀴 도는 데 필요한 각을 360° 로 표현

호도법 반지름이 r인 원에서 그 반지름과 같은 길이의 호 AB가 있다고 할 때 그 중심각의 크기는 항상 일정함. 이때의 각을 1 radian이라고 부르고 1 rad라고 표기함

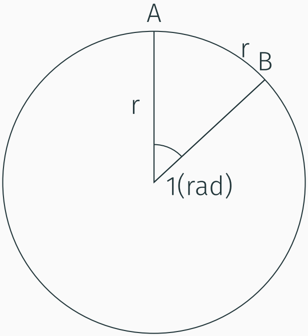


Figure 12: radian

삼각함수 iii

도수법	0°	30°	45°	60°	90°	120°	180°	360°
호도법	0	$\frac{1}{6}\pi$	$\frac{1}{4}\pi$	$\frac{1}{3}\pi$	$\frac{1}{2}\pi$	$\frac{2}{3}\pi$	π	2π

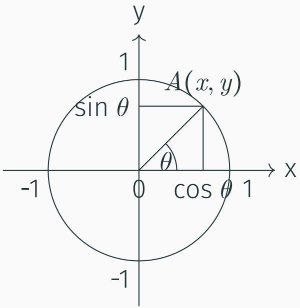


Figure 13: 단위 원과 삼각함수

$$sin\theta = y$$

$$cos\theta = x$$

$$tan\theta = \frac{y}{x}$$

삼각함수 iv

θ	0	$\frac{1}{6}\pi(=30^\circ)$	$\frac{1}{4}\pi(=45^\circ)$	$\frac{1}{3}\pi(=60^\circ)$	$\frac{1}{2}\pi(=90^\circ)$
$sin\theta$	0	$\frac{1}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{\sqrt{3}}{2}$	1
$cos\theta$	1	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{2}}{2}$	$\frac{1}{2}$	0
$tan\theta$	0	$\frac{\sqrt{3}}{3}$	1	$\sqrt{3}$	-

θ	$\frac{2}{3}\pi(=120^\circ)$	$\frac{5}{6}\pi(=150^\circ)$	$\pi(=180^\circ)$	$\frac{3}{2}\pi(=270^\circ)$	$\pi(=360^\circ)$
$sin\theta$	$\frac{\sqrt{3}}{2}$	$\frac{1}{2}$	0	-1	0
$cos\theta$	$-\frac{1}{2}$	$-\frac{\sqrt{3}}{2}$	-1	0	1
$tan\theta$	$-\sqrt{3}$	$-\frac{\sqrt{3}}{3}$	0	$-$	0

- 반지름 1인 원의 둘레 위의 한 점 A의 좌표 x, y는 $-1 \le x \le 1$과 $-1 \le y \le 1$ 범위 안에 있음
- 같은 이유로 $\sin\theta, \cos\theta$가 갖을 수 있는 값의 범위도 $-1 \le \sin\theta \le 1$, $-1 \le \cos\theta \le 1$ 이 됨
- $\tan\theta$는 임의의 실숫값을 갖음
- 함수가 값을 가질 수 있는 범위를 치역 (range)라고 함

$$\tan\theta = \frac{\sin\theta}{\cos\theta}$$
$$\sin^2\theta + \cos^2\theta = 1$$
$$1 + \tan^2\theta = \frac{1}{\cos^2\theta}$$

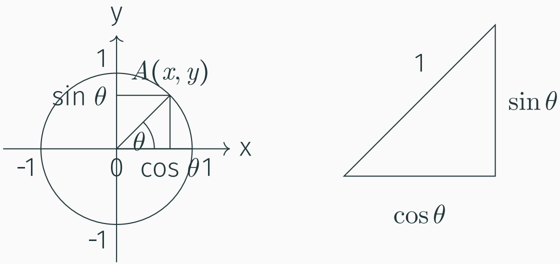


Figure 14: 단위 원과 삼각함수

절댓값 어떤 수와 0과의 수직선과의 거리, $|$를 사용해서 숫자의 앞 뒤를 감싸서 표현. $|3| = 3, |-3| = 3$과 같이 표현됨

유클리드 거리 좌표계 상의 두 점을 잇는 최단 거리의 선
a 의 좌표 (a_x, a_y) b의 좌표 (b_x, b_y)라고 할 때 이 두 점의 유클리드 거리는 다음과 같음.

$$\sqrt{(a_x - b_x)^2 + (a_y - b_y)^2}$$

수열 여러 숫자가 줄지어 나열된 것을 표현함. 공학에서는 일정한 패턴을 갖는 수열을 주로 다룸

항 수열을 구성하는 하나의 숫자

$$a_1, a_2, a_3, a_4, \ldots, a_n$$

a_1 을 첫 항 또는 초항이라고 하고 a_n을 끝 항 또는 말항이라고 함

등차수열 각 항의 차(공차, common difference)가 일정한 수열

등차수열의 일반항 초항 a, 공차가 d일 때, 등차수열의 제 n항 a_n은 다음과 같이 정의한다.

$$a_n = a + (n-1)d$$

수열 ii

등차수열의 합 초항이 a, 말항이 l, 항의 개수는 n, 초항에서 말항까지의 합이 S라고 할 때, 다음의 식이 성립함

$$S = \frac{1}{2}n(a + l)$$

등비수열 인접하는 항의 비율이 일정한 수열, $1, 2, 4, 8, 16, 32, \ldots$ 일반적으로 등비수열은 $a, ar, ar^2, ar^3, \ldots, ar^n$과 같은 형태가 됨.

공비 등비수열에서 인접하는 항의 비율 $a_{n+1} = 2 \times a_n$

등비수열의 일반항 초항이 a, 공비가 r일 때, 등비수열의 제 n항 a_n은 다음과 같이 정의함

$$a_n = ar^{n-1}$$

등비수열의 합 초항이 a, 공비가 r, 초항에서 제 n항까지의 합이 S_n이라고 할 때 다음과 같은 식이 성립함

$$\begin{aligned} \text{if} \quad r \neq 1, S_n &= \frac{a(1-r^n)}{1-r} = \frac{a(r^n-1)}{r-1} \\ \text{if} \quad r = 1, S_n &= na \end{aligned}$$

총합 $\sum$이라는 기호를 쓰고 각 항의 합을 뜻함

$$\begin{aligned} \sum_{k=1}^{4}(3k+1) &= (3\times 1+1)+(3\times 2+1) \\ &\quad +(3\times 3+1)+(3\times 4+1) \\ &= 4+7+10+13 \\ &= 34 \end{aligned}$$

기억할 공식 다음 4 개의 식은 기억해두면 좋음

$\sum_{k=1}^{n} k = \frac{1}{2}n(n+1)$ $\sum_{k=1}^{n} k^2 = \frac{1}{6}n(n+1)(2n+1)$

$\sum_{k=1}^{n} k^3 = \left(\frac{1}{2}n(n+1)\right)^2$ $\sum_{k=1}^{n} c = nc$, c는 상수

총승의 성질 다음과 같은 성질을 갖음

$$\sum_{k=1}^{n}(a_k + b_k) = \sum_{k=1}^{n} a_k + \sum_{k=1}^{n} b_k$$
$$\sum_{k=1}^{n} pa_k = p\sum_{k=1}^{n} a_k, p = \text{상수}$$

총승 $\prod$ 이라는 기호를 쓰고 각 항의 곱을 뜻함

$$
\begin{aligned}
Let \;\; a_k &= k-1 \\
\prod_{k=1}^{4} a_k &= a_1 \times a_2 \times a_3 \times a_4 \\
&= 1 \times 3 \times 5 \times 7 \\
&= 105
\end{aligned}
$$

$$y = \begin{bmatrix} x_1 & x_2 & x_3 & \dots & x_n \end{bmatrix} \times \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ x_n \end{bmatrix} + b$$

$$\begin{aligned} y &= x_1 \cdot w_1 + x_2 \cdot w_2 + \cdots + x_n \cdot w_n + b \\ &= \sum_{k=1}^{n} x_k w_k + b \end{aligned}$$

집합 set, 어떤 조건을 만족하는 것들을 중복되지 않도록 모두 모은 모둠. { 과 }으로 원소를 감싸는 모양으로 표기

원소 element, 이 집합에 들어가는 각각의 것

표기 방법

- $\{2, 4, 6, 8, 10\}$
- $\{x | x \text{ 조건 }\}$

만약 어떤 원소 x가 A에 속한다면 $x \in A$ 그렇지 않다면 $x \notin A$

두 개의 집합 A, B가 있을 때 두 집합의 원소가 서로 완전히 일치하면 $A = B$이라고 표현

집합과 원소 ii

부분집합 집합 B의 모든 원소가 집합 A의 원소라면 집합 B는 A의 부분집합이라고 표현

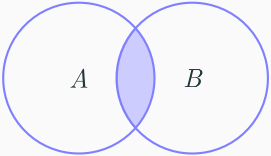
교집합 intersection, 두 개의 집합이 있을 때 두 개의 집합에 모두 속하는 원소들의 집합, $A \cap B$

합집합 union, 두 개의 집합이 있을 때 적어도 한 집합에 속하는 원소들의 집합, $A \cup B$

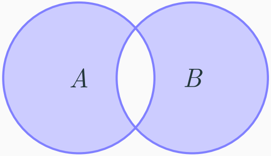
공집합 원소가 하나도 없는 집합. ϕ로 표현하고 ϕ는 모든 집합의 부분집합임. 어떤 집합 A가 있을 때 $\phi \subset A$라고 할 수 있음

집합과 원소 iii

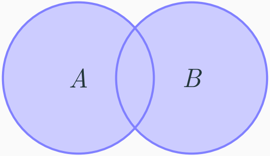
$A \cap B$



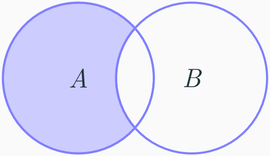
$\overline{A \cap B}$



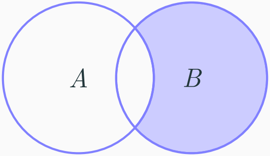
$A \cup B$



$A - B$



$B - A$



미분

극한

극한 수열이나 함숫값이 어떤 특정 값에 한없이 가까워지는 것을 의미

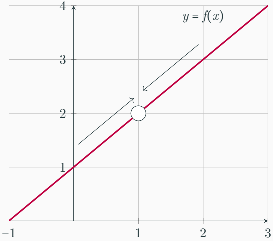
수렴 convergent, x의 값을 어떤 값 a에 최대한 가깝게 만들 때, 함수 $f(x)$의 어떤 값 a에 최대한 가까워지는 모양

극한값 limit, limiting value. 수렴하려는 값을 표현. a 는 함수 $f(x)$에서 $x \to a$일 때의 극한값이라고 표현

$$\lim_{x \to a} f(x) = \alpha, \qquad f(x) \to \alpha (x \to a)$$

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

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



$f(x)$는 $x = 1$일 때 분모가 0이 되기 때문에 y를 정의할 수 없음. $x \neq 1$일 때는 정의 가능

예제
강남역에서 인천공항까지 72.56km의 거리를 자동차로 이동하는데 1시간 반이 걸렸다. 이때의 이동 평균 속도를 구하라.

$$
\begin{aligned}
\text{평균 속도} v &= \frac{72.56km}{1.5h} \approx 48.37km/h \\
\text{순간 속도 } v &= \lim_{\Delta t \to 0} \frac{\Delta x}{\Delta t} \\
&= \lim_{\delta t \to 0} \frac{x(x+\delta x) - x(t)}{\Delta t} \\
\text{미분 표기법} & \quad \frac{dx(t)}{dt}
\end{aligned}
$$

미분 풀이

$$y = f(x)$$
$$y = \alpha x + \beta$$

$$\text{식 1: } f(a) = \alpha a + \beta$$
$$\text{식 2: } f(b) = \alpha b + \beta$$

$$f(b) - f(a) = \alpha(b - a)$$
$$\text{식 3: } \alpha = \frac{f(b) - f(a)}{b - a}$$

$$\beta = f(a) - \alpha a$$
$$= f(a) - \frac{f(b) - f(a)}{b - a} a$$

- 식 3: 기울기 α – 두 점 사이에서 평균적으로 변화한 정도. 평균변화율
- 이해: 함수 $f(x)$ 위의 점 $(a, f(x))$ 에서 순간적으로 변화는 정도 (기울기) $alpha = \frac{\mathrm{d}f(x)}{\mathrm{d}x}$
- 어떤 함수의 특정한 지점의 기울기를 구하는 것을 미분한다고 표현

미분한다

$$\begin{aligned}\frac{\mathrm{d}f(a)}{\mathrm{d}a} &= \lim_{\Delta x\to 0}\frac{\Delta f(a)}{\Delta x}\\ &= \lim_{\Delta h\to 0}\frac{f(a+h)-f(a)}{(a+h)-a}\\ &= \lim_{\Delta h\to 0}\frac{f(a+h)-f(a)}{h}\end{aligned}$$

- **접선**: 함수 $f(x)$ 위의 한 점 a 을 지나는 직선
- **미분계수**: 평균변화율의 극한 값인 α는 $x=a$일 때의 미분계수라고 함

- 상수 a는 변수 x가 가질 수 있는 수 많은 값들 중 하나임.
- 상수 a에 어떤 x를 대입하더라도 $\frac{\mathrm{d}f(a)}{\mathrm{d}x}x$의 값은 결정되므로 $\frac{\mathrm{d}f(a)}{\mathrm{d}x}x$는 x에 대한 일종의 함수
- $\frac{\mathrm{d}f(x)}{\mathrm{d}x}x$라 쓰고 도함수 (derivative)라고 부름.

$$\begin{aligned} y &= \frac{\mathrm{d}f(a)}{\mathrm{d}x}x+\left(f(a)-\frac{\mathrm{d}f(a)}{a}\right)\\ &= \frac{\mathrm{d}}{f(a)}(x-a)+f(a)\end{aligned}$$

상미분
Ordinary derivative, 변수가 하나만 있는 함수의 미분

1. $y = x^r$ 일 때 $\frac{\mathrm{d}f(x)}{x} = rx^{r-1}$, 이 때 r은 임의의 실수
2. $\frac{\mathrm{d}}{x}\left(f(x) + g(x)\right) = \frac{\mathrm{d}f(\mathrm{d}x)}{\mathrm{d}x} + \frac{\mathrm{d}g(x)}{\mathrm{d}x}$
3. $\frac{\mathrm{d}}{\mathrm{d}x}\left(kf(x)\right) = k\frac{\mathrm{d}f(x)}{x}$

전미분
Total derivative, 변수가 두 개 이상이 있는 함수의 미분

$$
\begin{aligned}
z &= f(x,y) \\
&= 3x^2 + 2xy + 2y^2 \\
\Delta z &= f(x+\Delta x, y+\Delta y) - f(x,y) \\
&= 3(x+\Delta x)^2 + 2(x+\Delta x)(y+\Delta y)2(x+\Delta y)^2 - (3x^2 + 2xy + 2y^2) \\
&= (6x+2y)\Delta x + (2x+4y)\Delta y + 3\Delta x^2 + 2\Delta x\Delta y + 2\Delta y^2
\end{aligned}
$$

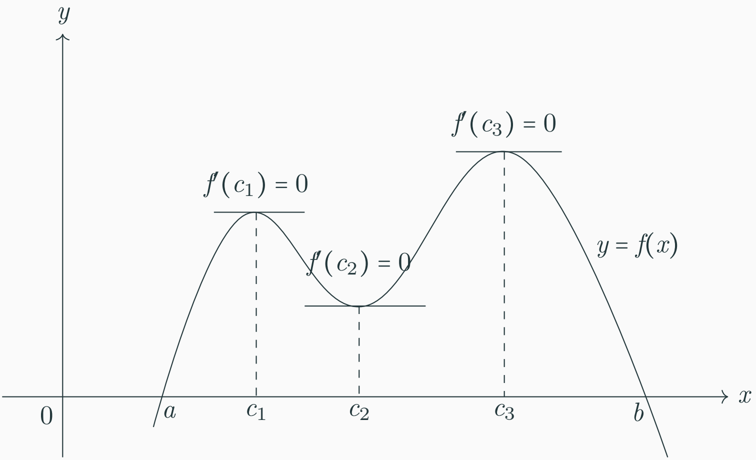
편미분
partial derivative, 변수가 두 개 이상일 때 하나의 변수 외에는 고정을 시킨 함수의 미분
$\Delta y = 0$일 때 $\Delta x \to 0$로 변화시키는 것

상미분과 편미분 iii

$$\begin{aligned}\frac{\delta f(x,y)}{\delta x} &= \lim_{\Delta x \to 0} \frac{\Delta z}{\Delta x} \\ &= \lim_{\Delta x \to 0} 6x + 2y + 3\Delta x \\ &= 6x + 2y\end{aligned}$$

$$\frac{\delta f(x,y)}{\delta y} = 2x + 4y$$

편미분을 간단하게 다음과 같이 표현할 수 있음
$\frac{\delta f(x,y)}{\delta x}$ 는 $f_x(x,y)$, $\frac{\delta f(x,y)}{\delta t}$ 는 $f_y(x,y)$


y
$f'(c_3) = 0$
$f'(c_1) = 0$
$f'(c_2) = 0$
$y = f(x)$
x
0
a
c_1
c_2
c_3
b

그래프 그리기 ii

$\frac{\mathrm{d}x}{\mathrm{d}t}$	$\frac{\mathrm{d}^2x}{\mathrm{d}t^2}$	화살표	의미
0	NA	→	x는 일정 $\left(\frac{\mathrm{d}x}{\mathrm{d}t} = 0\right)$
+	+	↗	x는 증가 $\left(\frac{\mathrm{d}x}{\mathrm{d}t} > 0\right)$ 하고, 증가율이 증가 $\left(\frac{\mathrm{d}^2x}{\mathrm{d}t^2} > 0\right)$
+	0	↗	x는 증가 $\left(\frac{\mathrm{d}x}{\mathrm{d}t} > 0\right)$ 하고, 증가율이 일정 $\left(\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = 0\right)$
+	−	↗	x는 증가 $\left(\frac{\mathrm{d}x}{\mathrm{d}t} > 0\right)$ 하고, 증가율이 감소 $\left(\frac{\mathrm{d}^2x}{\mathrm{d}t^2} < 0\right)$
−	+	↘	x는 감소 $\left(\frac{\mathrm{d}x}{\mathrm{d}t} > 0\right)$ 하고, 감소율이 감소 $\left(-\frac{\mathrm{d}^2x}{\mathrm{d}t^2} < 0\right)$
−	0	↘	x는 감소 $\left(\frac{\mathrm{d}x}{\mathrm{d}t} < 0\right)$ 하고, 감소율이 일정 $\left(-\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = 0\right)$
−	−	↘	x는 감소 $\left(\frac{\mathrm{d}x}{\mathrm{d}t} < 0\right)$ 하고, 감소율이 증가 $\left(-\frac{\mathrm{d}^2x}{\mathrm{d}t^2} > 0\right)$

변곡점: 그래프 상에서 곡선의 방향이 바뀌는 지점을 말함. 변곡점에서 2계미분을 하면 값이 0이 됨. 값의 앞과 뒤 점에서의 부호가 반전됨.

극값, 극대 또는 극소: 변곡점 중 최대 또는 최소 값을 갖는 점을 극대 또는 극소라 하고 통틀어서 극값이라고 함.

함수의 최댓값과 최소값

최댓값과 최솟값 극점 또는 함수가 정의된 구간 양끝단에서 나옴

x	-3	$\cdots$	1	$\cdots$	10
$\frac{\mathrm{d}f(x)}{\mathrm{d}x}$	-	-	0	+	+
$\frac{\mathrm{d}^2 f(x)}{\mathrm{d}x^2}$	+	+	+	+	+
$f(x)$	20	$\searrow$	4	$\nearrow$	85

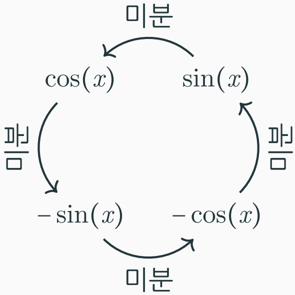
초등함수 x^r (멱함수), a^x (지수함수), $\log_c x$ (로그함수), 삼각함수 등을 통틀어 초등함수(elementary function)라 부름

초등함수의 미분 공식

원래 함수			원래 함수를 x로 미분한 것
역함수		x^r	rx^{r-1}
지수함수		$e^x, \exp(x)$	$e^x, \exp(x)$
		a^x	$a^x \log_e^a$
로그함수		$\log_e x \quad (x > 0)$	$\frac{1}{x}$
삼각함수	사인함수	$\sin(x)$	$\cos(x)$
	코사인함수	$\cos(x)$	$-\sin(x)$
	탄젠트함수	$\tan(x)$	$\frac{1}{\cos(x)}$

초등함수와 합성함수의 미분, 그리고 곱의 법칙 ii

사인과 코사인의 관계



공식

- **합성함수 미분(변수 한 개)**: $y = f(x)$

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}y}{\mathrm{d}u} \cdot \frac{\mathrm{d}u}{\mathrm{d}x}$$

- **합성함수 미분(변수 여러 개)**: 연쇄 법칙 (chain rule)이 적용됨 $z = f(x, y)$

$$\frac{\mathrm{d}z}{\mathrm{d}x} = \frac{\mathrm{d}z}{\mathrm{d}u} \cdot \frac{\mathrm{d}u}{\mathrm{d}x} + \frac{\mathrm{d}z}{\mathrm{d}v} \cdot \frac{\mathrm{d}v}{\mathrm{d}x}$$

- **곱의 법칙**

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(f(x)g(x)\right) = \frac{\mathrm{d}(x)}{\mathrm{d}x} g(x) + f(x)\frac{\mathrm{d}g(x)}{\mathrm{d}x}$$

예: $f(x) = (3x-4)^{50}$ 의 미분
$u = 3x - 4$ 라고 하자. 이 때 다음과 같이 표현할 수 있음.

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x} = \frac{\mathrm{d}f(x)}{\mathrm{d}u} \cdot \frac{\mathrm{d}u}{\mathrm{d}x}$$

계산하면 다음과 같음.

$$\begin{aligned}
\frac{\mathrm{d}f(x)}{\mathrm{d}x} &= \frac{\mathrm{d}u^{50}}{\mathrm{d}u} \cdot \frac{\mathrm{d}(3x-4)}{\mathrm{d}x} \\
&= 50u^{49} \cdot 3 \\
&= 150(3x-4)^{49}
\end{aligned}$$

시그모이드 함수

$$\varsigma_a(x) = \frac{1}{1 + \exp(-ax)}$$

미분

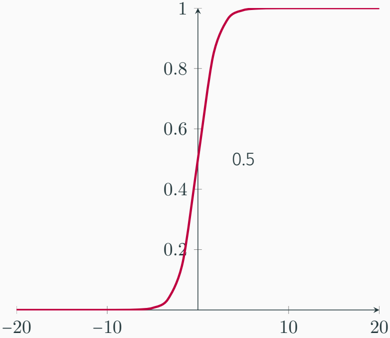
$$\begin{aligned} \frac{\mathrm{d}\varsigma_a(x)}{\mathrm{d}x} &= \frac{a \cdot \exp(-ax)}{\left(1 + \exp(-ax)\right)^2} \\ &= a\varsigma_n(x)\left(1 - \varsigma_a(x)\right) \end{aligned}$$

시그모이드 함수의 2차 미분

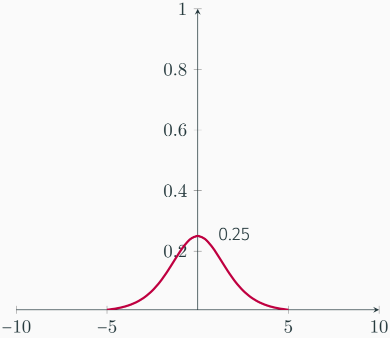
$$
\begin{aligned}
\frac{\mathrm{d}^2 \varsigma_a(x)}{\mathrm{d}x^2} &= \frac{\mathrm{d}\left(a\varsigma_a(x)\left(1-\varsigma_a(x)\right)\right)}{\mathrm{d}x} \\
&= a\frac{\mathrm{d}\varsigma_a(x)}{\mathrm{d}x}\left(1-\varsigma_a(x)\right) + a\varsigma_a(x)\frac{\mathrm{d}\left(1-\varsigma_a(x)\right)}{\mathrm{d}x} \\
&= a\frac{\mathrm{d}\varsigma_a(x)}{\mathrm{d}x}\left(\varsigma_a(x)\right) + a\varsigma_a(x)\frac{\mathrm{d}\left(1-\varsigma_a(x)\right)}{\mathrm{d}x} \\
&= a\frac{\mathrm{d}\varsigma_a(x)}{\mathrm{d}x}\left(1-2\varsigma_a(x)\right) \\
&= a^2\varsigma_a(x)\left(1-\varsigma_a(x)\right)\left(1-2\varsigma_a(x)\right)
\end{aligned}
$$

특수 함수의 미분 iii

표준 시그모이드 함수 ($a = 1$)



표준 시그모이드 함수의 미분

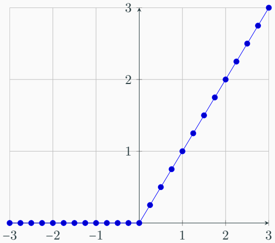


ReLU (Rectified Linear Unit) 함수

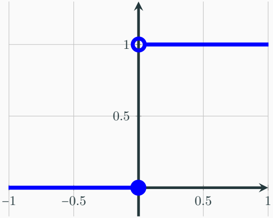
$$\varphi(x) = \max(0, x) = \begin{cases} x & (x > 0) \\ 0 & (x \le 0) \end{cases}$$

$$\varphi'(x) = \begin{cases} 1 & (x > 0) \\ 0 & (x \le 0) \end{cases}$$

ReLU 함수 $\varphi(x)$의 그래프



ReLU 함수의 미분 $\varphi'(x)$ 그래프



선형대수

벡터

벡터를 표현하는 세 가지 방법

- $\mathbf{a}$ 출판물의 활자에 자주 사용됨
- $\vec{a}$ 벡터 표기할 때 자주 사용됨
- $\mathbb{A}$ 실수, 자연수 등을 표현하는데 사용되기 때문에 잘 안 쓰임

행벡터: 성분을 나열한 방식이 가로인 경우

$$\mathbf{a} = \begin{pmatrix} a_1 & a_2 & a_3 & \dots & a_n \end{pmatrix}$$

열벡터: 성분을 나열한 방식이 세로인 경우

$$\mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \dots \\ b_n \end{pmatrix}$$

덧셈과 뺄셈, 그리고 스칼라배

차원이 같은 경우만 덧셈, 뺄셈이 가능함

벡터의 덧셈

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

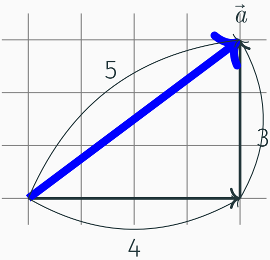
벡터의 뺄셈

$$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} + \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix} = \begin{pmatrix} 1-4 \\ 2-5 \\ 3-6 \end{pmatrix} = \begin{pmatrix} -3 \\ -3 \\ -3 \end{pmatrix}$$

벡터의 스칼라 배

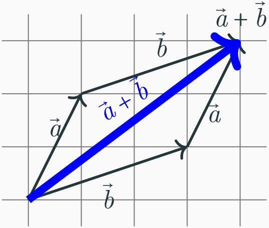
$$2 \cdot \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 2 \times 1 \\ 2 \times 2 \\ 2 \times 3 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \\ 6 \end{pmatrix}$$

유향성분 $\vec{a} = (4,3)$은 오른쪽으로 4, 위쪽으로 3 움직이는 것과 같은 의미임. 이때 원점에서 특정 점 $(4,3)$을 잇는 최단 방향과 그 거리를 나타내는 화살표를 뜻함



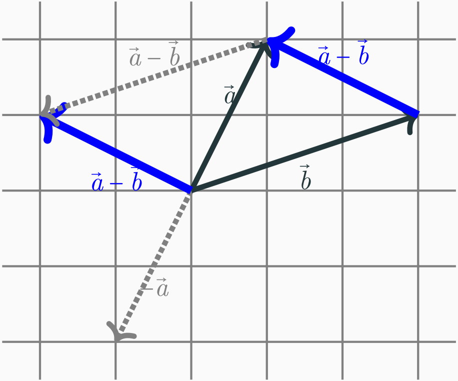
$\vec{a} = (1,2), \vec{b} = (3,1)$ 일 때

$$\vec{a} + \vec{b} = (4,3)$$



$\vec{a} = (1, 2), \vec{b} = (3, 1)$ 일 때

- $-\vec{a}$
- $\vec{a} - \vec{b}$
- $-\vec{b}$



내적 벡터에서 서로 대응하는 성분끼리 곱한 다음 그것들을 모두 더한 값. $\vec{a}\cdot\vec{b}$ 로 표현되기도 하고 $\langle\vec{a},\vec{b}\rangle$ 로 표기하기도 함 벡터의 내적은 벡터가 아니라 스칼라가 됨

$$\vec{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}, \vec{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \text{일 때}$$

$$\begin{aligned} \langle\vec{a},\vec{b}\rangle &= a_1 b_1 + a_2 b_2 + \ldots + a_n b_n \\ &= \sum_{i=1}^{n} = a_i b_i \end{aligned}$$

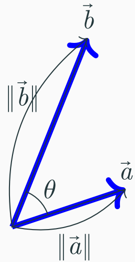
- 내적의 다른 말: 점곱 (dot product), 스칼라곱 (scalar product), 영사곱 (projection product)
- 외적의 다른 말: 교차곱 (cross product), 벡터곱 (vector product), 텐서곱 (tensor product)

기하학적 특징

벡터 $\vec{a}$와 $\vec{b}$가 이루는 각이 θ일 때, $\langle \vec{,} a, \vec{b} \rangle$는 다음과 같이 정의할 수 있음

$$\langle \vec{a}, \vec{b} \rangle = \|\vec{a}\| \|\vec{b}\| \cos\theta$$

이 때 θ 는 $\vec{a}$와 $\vec{b}$의 시작점을 일치시킬 때 생기는 사이각을 말함. $\|\vec{a}\|$는 벡터의 길이 또는 유클리드 거리를 의미함.



$$\langle \vec{a}, \vec{b} \rangle = 2 \times 1 + 1 \times 3 = 2 + 3 = 5$$

$$\|\vec{a}\| = \sqrt{2^2 + 1^2} = \sqrt{5}, \|\vec{b}\| = \sqrt{1^2 + 3^2} = \sqrt{10}$$

$$\langle \vec{,} a, \vec{b} \rangle = \|\vec{a}\| \|\vec{b}\| \cos 45^\circ = \sqrt{5} \times \sqrt{10} \times \frac{\sqrt{2}}{2} = 5$$

직교 조건

두 개의 벡터가 직교한다(수직으로 만난다)는 것은 두 벡터가 이루는 각이 $90°$ 라는 뜻임

$$\langle \vec{a}, \vec{b} \rangle = \|\vec{a}\| \|\vec{b}\| \cos 90°$$

이 때 $\cos 90° = 0$ 이기 때문에 두 벡터의 내적은 0 이 됨.

법선벡터

예를 들어 구 표면 위의 하나의 점에 접하는 접선을 구했을 때, 접선이 존재하는 평면. 이를 접평면 (tangent plane)이라고 하고, 접평면과 구와 접하는 점을 접점이라고 함

접선과 직교하는 벡터를 통해 접평면을 다룰 수 있는 개념으로 법선벡터을 사용함

벡터의 노름 (norm)

노름 (norm) 벡터의 시작점에서 도착점까지 도달하기 위한 x, y에서 이동한 거리. 예: $\vec{a} = (4, 3)$ 일 때, 오른쪽으로 4, 위로 3 이동하여 총 7만큼 이동함. 이렇게 구한 움직인 거리를 $L1 norm$이라고 함. 또는 맨해튼 거리라고도 함.

$\vec{a}$에 대한 $L1$ norm

$$\|a\|_1 = |a_1| + |a_2| + \cdots + |a_n| = \sum_{i=1}^{n} |a_i|$$

$\vec{a}$에 대한 $L2$ norm

$$\|a\|_2 = \sqrt{\sum_{i=1}^{n} a_i^2} = \sqrt{a_1^2 + a_2^2 + \cdots + a_n^2}$$

$\|\vec{a}\|_2 = \sqrt{\langle \vec{a}, \vec{a} \rangle}$와 같이 표현할 수 있음

코사인 유사도 벡터의 내적과 $L2$ norm을 활용하여 $\cos\theta$ 항을 통해 두 벡터의 닮음을 판단

기본 식

$$\langle \vec{a}, \vec{b} \rangle = \sum_{i=1}^{n} a_i b_i$$

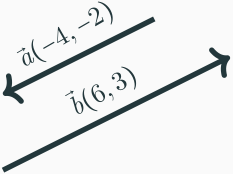
$$\langle \vec{a}, \vec{b} \rangle = \|\vec{a}\| \|\vec{b}\| \cos\theta$$

$$\cos(\vec{a}, \vec{b}) = \frac{\langle \vec{a}, \vec{b} \rangle}{\|\vec{a}\| \|\vec{b}\|}$$

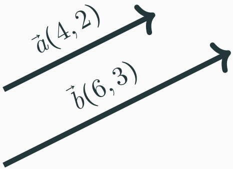
전개

$$\cos(\vec{a}, \vec{b}) \quad = \quad \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \sqrt{\sum_{i=1}^{n} b_i^2}}$$

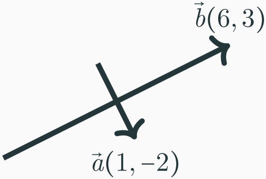
코사인 유사도 ii



코사인 유사도 -1



코사인 유사도 1



코사인 유사도 0

행렬 숫자를 가로와 세로로 늘어 놓은 것

$$3 \times 4\text{행렬}, \begin{bmatrix} 3 & 4 & 0 & 10 \\ 0 & 1 & 0 & -3 \\ -1 & 5 & 9 & 0 \end{bmatrix}$$

위의 행렬은 3행 4열의 행렬 또는 3×4 행렬이라고 함. 이 때 3행 2열의 성분은 5

각 행과 열의 원소 끼리 덧셈 또는 뺄셈을 할 수 있으며 행렬의 크기가 같을 때만 더하거나 뺄 수 있음

행렬은 벡터의 개념이 확장된 것임. 개념 이해를 위해 다음의 행벡터 $\vec{a}$와 $\vec{b}$ 열벡터를 살펴보자.

$$\vec{a} = (-1, 2), \vec{b} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

$\vec{a}$ 와 $\vec{b}$는 각각 1×2 과 2×1 행렬이라고 할 수 있음.

행렬의 곱셉은 벡터의 내적과 같음

행렬의 곱셈 ii

$$
\begin{aligned}
\vec{a}\vec{b} &= \langle \vec{a}, \vec{b} \rangle \\
\vec{a}\vec{b} &= \begin{pmatrix} -1 & 2 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix} \\
&= -1 \times 3 + 2 \times 2 = 1
\end{aligned}
$$

공식

$$
\begin{aligned}
\vec{a} &= \begin{pmatrix} a_1 & a_2 & \dots & a_n \end{pmatrix}, \vec{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \text{일 때} \\
\vec{a}\vec{b} &= \langle \vec{a}, \vec{b} \rangle \\
&= a_1 b_1 + a_2 b_2 + \cdots + a_n b_n = \sum_{i=1}^{n} a_i b_i
\end{aligned}
$$

예제

$$\vec{a_1} = (-1, 2)$$
$$\vec{a_2} = (1, 1)$$
$$A = \begin{bmatrix} \vec{a_1} \\ \vec{a_2} \end{bmatrix} = \begin{bmatrix} -1 & 2 \\ 1 & 1 \end{bmatrix}$$

$$\vec{b} = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

행렬 A와 열벡터 b_1 은

$a_1 b_1 = \langle \vec{a_1}, \vec{b_1} \rangle = -1 \times 3 + 2 \times 2 = 1$

$a_2 b_1 = \langle \vec{a_2}, \vec{b_1} \rangle = 1 \times 3 + 1 \times 2 = 5$

곱셉 결과는 $\vec{a_1}\vec{b_1}$ 과 $\vec{a_2}\vec{b_1}$ 를 세로로 쌓아서 만든 열벡터가 됨

$$\begin{aligned} A\vec{b_1} &= \begin{bmatrix} \vec{a_1} \\ \vec{a_2} \end{bmatrix} \vec{b_1} \\ &= \begin{bmatrix} \langle \vec{a_1}, \vec{b_1} \rangle \\ \langle \vec{a_2}, \vec{b_1} \rangle \end{bmatrix} \\ &= \begin{bmatrix} -1 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} \\ &= \begin{bmatrix} 1 \\ 5 \end{bmatrix} \end{aligned}$$

$$
\begin{aligned}
A &= \begin{bmatrix} \vec{a_1} \\ \vec{a_2} \\ \vdots \\ \vec{a_m} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \vec{b} = \begin{bmatrix} \vec{b_1} \\ \vec{b_2} \\ \vdots \\ \vec{b_n} \end{bmatrix} \\
A\vec{b} &= \begin{bmatrix} \vec{a_1} \\ \vec{a_2} \\ \vdots \\ \vec{a_m} \end{bmatrix} \vec{b} = \begin{bmatrix} \langle \vec{a_1}, \vec{b} \rangle \\ \langle \vec{a_2}, \vec{b} \rangle \\ \vdots \\ \langle \vec{a_n}, \vec{b} \rangle \end{bmatrix} \\
&= \begin{bmatrix} \sum_{i=1}^{n} a_{1i} b_i \\ \sum_{i=1}^{n} a_{2i} b_i \\ \vdots \\ \sum_{i=1}^{n} a_{mi} b_i \end{bmatrix} = \begin{bmatrix} a_{1i} b_1 + a_{12} b_2 + \cdots + a_{1n} b_n \\ a_{2i} b_1 + a_{22} b_2 + \cdots + a_{2n} b_n \\ \vdots \\ a_{1m} b_i + a_{m2} b_2 + \cdots + a_{mn} b_n \end{bmatrix}
\end{aligned}
$$

행렬의 곱셈 vi

$$
\begin{aligned}
A &= \begin{bmatrix} \vec{a}_1 \\ \vec{a}_2 \\ \vdots \\ \vec{a}_n \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \\
B &= \begin{bmatrix} b_1, b2, \cdots, b_l \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1l} \\ b_{21} & b_{22} & \cdots & b_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nl} \end{bmatrix} \\
AB &= \begin{bmatrix} \vec{a}_1 \\ \vec{a}_2 \\ \vdots \\ \vec{a}_n \end{bmatrix} \begin{bmatrix} b_1, b2, \cdots, b_l \end{bmatrix} = \begin{bmatrix} \langle \vec{a}_1, \vec{b}_1 \rangle & \langle \vec{a}_1, \vec{b}_2 \rangle & \cdots & \langle \vec{a}_1, \vec{b}_l \rangle \\ \langle \vec{a}_2, \vec{b}_1 \rangle & \langle \vec{a}_2, \vec{b}_2 \rangle & \cdots & \langle \vec{a}_2, \vec{b}_l \rangle \\ \vdots & \vdots & \cdots & \vdots \\ \langle \vec{a}_n, \vec{b}_1 \rangle & \langle \vec{a}_n, \vec{b}_2 \rangle & \cdots & \langle \vec{a}_n, \vec{b}_l \rangle \end{bmatrix}
\end{aligned}
$$

AB의 제p행 q 열 성분은
$\langle \vec{a}_p, \vec{b}_q \rangle = \sum_{i=1}^{n} a_{pi} b_{iq} = a_{p1} b_{1q} + a_{p2} b_{2q} + \cdots + a_{pn} b_{nq}$ 임

행렬의 스칼라 배

$$
\begin{aligned}
A &= \begin{bmatrix} \vec{a_1} \\ \vec{a_2} \\ \vdots \\ \vec{a_m} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \\
kA &= k\begin{bmatrix} \vec{a_1} \\ \vec{a_2} \\ \vdots \\ \vec{a_m} \end{bmatrix} = \begin{bmatrix} ka_{11} & ka_{12} & \cdots & ka_{1n} \\ ka_{21} & ka_{22} & \cdots & ka_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ ka_{m1} & ka_{m2} & \cdots & ka_{mn} \end{bmatrix}
\end{aligned}
$$

AB 에서 성분 $\langle \vec{a_p}, \vec{b_q} \rangle$를 정의할 수 있으려면 행렬 A의 열의 개수와 행렬 B의 개수가 같아야 함. 즉 $m \times n$ 행렬과 $n \times l$ 행렬을 곱하는 형태여야만 곱셈을 할 수 있고 그 결과는 $m \times l$ 행렬이 됨

행렬의 곱셈에서는 교환법칙 $(AB \neq BA)$이 성립하지 않음

영행렬 성분 전체가 0인 행렬

단위 행렬 왼쪽 위에서 오른쪽 아래 방향으로 대각선상의 모든 성분이 1이고 그 밖의 성분은 0으로 채워진 정방행렬. 어떤 행렬이나 벡터에 단위행렬을 곱하면 그 결과가 달라지지 않고 원래 행렬이나 벡터가 나옴. 항등사상

$$E = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

역행렬 행렬에는 나눗셈 연산을 할 수 없으므로 행렬의 역수 (reciprocal)를 만들어 곱셈으로 같은 효과를 낼 수 있음.

$$\frac{1}{2} \div \frac{3}{5} = \frac{1}{2} \times \frac{5}{3} = \frac{5}{6}$$

$$a \times a^{-1} = a^{-1} \times a = 1$$

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

역행렬 ii

행렬식 역행렬의 존재 여부를 확인하기 위해 사용됨. 행렬식이 0인 경우 역행렬이 존재하지 않음.

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

$$\text{행렬식} = \det A = |A|$$

$$|A| = \det A = ad - bc$$

역행렬 2×2 행렬의 역행렬은 다음과 같은 식으로 구할 수 있음

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

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

2×2 보다 큰 경우 가우스 소거법 (Gaussian elimination) 이나 여인자 전개 (cofactor expansion)과 같은 방법을 써야 함

선형 변환 i

선형 변환 수학적으로 벡터에 행렬을 곱해 또 다른 벡터를 만드는 함수. 하나의 벡터 공간에서 다른 벡터 공간으로, 벡터의 특징을 유지한 채 변환하는 방법

표준 기저 standard basis, x축이나 y축, 그리고 z축처럼 좌표계를 정할 수 있는 벡터의 집합

$$A = \begin{bmatrix} -1 & 2 \\ 1 & 1 \end{bmatrix}, \vec{b_1} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}, \text{일 때} A\vec{b_1} = \begin{bmatrix} -1 & 2 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 5 \end{bmatrix}$$

A와 $\vec{b_1}$의 표준 기저

$$A = (\vec{e_1}, \vec{e_2})$$

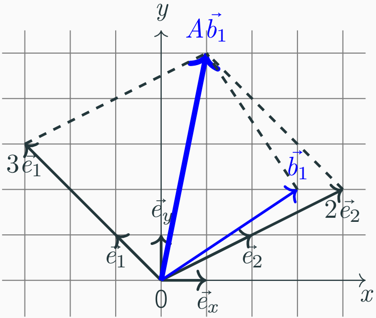
$$\vec{e_1} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \vec{e_2} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$\vec{b_1} = 3\begin{bmatrix} 1 \\ 0 \end{bmatrix} + 2\begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

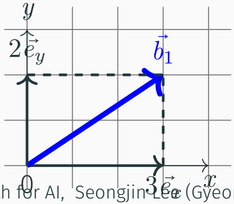
$$= 3\vec{e_x} + 2\vec{e_y}$$

선형 변환 ii

$$
\begin{aligned}
A\vec{b_1} &= \begin{bmatrix} -1 & 2 \\ 1 & 1 \end{bmatrix}\left(3\begin{bmatrix} 1 \\ 0 \end{bmatrix} + 2\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right) \\
&= 3\begin{bmatrix} -1 & 2 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} + 2\begin{bmatrix} -1 & 2 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix} \\
&= 3\begin{bmatrix} -1 \\ 1 \end{bmatrix} + 2\begin{bmatrix} 2 \\ 1 \end{bmatrix} = 3\vec{e_1} + 2\vec{e_2} = \begin{bmatrix} 1 \\ 5 \end{bmatrix}
\end{aligned}
$$



그림으로 표현한 선형 변환
회전하거나 확대 또는 축소 할 수 있음



정방행렬 A가 있고 다음의 식을 만족하는 열벡터 $\vec{x}$(단$x \neq 0$)가 존재할 때 λ를 행렬 A 의 고윳값(eigenvalue)이라고 하고 $\vec{x}$를 고유벡터 (eigenvector)라고 함

$$A\vec{x} = \lambda E\vec{x}$$

$$(A - \lambda E)\vec{x} = 0$$

만약, $(A - \lambda E)$가 역행렬 $(A - \lambda E)^{-1}$ 을 갖는다면, 다음과 같이 양변에 역행렬을 곱해줄 수 있음

$$(A - \lambda E)^{-1}(A - \lambda E)\vec{x} = (A - \lambda E)^{-1}0$$

$$\vec{x} = (A - \lambda E)^{-1}0 = 0$$

단, $\det(A - \lambda E) = 0$

벡터가 회전하지 않고 확대나 축소만 할 때, 변화한 벡터의 길이 비율이 고윳값이고, 벡터의 방향이 고유 벡터가 됨

$$
\begin{aligned}
\det\left(\begin{bmatrix} 2 & 4 \\ -1 & -3 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) &= 0 \\
\det\begin{bmatrix} 2-\lambda & 4 \\ -1 & -3-\lambda \end{bmatrix} &= 0 \\
(2-\lambda)(-3-\lambda) - 4(-1) &= 0 \\
\lambda^2 + \lambda - 2 &= 0 \\
(\lambda+2)(\lambda-1) &= 0
\end{aligned}
$$

case 1 $\lambda = -2$

$$(A-(-2)E)\vec{x} = \begin{bmatrix} 2-(-2) & 4 \\ -1 & -3-(-2) \end{bmatrix}\vec{x} = \begin{bmatrix} 4 & 4 \\ -1 & -1 \end{bmatrix}\vec{x} = 0$$

이 식을 만족하는 해가 고유벡터 $\vec{x}$임.

$\vec{x} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$라고 가정하고 $\begin{bmatrix} 4 & 4 \\ -1 & -1 \end{bmatrix}\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ 을 풀면 $\alpha + \beta = 0$ 이 나옴

상수 t를 가정하고 $\alpha = t$ 그리고 $\beta = -t\begin{bmatrix} 0 \\ 0 \end{bmatrix}$라고 풀어 쓸수 있음.

$\vec{x} = t\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ 이고 x는 $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$의 상수 배가 됨

case 2 $\lambda = 1$

$$(A-(1)E)\vec{x} = \begin{bmatrix} 2-(-1) & 4 \\ -1 & -3-(-1) \end{bmatrix}\vec{x} = \begin{bmatrix} 1 & 4 \\ -1 & -4 \end{bmatrix}\vec{x} = 0$$

이 식을 만족하는 해가 고유벡터 $\vec{x}$임.

$\vec{x} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$라고 가정하고 $\begin{bmatrix} 1 & 4 \\ -1 & -4 \end{bmatrix}\begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ 을 풀면 $\alpha + 4\beta = 0$ 이 나옴

정리하면 $\alpha = -4\beta$ 이고, 이를 풀면 $\alpha = t$ 그리고 $\beta = -\frac{1}{4}t$ 가 됨

$x = t\begin{bmatrix} 1 \\ -\frac{1}{4} \end{bmatrix} = t\frac{1}{4}\begin{bmatrix} 4 \\ -1 \end{bmatrix}$ 이 되어 $\vec{x}$는 $\begin{bmatrix} 4 \\ -1 \end{bmatrix}$ 의 상수 배가 됨

고유값과 이에 대응하는 고유벡터는 각각 두 개씩 있음

인공지능 알고리즘 중 비지도 학습에서 주성분 분석 (PCA, Principal Component Analysis)라는 기법을 쓰는데, 이 때 고윳값과 고유벡터를 활용함. 기여율을 사용하여 각 주성분(고유벡터에 대응하는 고윳값을 전체 고윳값들의 총합으로 나눈 값)이 데이터를 얼마나 잘 설명하는지 평가하는 척도로 사용

확률과 통계

확률 어떤 사건이 우연히 발생할 가능성을 표현한 것. P

$$\text{확률} = \frac{\text{어떤 사건이 발생할 수 있는 경우의 가짓수}}{\text{모든 경우의 가짓수}}$$

조합 combination, 서로 다른 n개로부터 중복 없이 k개를 골라내는 경우의 수

$$_nC_k = \frac{n\cdot(n-1)\cdots(n-k+1)}{1\cdot 2\cdot\ \cdots\ \cdot(k-1)\cdot k}$$

예: 트럼프 카드에서 다섯 장의 카드를 동시에 뽑았을 때, 다섯 장 모두가 하트인 경우는 몇 가지인가. $_{13}C_5$

여사건 complimentary event, 사건 A가 발생할 확률이 P일 때, 사건 A의 여사건, $\bar{A}$가 발생할 확률

$$P(\bar{A}) = 1 - P$$

예: 트럼프 카드에서 4장의 카드를 동시에 뽑았을 때, 적어도 1장이 스페이드인 경우의 확률. $1 - \frac{{}_{39}C_4}{{}_{52}C_4}$

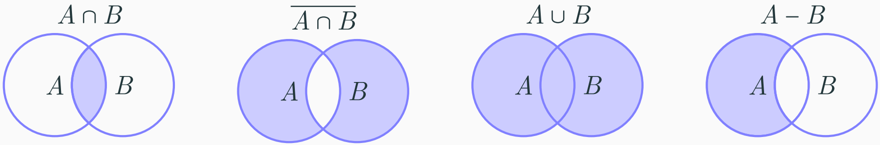
확률 iii

합사건 사건 A와 사건 B가 동시에 발생하는 사건: $A \cap B$ (A and B)

$$P(A \cap B) = P(A)P(B)$$

곱사건 사건 A와 사건 B중에서 어느 한쪽이 발생할 사건 $A \cup B$ (A or B)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



확률변수 random variable, 어떤 변수 X를 사용할 때 확률 $P(X)$의 확률로 구할 수 있다면 X는 확률변수임. 예: 주사위를 두 번 던져 합이 2가 될 확률 $P(X_2 = 2) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$

이산확률변수 discrete random variable, 확률변수 중에 연속되지 않고 셀 수 있을 만큼 흩어진 경우를 말함. 예: 주사위에 적힌 숫자, 사건이 일어나는 시행 횟수

연속확률변수 continuous random variable, 값이 특정 범위 내에서 실수 형태로 존재하며, 소수점 이하까지 내려가는 경우. 예: 몸무게, 경과 시간과 같이 끊김 없이 연속적으로 이어지는 값

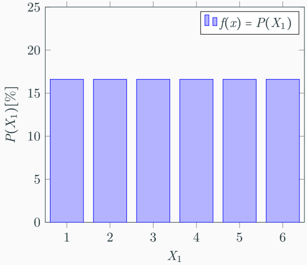
이산확률분포 discrete probability distribution, 어떤 사건의 이산확률변수가 X일 때, 그에 대한 확률 P는 이산확률분포 $f(x)$를 따름

$$P(X) = f(x)$$

주사위를 두 개 던졌을 때 숫자의 합과 그에 대한 확률

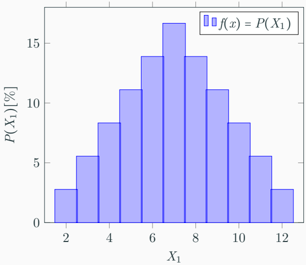
X_2	2	3	4	5	6	7	8	9	10	11	12
$P(X_2)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

주사위가 하나일 때 히스토그램



균등분포

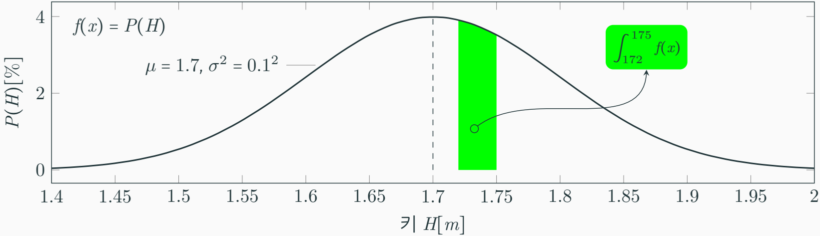
주사위가 두 개일 때 히스토그램



종형 곡선 (bell curve)

연속확률분포 어떤 사건의 연속확률변수가 X일 때, 그에 대한 확률 P는 연속확률분포 $f(x)$를 지정한 X의 구간 안에서 적분한 값과 같음

$$P(a \leq X \leq b) = \int_a^b f(x)\mathrm{d}x$$



연속확률분포 종류: 정규분포 (normal distribution), 지수분포 (exponential distribution), 스튜던트 t 분포 (student's t-distribution), 파레토 분포 (Pareto distribution), 로지스틱 분포 (logistic distribution)

결합확률과 조건부확률

결합확률 사건 A와 사건 B가 서로 독립된 사건일 때, 두 사건의 결합확률(동시에 일어날 확률)은 다음과 같음

$$P(A \cap B) = P(A, B) = P(A)P(B)$$

조건부 확률 사건 B가 일어났을 때, 사건 A가 일어날 조건부 확률은 다음과 같음

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

$$P(A \cap C) = P(A) \cdot P(\bar{B})$$

$$P(C) = P(A \cap \bar{B}) + P(\bar{A} \cap B)$$

기댓값 모든 이산확률변수 X에 대한 기댓값 $E(X)$는 다음과 같음. 이때, 확률은 $P(X)$

$$E(X) = \sum P(X) \cdot X$$

X와 Y가 서로 독립된 확률변수이고 k는 상수라고 할 때 다음 식이 성립함

1. $E(k) = k$: 상수의 기댓값은 상수
2. $E(kX) = kE(X)$: 확률변수를 상수 배하면 기댓값도 상수배가 됨
3. $E(X + Y) = E(X) + E(Y)$: 확률변수의 합의 기댓값은 각 기댓값의 함과 같음
4. X 와 Y 가 서로 독립일 때 $E(XY) = E(X) \cdot E(Y)$: 독립적인 확률변수의 곱에 대한 기댓값은 각 기댓값의 곱과 같음

평균값 n개의 확률변수가 각각 $x_1, x_2, \ldots, x_n$이라는 값을 가질 때 평균값 $\bar{x}$는 다음과 같음

$$\bar{x} = \sum_{k_1}^{n} \frac{1}{n} \cdot x_k = \frac{1}{n} \sum_{k_1}^{n} x_k$$

분산 n개의 확률변수가 각각 $x_1, x_2, \ldots, x_n$ 이라는 값을 가지고 평균값이 $\bar{x}$일 때 분산 σ^2 은 다음과 같음

$$\sigma^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - \bar{x})^2$$

표준편차 편차는 $+$와 $-$ 같은 부호를 갖고 있음. 평균을 기준으로 떨어진 정도를 나타냄. σ는 다음과 같음.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n} \sum_{k=1}^{n} (x_k - \bar{x})^2}$$

평균과 분산 그리고 공분산 ii

공분산 두 확률변수의 상관관계를 확인할 때 사용. 두 가지 데이터에 대한 n조의 확률변수 $(X, Y) = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$이 있다고 가정함. X의 평균이 μ_x이고 Y의 평균이 μ_y라고 할 때 공분산은 $Cov(X, Y)$ 다음과 같음

$$Cov(X, Y) = \frac{1}{n} \sum_{k=1}^{n} (x_k - \mu_x)(y_k - \mu_y)$$

상관계수

상관계수 확률변수 X와 Y의 분산이 양수이고 각각의 표준편차가 σ_X, σ_Y 공분산이 σ_{XY}라고할 때의 상관계수는 다음과 같음 (이때, $-1 \le \rho \le 1$)

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

공분산을 상관계수로 변환하였기 때문에 상관계수 끼리 강약을 비교할 수 있음

최대가능도추정 maximum likelihood estimation. 가장 그럴듯하게 값을 추정.
최대가능도추정이란 어떤 파라미터 θ의 값을 추정하는 방법이며, θ에 대한 가능도 함수 $L(\theta)$ 을 최대로 만드는 θ을 찾는 것. 이때의 θ에 대한 추정값은 다음 방정식을 만족함

$$\frac{\mathrm{d}L(\theta)}{\mathrm{d}\theta} = 0$$

이산확률분포의 식은 확률의 곱으로 표현이 되기 때문에 미분이 쉽지 않음. 가능도함수에 자연로그를 붙여 주어 로그가능도함수 $\log_e L(\theta)$를 만들면 됨

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log_e L(\theta) = 0$$

선형회귀

선형모델의 변수

목적변수 objective variable 추정하고 싶은 값

종속변수 dependent variable 추정하고 싶은 값

설명변수 explanatory variable 추정하는 데 필요한 정보

독립변수 independent variable 추정하는 데 필요한 정보

데이터 척도

분류	척도	설명
질적 데이터	명목척도 (nominal scale)	분류나 구별을 하기 위한 척도. 더미 변수라고도 함 (예: 남성 0, 여성 1)
	서열척도 (ordinal scale)	대소 관계만 의미가 있는 척도 (예: 나쁨: 0, 보통: 1, 좋음: 2)
양적 데이터	등간척도 (interval scale)	간격에 의미가 있는 변수. 덧셈, 뺄셈만 의미가 있음 (예: 서기)
	비율척도 (ratio scale)	비례에도 의미가 있는 변수. 덧셈, 뺄셈, 곱셈, 나눗셈 전체에 의미가 있음. (예: 속도, 키, 체중)

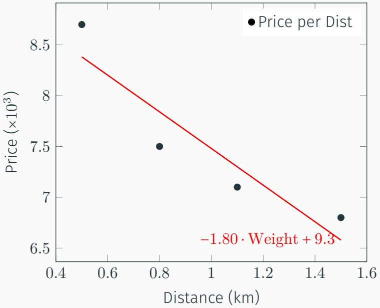
선형회귀 모델

회귀 모델 하나의 목적변수(종속변수)를 하나 이상의 설명변수(독립변수)로 기술한 관계식. 계수(가중치)는 일반적으로 $w_0, w_1, \ldots, w_n$로 표현하고 설명변수는 일반적으로 $x_1, x_2, \ldots, x_n$로 표현함

$$
\begin{aligned}
y &= w_0 + \sum_{k=1}^{l} w_k x_k \\
y &= w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_n x_n \\
\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} &= \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1l} \\ 1 & x_{21} & x_{22} & \cdots & x_{2l} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nl} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} \\
y &= XW
\end{aligned}
$$

최소제곱법 least squared method 수치 데이터들을 1차함수와 같은 특정 함수를 사용하여 근사적으로 표현하는 방법. 수치 데이터 값과 함수의 결과값 사이에 오차가 최소가 되도록 하는 것. 이 과정에서 오차가 가장 작게 나오는 가중치를 찾으면 이를 모델식의 계소로 사용함.

실제 데이터와 추세선 간의 거리의 차



$$
\begin{aligned}
D &= \sum_{l=1}^{29} \left| y_l - (w_0 + w_1 x_l) \right| \\
&= \sum_{l=1}^{29} \left(y_l - (w_0 + w_1 x_l) \right)^2
\end{aligned}
$$

최소제곱법으로 파라미터 도출 ii

$$
\begin{aligned}
D &= \sum_{l=1}^{29}(y_l - (w_0 + w_1 x_l))^2 \\
D &= (8.7 - (w_0 + 0.5w_1))^2 + (7.5 - (w_0 + 0.8w_1))^2 \\
&\quad +(7.1 - (w_0 + 1.1w_1))^2 + (6.8 - (w_0 + 1.5w_1))^2 \\
D &= 4w_0^2 + 4.35w_1^2 + 7.8w_0w_1 - 60.2w_0 - 56.72w_1 + 228.59
\end{aligned}
$$

이때 D가 최솟값을 가지려면 w_0과 w_1로 편미분했을 때 값이 0이 되어야 함. 그러므로 다음과 같은 식을 만들 수 있음

$$\begin{aligned}\frac{\delta D}{\delta w_0} &= 8w_0 + 7.8w_1 - 60.2 = 0\\ \frac{\delta D}{\delta w_1} &= 8.7w_0 + 7.8w_1 - 56.72 = 0\end{aligned}$$

연립방정식을 풀면

$$-1.8037x + 9.2836$$

일반적으로는 목적변수 Y, 설명변수 $X_1, X_2, \ldots, X_l$ 그리고 모델식 $f(X_1, X_2, \ldots, X_l)$ 이라고 할 때 최소제곱법을 적용하는 과정은 오차의 제곱합 D를 최소화하는 $f(X_1, X_2, \ldots, X_l)$를 구하는 문제임.

n개의 데이터 세트에서 k번째의 데이터를 $(x_k1, x_k2, \ldots, x_kl, y_k)$ 오차의 제곱합은 다음과 같은 식으로 표현할 수 있음.

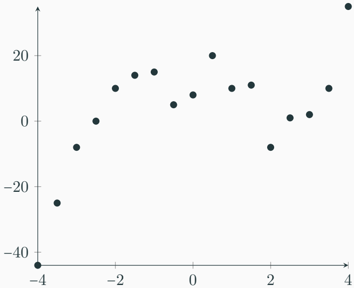
$$D = \sum_{k=1}^{n} (y_k - f(x_{k1}, x_{k2}, \cdots, x_{kl})^2$$

모델 식은 다음과 같음.

$$f(x_k1, x_k2, \ldots, x_kl) = \sum_{m=1}^{l} w_m x_{km} + w_0$$

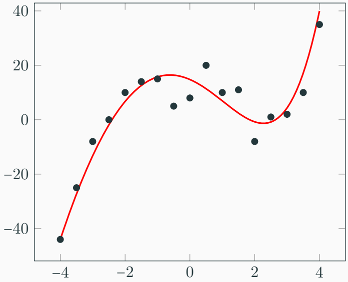
가중치$(w_0, w_1, \ldots, w_l)$를 변화시키면서 함수 $D(w_0, w_1, \ldots, w_l)$d이 최소가 되는 값을 만드는 가중치의 조합을 찾기!! 각 가중치의 편미분이 0이 되는 해 찾기

샘플 데이터

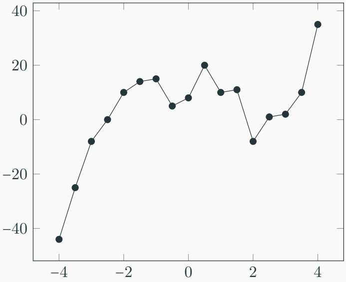


정규화로 과학습 줄이기 ii

일반화 능력 generalization ability, 약간의 노이즈는 허용하면서 전체적인 데이터의 특성은 잘 반영한 식



과학습/과적합 overfitted, 주어진 데이터에 너무 정확히 들어맞아 지나치게 복잡하게 표현된 상태



정규화 선형회귀에서 과학습을 피하는 방법. 모델이 복잡해질 수록 일종의 패널티를 적용하여 과학습을 억제

정규화 방법: $L1$ 정규화, $L2$ 정규화, Norm과 같음

선형회귀 모델 : $y = w_0 + \sum_{k=1}^{l} w_k x_k$

정규화를 위해 추가할 항: $\lambda E(w)$

정규화로 과학습 줄이기 iv

$L1$ 정규화에서는 파라미터의 $L1$ norm에 계수를 곱한 다음과 같은 항을 사용

$$\lambda E(w) = \sum_{k=1}^{l} |w_k|$$

최소제곱오차 식 $D = \sum_{k=1}^{n} (y_k - f(x_{k1}, x_{k2}, \ldots, x_{kl}))^2$ 에 추가

Lasso 회귀

$$D = \sum_{k=1}^{n} (y_k - f(x_{k1}, x_{k2}, \ldots, x_{kl}))^2 + \lambda E(w)$$

$L2$ 정규화에서는 파라미터의 $L2$ norm에 계수를 곱한 다음과 같은 항을 사용

$$\lambda E(w) = \sum_{k=1}^{l} |w_k^2|$$

그리고 $L1$ 정규화를 할 때와 같이 최소제곱오차를 구하는 식에 항을 추가

$$D = \sum_{k=1}^{n} (y_k - f(x_{k1}, x_{k2}, \ldots, x_{kl}))^2 + \lambda E(w)$$

$L2$ 정규화에서는 $L1$정규화와 달리 절댓값을 사용하지 않음. 미분이 상대적으로 쉬움. 이렇게 정규화된 선형회귀를 Ridge 회귀라고 함

$L1$ 정규화는 $L2$정규화는 기존의 모델식에 조합해서 쓸 수 있고 이 둘을 조합할 수 도 있음. 그렇게 만들어진 회귀 모델을 Elastic Net이라고 함

정규화로 과학습 줄이기 vi

scikit-learn에서는 특별히 지정하지 않은 한 기본적으로 $\lambda = 1.0$으로 계산함. 정규화 강약을 조정하려면 λ를 조절

모델 평가 i

모델을 검증하기 위한 데이터 세트를 만드는 방법

홀드아웃 교차 검증법 holdout cross validation, 하나의 데이터 세트를 학습용 데이터와 테스트용 데이터 두 가지로 나누는 방법

k-분할 교차 검증법 k-fold cross validation, 데이터 세트를 k개로 분할한 다음, k번에 걸쳐 학습 데이터와 테스트 데이터의 조합을 바꿔쓰는 방법

모델의 성능은 시각화를 통해 눈으로도 확인할 수 있는 잔차 (residual) 그래프에 표시

잔차 추정된 회귀식과 실제 데이터 사이의 차이

회귀식을 $y = w_0 + \sum_{k=1}^{l} w_k x_k$라고 i번째의 잔차를 e_i라고 할 때 잔차를 구하는 식은 다음과 같음

$$e_i = y_i - \left(w_0 + \sum_{k=1}^{l} w_k x_{ki} \right)$$

n개의 데이터가 있을 때 평균제곱오차와 결정계수는 다음과 같은 식으로 구할 수 있음

$$
\begin{aligned}
MSE &= \frac{1}{n}\sum_{i=1}^{n}(y_{\text{실측값}i} - y_{\text{예측값}i})^2 \\
R^2 &= 1 - \frac{MSE}{\frac{1}{n}\sum_{i=1}^{n}(y_{\text{실측값}i} - \bar{y}_{\text{실측값}i})^2}, (0 \le R^2 \le 1)
\end{aligned}
$$

R^2 의 분모는 y의 분산이고, $\bar{y}_{\text{실측값}i}$은 $y_{\text{실측값}i}$의 평균을 뜻함. 이 지표는 0 이상 1 이하의 값을 갖고 있으며 1에 가까울 수록 잘 맞는 모델임.

Classification

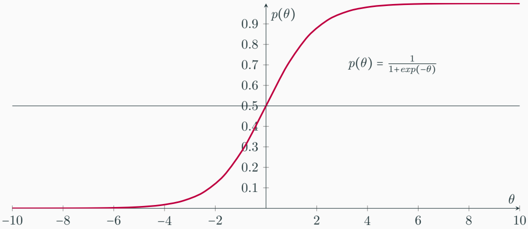
로지스틱 회귀 Logistic regression, 1이 될 확률 p이고 0이 될 확률이 $1-p$인 이산확률분포를 사용하여 1이나 0의 값을 확률적으로 얻는 방법. 이산확률분포로는 베르누이 분포 (Bernoulli distribution) 등이 있음

x가 실수 입력값이고 $y = 0, 1$이 출력값일 때, 출력값은 반드시 0과 1중 하나가 나옴. 이때, 출력값 $y = 1$이 되는 조건부확률 $p(y = 1|x; \theta)$는 다음과 같음. 이때 θ는 실수 파라미터임

$$p(y = 1|x; \theta) = \frac{1}{1 + exp(-\theta^T x)}$$

로지스틱 회귀 ii

$p(\theta) = \frac{1}{1+exp(-\theta)}$ 를 로지스틱 함수라고 함 치역은 0에서 1이고 평균 값인 0.5가 되는 함수. 시그모이드 함수로 소개됨



이 함수의 정의역은 실수 전체로 $\theta > 0$일 때 $y = 1$이 될 확률 $p(\theta)$가 0.5 보다 커지는 특징이 있음. 이 특징을 사용하여 어떤 대상을 분류할 때 사용함

학습데이터로 쓸 데이터 세트 (x_i, y_i)가 $1 \le i \le m$만큼 주어졌다고 가정할 때 다음과 같은 식이 성립함

$$p_i(y = y_i | x_i; \theta) \ge 0.5\text{일때, } y_i = 1$$
$$p_i(y = y_i | x_i; \theta) < 0.5\text{일때, } y_i = 0$$

확률표현을 $p_i(y = y_i|x_i;\theta)$와 같이 간단히 표현한 후, 목적 함수 $J(\theta)$를 최소로 만드는 θ를 구한다고 할 때, 다음과 같은 목적 함수를 만들 수 있음

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(p_{x_i} - y_i)^2$$

일반적으로는 이 식을 최소화하는 θ를 구하기 위해 전개해야 함. 이 경우는 $0 \le p_{x_i} \le 1$이고, y_i는 0 또는 1만 나온다는 사실을 알고 있기 때문에 다음과 같은 손실 함수 $L(\theta)$로 바꿔서 계산하는 것이 효과적임

$$L(\theta) = -\sum_{i=1}^{m}(y_i log(p_{x_i}) + (1 - y_i)\log(1 - px_i))$$

$L2$ 정규화를 적용한 로그 목적함수는 다음과 같음. 이때 λ는 정규화의 강도를 나타내는 파라미터임

$$L(\theta) = -\sum_{i=1}^{m}(y_i log(p_{x_i}) + (1 - y_i)\log(1 - px_i)) + \frac{1}{2\lambda}\sum_{j=1}^{n}\theta_j^2$$

다중 클래스 분류 (multi-class classification)의 경우 one-vs-rest 기법을 사용하여 여러 개의 이진 클래스 분류기 문제로 나눠서 해결함.
$(y = i, y \neq i)$

정밀도, 재현율, F값

		예측 결과	
		Positive	False
실제 결과	True	진양성, True positive, TP	위음성, False negative, FN
	거짓	위양성, False Positive, FP	진음성, True negative, TN

정밀도

$$\text{Precision} = \frac{TP}{TP + FP}$$

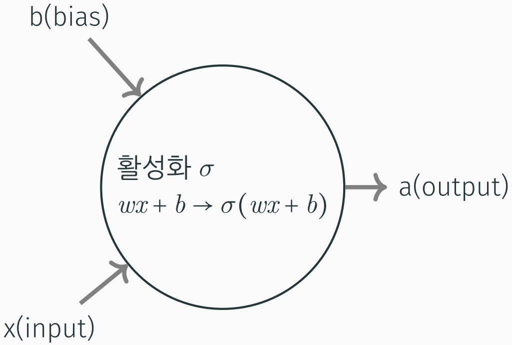
재현율

$$\text{Recall} = \frac{TP}{TP + FN}$$

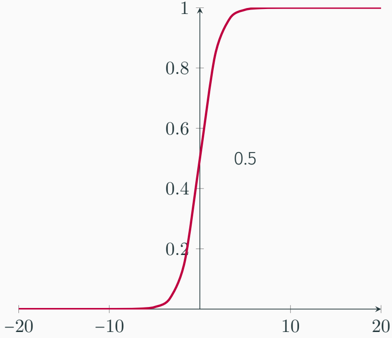
F 값

$$F = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} = \frac{2TP}{2TP + FN + FP}$$

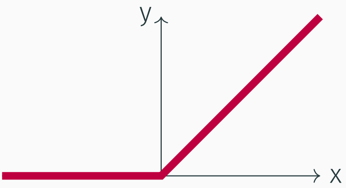
Neural Networks


b(bias)
활성화 σ
$wx + b \to \sigma(wx + b)$
a(output)
x(input)

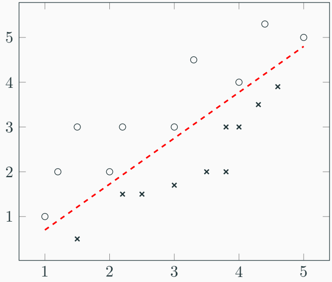
$$\varsigma_a(x) = \frac{1}{1 + \exp(-ax)}$$



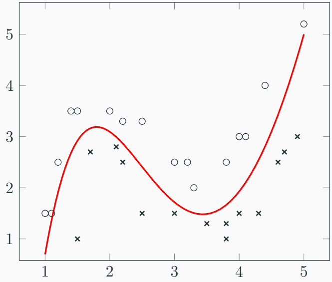
$$\varphi(x) = \max(0, x) = \begin{cases} x & (x > 0) \\ 0 & (x \leq 0) \end{cases}$$



선형 분리가 가능한 경우

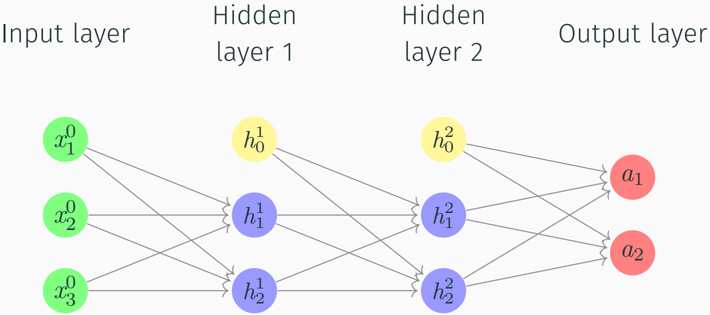


선형 분리가 불가능한 경우



비선형 변환 활성화 함수를 사용하여 분리 할 수 있도록 만들어야 함

순전파 i



입력값	출력값	가중치	바이어스	활성화 함수
x_β^α	a_β^α	$w_{\beta\gamma}^\alpha$	b_β^α	σ_α
α : 계층번호 β : 노드번호	α : 계층번호 β : 노드번호	α : 다음 계층번호 β : 다음 계층 노드번호 γ : 이전 계층의 노드번호	α : 다음 계층번호 β : 다음 계층 노드번호	α : 계층번호

최초의 입력층에서는 별다른 처리를 하지 않음

$$x_1^0 = a_1^0, x_2^0 = a_2^0$$

은닉층의 입력값은 다음과 같이 표현됨

$$\begin{aligned} x_1^1 &= w_{11}^1 a_1^0 + w_{12}^1 a_2^0 + w_{13}^1 a_3^0 + b_1^1 \\ x_2^1 &= w_{21}^1 a_1^0 + w_{22}^1 a_2^0 + w_{23}^1 a_3^0 + b_2^1 \end{aligned}$$

행렬로 표현하면 다음과 같음

$$x^1 = \begin{pmatrix} x_1^1 \\ x_2^1 \end{pmatrix}, \qquad W^1 = \begin{pmatrix} w_{11}^1 & w_{12}^1 & w_{13}^1 \\ w_{21}^1 & w_{22}^1 & w_{23}^1 \end{pmatrix}, \qquad a^0 = \begin{pmatrix} a_1^0 \\ a_2^0 \\ a_3^0 \end{pmatrix}, \qquad b^1 = \begin{pmatrix} b_1^1 \\ b_2^1 \end{pmatrix}$$

$$x^1 = W^1 a^0 + b^1$$

은닉층의 출력값 (σ 함수는 시그모이드 함수로 가정)

$$\begin{aligned} a_1^1 &= \sigma_1(x_1^1) \\ a_2^1 &= \sigma_1(x_2^1) \end{aligned}$$

예를 들어 $x_1^1 = 0$ 일때 시그모이드 함수를 사용하면 $a_1^1 = \sigma_1(0) = 0.5$ 가 됨

은닉층에서 출력층으로 정보 전달

$$x^2 = \begin{pmatrix} x_1^2 \\ x_2^2 \\ x_3^2 \end{pmatrix}, \qquad W^2 = \begin{pmatrix} w_{11}^2 & w_{12}^2 \\ w_{21}^2 & w_{22}^2 \\ w_{31}^2 & w_{32}^2 \end{pmatrix}, \qquad a^1 = \begin{pmatrix} a_1^1 \\ a_2^1 \end{pmatrix}, \qquad b^2 = \begin{pmatrix} b_1^2 \\ b_2^2 \\ b_3^2 \end{pmatrix}$$

$$x^2 = W^2 a^1 + b^2$$

softmax 함수 n차원의 실수 벡터 $\vec{x} = (x_1, x_2, \ldots, x_n)$이 있다고 가정할 때, 다음 식에서 n 차원의 실수 벡터 $\vec{y} = (y_1, y_2, \ldots, y_n)$을 결괏값으로 내는 함수를 softmax함수라고 부름

$$y_i = \frac{\exp(x_i)}{\exp(x_1) + \exp(x_2) + \cdots + \exp(x_n)} \qquad (1 \le i \le n)$$

softmax를 사용하면 결괏값을 확률적인 표현으로 만들 수 있음

2계층에서 적용된 softmax 함수를 σ_2라고 할 때, 이 함수의 결과로 나오는 확률 a_1^2, a_2^2, a_3^2는 다음과 같이 표현됨

$$a_1^2 = \sigma_2(x_1^2), \qquad a_2^2 = \sigma_2(x_2^2), \qquad a_3^2 = \sigma_2(x_3^2)$$

a_1^2, a_2^2, a_3^2 중에서 가장 큰 값이 나오는 카테고리가 판별의 결과

가중치와 바이어스를 조정할 때는 손실 함수의 값이 최소가 되도록 만들어야 함

손실함수 신경망이 출력한 값과 실제 값과의 오차에 대한 함수

다양한 손실함수가 있음. 그 중 평균제곱오차 (MSE, mean squred error) E는 다음과 같이 구함. 이때, t는 정답 레이블 y는 신경망의 출력

$$E = \frac{1}{2}\|t - y\|^2$$

예시

3계층 신경망에서 2계층의 출력 y^2가 $y^2 = (0.1w, 0.5w, 1 - 0.6w)$이고 정답 t는 $t = (0, 1, 0)$이라고 할 때 평균 제곱오차 E를 구하고 이 값을 최소로 만드는 w를 구하시오

$$
\begin{aligned}
E &= \frac{1}{2}((0-0.1w)^2 + (1-0.5w)^2 + (0-1(1-0.6w))^2) \\
E &= 0.31w^2 - 1.1w + 1
\end{aligned}
$$

평균제곱오차를 최소화시키려면 w로 미분했을 때의 값이 0이 되도록 해야 함 (답 $w \approx 1.774$)

$$\frac{\mathrm{d}E}{\mathrm{d}w} = 0.62w - 1.1 = 0$$

손실 함수 iii

softmax의 결과 예시 2

	0	1	2	3	4	5	6	7	8
출력층 y	0.01	0.02	0.05	0.02	0.67	0.13	0.05	0.01	0.01
레이블 t	0	0	0	0	1	0	0	0	0

이때의 평균제곱오차는 다음과 같이 구함

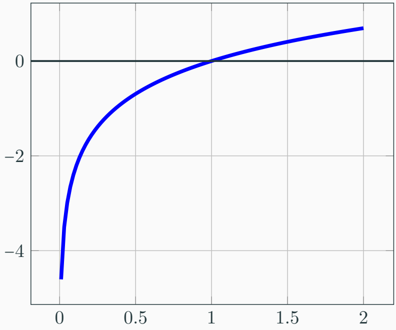
(정답인 경우 1 그 외는 0으로 표기하는 one-hot 기법 사용)

$$\begin{aligned} E &= \frac{1}{2}((0-0.01)^2+(0-0.02)^2+(0-0.05)^2+(0-0.02)^2+(1-0.67)^2 \\ &\quad +(0-0.13)^2+(0-0.05)^2+(0-0.01)^2+(0-0.01)^2+(0-0.03)^2) \\ &= 0.0664 \end{aligned}$$

교차 엔트로피

$$E = -\sum t \log_e y$$

평균제곱오차에서 사용한 one-hot 표현법은 0 과 1 뿐이 없기 때문에 교차 엔트로피를 손실함수로 사용함

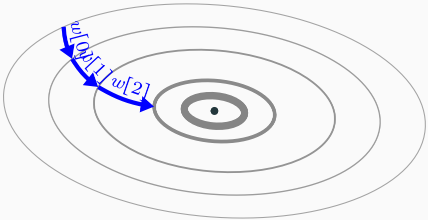


그래프를 보면 y가 1일 때 $E = 0$이 나오고 y가 0에 가까워질수록 E의 값은 0보다 작아짐. 0에 가까워질 수록 음수가 나오는데 이것을 양수로 만들기 위해 교차 엔트로피의 손실 함수에 마이너스 부호를 붙임.

출력층의 활성화 함수로 softmax함수를 사용할 때 출력값은 $0 \leq y \leq 1$과 같은 확률 표현됨. 결과 y는 1을 넘지 않으므로 $\log_e y > 0$이 되지 않음

입력층, 은닉층 출력층이 많아질 수록 평균제곱오차를 최소화하기 위해 고려해야 하는 변수의 수가 기하급수적으로 많아짐 (예: 입력층 3, 은닉층 2, 출력층 3개 일때 변수 17개)

경사하강법 함수의 그래프를 따라 움직이면서 기울기를 조사하고 이때 구한 기울기의 값이 작아지는 방향으로 조금씩 이동하는 방법



경사하강법 수식

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

$h = \Delta x$라고 가정하면

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x} = \lim_{\Delta\to 0}\frac{f(x+\Delta x)-f(x)}{\Delta x}$$

Δx 가 충분히 작은 값이라고 한다면 $\lim_{\Delta\to 0}$ 에 따라서 다음과 같은 근사식이 됨

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x} \approx \lim_{\Delta\to 0}\frac{f(x+\Delta x)-f(x)}{\Delta x}$$

양변에 Δx을 곱해 전개하면 다음과 같이 됨 (근사공식)

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x}\Delta x \approx \lim_{\Delta\to 0} f(x+\Delta x)-f(x)$$

경사하강법 iii

함수 $f(x)$에 대해 x에 Δ만큼의 변화를 주었을 때 $f(x)$의 변화량을 Δf이라고 하면 다음과 같이 정리할 수 있음

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

앞의 식에 대입하면 다음과 같이 정리됨

$$\Delta f(x) \approx \frac{\mathrm{d}f(x)}{\mathrm{d}x} \Delta x$$

원하는 것은 $\Delta f(x)$가 음이 되는 방향으로 Δx를 조금씩 이동하며 최솟값을 찾아야 함.

최솟값을 찾을 조건

- 접선의 기울기 ($\frac{\mathrm{d}f(x)}{\mathrm{d}x}\Delta x$)가 음일때 Δx가 양이라면 $\Delta f(x)$는 음
- 접선의 기울기 ($\frac{\mathrm{d}f(x)}{\mathrm{d}x}\Delta x$)가 양일때 Δx가 음이라면 $\Delta f(x)$는 음

다음과 같은 수식이 성립한다면 함수의 최솟값을 찾기 위해 그래프를 타고 내려갈 수 있음

$$\Delta x = -\eta \frac{\mathrm{d}f(x)}{\mathrm{d}x}$$

이때 η를 학습률(learning rate)이라고 부르고 너무 크거나 너무 작으면 최솟값에 이르지 못할 수 있음

이동하기 전후의 위치를 x_{old}, x_{new}라고 하면 다음과 같이 표현할 수 있음 (갱신식)

$$\Delta x = x_{new} - x_{old}$$

앞의 식에 대입하면 다음과 같음

$$x_{new} = x_{old} - \eta \frac{\mathrm{d}f(x)}{\mathrm{d}x}$$

다변수 함수인 경우의 경사하강법 식은 다음과 같이 정리됨 (우변의 괄호안을 손실 함수 E의 기울기라고 표현함)

$$E = -\eta \left(\frac{\partial E}{\partial w_{11}^1}, \frac{\partial E}{\partial w_{21}^1}, \frac{\partial E}{\partial w_{31}^1}, \cdots \right)$$

경사하강법 vi

경사하강법을 사용하여 모든 학습 데이터의 오차 계산을 해야하면 학습시간이 너무 오래 걸리는 문제가 있음.

확률적 경사하강법을 사용하여 이 문제를 해결함.

학습데이터 중 N의 데이터를 골라 학습시킨 후 그 결과로 나온 손 함수에 경사하강법을 적용하여 가중치를 구하는 방법

이 과정을 반복하면 N개의 데이터마다 가중치를 갱신할 수 있게 되는데, 이때 처리하는 N개의 데이터 개수를 배치 사이즈라고 함

학습 데이터를 몇 차례 다시 사용하면서 정확도를 높일 수 있는데, 이때 반복하는 횟수를 에포크(epoch)라고 함

오차역전파법 i

- 손실 함수의 기울기를 구하는 것은 쉽지 않은 일임. 가중치나 바이어스 변수가 너무 많고, 이를 미분할 때도 계산량이 너무 많기 때문임
- 손실 함수의 기울기를 좀 더 쉽게 구할 수 있는 방법이 필요함
- 오차역전파법이 이를 개선하기 위해 등장

순전파방식
입력값에 가중치를 곱하고, 그 값을 다음 계층으로 전달하는 과정을 반복하는 방식

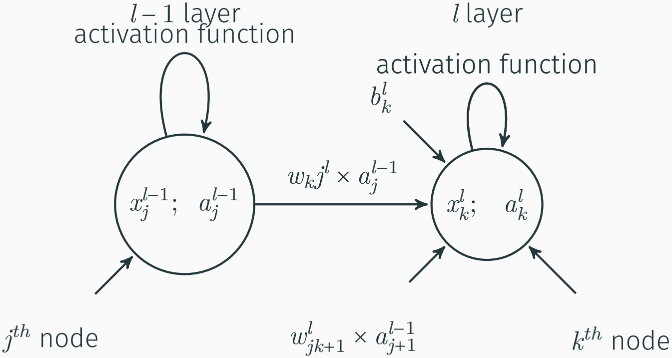
역전파방식
출력값과 정답 사이의 오차를 먼저 구한 후, 그 정보를 바탕으로 바로 직전 단계의 계층의 가중치와 바이어스를 조정하는 방식

손실함수는 평균제곱오차 식을 사용하고 뉵층의 활성화 함수로 표준 시그모이드 함수를 사용

우리의 목표
함수의 기울기를 구하기 위해 다변수 함수의 기울기$\left(\frac{\partial E}{\partial w_{11}^1}, \frac{\partial E}{\partial w_{21}^1}, \frac{\partial E}{\partial w_{31}^1}, \cdots\right)$를 최소화하는 것

일반화하면 $\frac{\partial E}{\partial w_{kj}^l}$ 을 최소화하는 것

개념 도식


$l-1$ layer
activation function
l layer
activation function
b_k^l
$w_kj^l \times a_j^{l-1}$
$x_j^{l-1}; \quad a_j^{l-1}$
$x_k^l; \quad a_k^l$
j^{th} node
$w_{jk+1}^l \times a_{j+1}^{l-1}$
k^{th} node

미분의 연쇄법칙 적용

$$\frac{\partial E}{\partial w_{kj}^{l}} = \frac{\partial E}{\partial x_k^l}\frac{\partial x_k^l}{\partial w_{kj}^{l}}$$

x_k^l 을 풀어 쓰면 다음과 같음

$$x_k^l = w_{kl}^l a_1^{l-1} + w_{k2}^l a_2^{l-1} + \cdots + w_{kj}^l a_j^{l-1} + \cdots + b_k^l$$

x_k^l 를 w_{kl}^l로 미분하면 다음과 같은 식을 얻음

$$\frac{\partial x_k^l}{\partial w_{kj}^{l}} = a_j^{l-1}$$

앞의 식에 적용하면 다음과 같은 결과를 얻음

$$\frac{\partial E}{\partial w_{kj}^{l}} = \frac{\partial E}{\partial x_k^l} a_j^{l-1}$$

a_j^{l-1} 는 직전 계층의 출력이기 때문에 쉽게 얻을 수 있음.

이해를 위해 오차 δ_k^l 를 $\delta_k^l = \frac{\partial E}{\delta x_k^l}$ 로 바꿔서 다음과 같이 표현

$$\frac{\partial E}{\delta x_k^l} = \delta_k^l a_j^{l-1}$$

δ_k^l 는 계층에 따라 구하는 방법이 다름

- case1: 마지막 계층 일때
- case2: 마지막 계층이 아닐 때

마지막 계층일 때 δ_k^l i

혼돈을 막기 위해 l을 L로 바꿈 $\rightarrow \delta_k^L$

$$\delta_k^L = \frac{\partial E}{\delta x_k^L}$$

연쇄법칙을 적용

$$\delta_k^L = \frac{\partial E}{\delta x_k^L} = \frac{\partial E}{\partial a_k^L}\frac{\partial a_k^L}{\delta x_k^L}$$

$\frac{\partial a_k^L}{\delta x_k^L}$ 는 다음과 같이 표현할 수 있음

$$\frac{\partial a_k^L}{\delta x_k^L} = \frac{\partial \varsigma(x_k^L)}{\partial x_k^L} = \varsigma'(x_k^L)$$

마지막 계층일 때 δ_k^l ii

$\frac{\partial E}{\partial a_k^L}$ 는 다음과 같이 표현할 수 있음

$$\frac{\partial E}{\partial a_k^L} = \frac{\partial \frac{1}{2}(a_k^L - y_k)^2}{\partial a_K^L} = (a_k^L - y_k)$$

정리하면 다음과 같은 식을 만들 수 있음

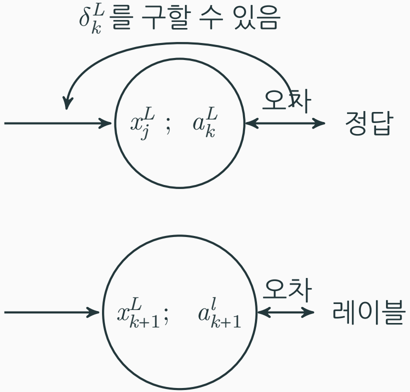
$$\begin{aligned} \delta_k^L &= \frac{\partial E}{\partial x_k^L} = \frac{\partial E}{\partial a_k^L}\frac{\partial a_k^L}{\delta x_k^L} \\ &= \frac{\partial \frac{1}{2}(a_k^L - y_k)^2}{\partial a_K^L}\varsigma'(x_k^L) \\ &= (a_k^L - y_k)\varsigma'(x_k^L) \end{aligned}$$

마지막 계층일 때 δ_k^l iii

a_k^L는 마지막 계층일 때의 출력, y_k는 마지막 계층일 때의 정답 레이블, 그리고 $\varsigma'(x_k^L)$ 는 마지막 계층일 때의 입력을 활성화 함수에 대입 후 미분한 것

마지막 계층일 때 δ_k^l iv

Case 1: 마지막 계층일 때의 δ_k^L 의 도식화



δ_k^l를 구하는 방법,

다음 식의 값을 구하는 것이 목표임: $\delta_k^l = \frac{\partial E}{\partial x_k^l}$

$$\delta_1^2 = \frac{\partial E}{\partial x_1^3}\frac{\partial x_1^3}{\partial a_1^2}\frac{\partial a_1^2}{\partial x_1^2} + \frac{\partial E}{\partial x_2^3}\frac{\partial x_2^3}{\partial a_1^2}\frac{\partial a_1^2}{\partial x_1^2} + \frac{\partial E}{\partial x_3^3}\frac{\partial x_3^3}{\partial a_1^2}\frac{\partial a_1^2}{\partial x_1^2}$$

각 항에 나오는 첫 번째 부분

$\frac{\partial E}{\partial x_1^3}$ 은 노드의 오차 δ_j^l의 정에 의해서 $\frac{\partial E}{\partial x_1^3} = \delta_1^3$과 같이 표현할 수 있음

$$\frac{\partial E}{\partial x_1^3} = \delta_1^3, \qquad \frac{\partial E}{\partial x_2^3} = \delta_2^3, \qquad \frac{\partial E}{\partial x_3^3} = \delta_3^3$$

각 항에 나오는 두 번째 부분 $\frac{\partial x_1^3}{\partial a_1^2}$ 은 $x_1^3 = a_1^2 w_{11}^3 + a_2^2 w_{12}^3 + b_1^3$ 를 적용하면 $\frac{\partial x_1^3}{\partial a_1^2} = w_{11}^3$ 과 같이 표현할 수 있음

$$\frac{\partial x_1^3}{\partial a_1^2} = w_{11}^3, \qquad \frac{\partial x_2^3}{\partial a_1^2} = w_{21}^3, \qquad \frac{\partial x_3^3}{\partial a_1^2} = w_{31}^3$$

각 항에 나오는 세 번째 부분 $\frac{\partial a_1^2}{\partial x_1^2}$ 의 경우를 보면, $a_1^2 = \varsigma(x_1^2)$

$$\frac{\partial a_1^2}{\partial x_1^2} = \frac{\partial \varsigma(x_1^2)}{\partial x_1^2} = \varsigma'(x_1^2)$$

마지막 계층이 아닐 때 δ_k^l iii

정리하면 δ_1^2는 다음과 같이 정리할 수 있음

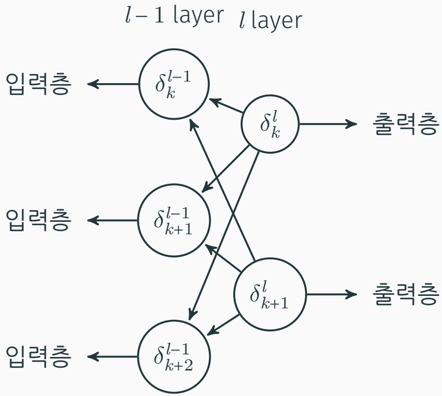
$$
\begin{aligned}
\delta_1^2 &= \frac{\partial E}{\partial x_1^3}\frac{\partial x_1^3}{\partial a_1^2}\frac{\partial a_1^2}{\partial x_1^2} + \frac{\partial E}{\partial x_2^3}\frac{\partial x_2^3}{\partial a_1^2}\frac{\partial a_1^2}{\partial x_1^2} + \frac{\partial E}{\partial x_3^3}\frac{\partial x_3^3}{\partial a_1^2}\frac{\partial a_1^2}{\partial x_1^2} \\
&= \delta_1^3 w_{11}^3 \varsigma'(x_1^2) + \delta_2^3 w_{21}^3 \varsigma'(x_1^2) + \delta_3^3 w_{31}^3 \varsigma'(x_1^2) \\
&= (\delta_1^3 w_{11}^3 + \delta_2^3 w_{21}^3 + \delta_3^3 w_{31}^3)\varsigma'(x_1^2)
\end{aligned}
$$

이 식을 일반화하면 다음과 같이 표현할 수 있음

$$
\begin{aligned}
\delta_k^l &= (\delta_1^{l+1} w_{1k}^{l+1} + \delta_2^{l+1} w_{2k}^{l+1} + \cdots + \delta_m^{l+1} w_{mk}^{l+1})\varsigma'(x_k^l) \\
\delta_k^l &= \sum_{i=1}^{m}(\delta_i^{l+1} w_{ik}^{l+1})\varsigma'(x_k^l)
\end{aligned}
$$

m은 $l+1$ 계층에 있는 노드의 개수

Case 2: 마지막 계층이 아닐 때의 δ_k^l의 도식화



l계층의 δ_k^l에서 $l-1$계층의 오차 δ_k^{l-1}를 구할 수 있음

$$\delta_k^l = \begin{cases} (a_k^l - y_k)\varsigma'(x_k^L) & \text{L} \\ \sum_{i=1}^m (\delta_i^{l+1} w_{ik}^{l+1})\varsigma'(x_k^l) & l \end{cases}$$

바이어스를 구하는 식은 가중치를 구하는 식과 같음

$l-1$계층의 출력이 1일 때 $a_j^{l-1} = 1$ 이 됨

$$\frac{\partial E}{\partial b_k^l} = \delta_k^l$$

오차역전파법 공식 정리 i

$$\frac{\partial E}{\partial w_{kj}^l} = \delta_k^l a_j^{l-1}$$
$$\frac{\partial E}{\partial b_k^l} = \delta_k^l j^l$$

$$\delta_k^l = \begin{cases} (a_k^l - y_k)\varsigma'(x_k^L) & l\text{이 마지막 계층일 때} \\ \sum_{i=1}^m (\delta_i^{l+1} w_{ik}^{l+1})\varsigma'(x_k^l) & l\text{이 마지막 계층이 아닐 때} \end{cases}$$

E: 손실함수
w_{kj}^l: l계층 k번째 노드의 $l-1$계층 j번째 노드로부터의 가중치
δ_k^l: l계층 k번째 노드의 오차
$a-j^{l-1}$: $l-1$계층 j번째 노드의 출력

b_k^l: l계층 k번째 노드의 바이어스
y_k: k번째 노드의 정답 레이블
$\varsigma(x_k^l)$: 활성화 함수
m: $l+1$계층의 노드 개수(시그모이드 함수 등)

오차역전파법 수식과 순서 정리 i

1. 손실 함수를 구한 후, 그 값을 최소화하기 위한 w와 b를 구함

수식

$$E = \frac{1}{2}\|\vec{t} - \vec{y}\|^2, \qquad \vec{y} = W\vec{x} + \vec{b}$$

t: 정답 레이블
y: 신경망의 출력
W: 가중치
x: 출력
b: 바이어스

문제점: 최소화하고 싶은 변수 w와 b의 개수가 너무 많아 미분할 때 0이 되는 연립방정식을 풀기가 어려움

2. 경사하강법을 사용하여 손실 함수의 값이 작아지는 방향을 확인

수식

$$w_{new} = w_{old} - \eta \frac{\partial E}{\partial w_{old}}, \qquad b_{new} = b_{old} - \eta \frac{\partial E}{\partial b_{old}}$$

w_{new}: 이동 후의 가중치
w_{old}: 이동 전의 가중치
η: 학습률

b_{new}: 이동 후의 바이어스
b_{old}: 이동전의 바이어스

문제점 : 값을 움직일 양을 구하기 위해 $\frac{\partial E}{\partial w}$ 와 $\frac{\partial E}{\partial b}$ 를 계산하기에는 너무 많은 미분 대상이있어서 어려움

3. 오차역전파법을 사용하여 가중치를 결정

수식

$$\frac{\partial E}{\partial w^l_{kj}} = \delta^l_k a^{l-1}_j, \qquad \frac{\partial E}{\partial b^l_k} = \delta^l_k j^l$$

$$\delta^l_k = \begin{cases} (a^l_k - y_k)\varsigma'(x^L_k) & l\text{이 마지막 계층일 때} \\ \sum_{i=1}^m (\delta^{l+1}_i w^{l+1}_{ik})\varsigma'(x^l_k) & l\text{이 마지막 계층이 아닐 때} \end{cases}$$

E: 손실함수
w^l_{kj}: l계층 k번째 노드의 $l-1$계층 j번째 노드로부터의 가중치
δ^l_k: l계층 k번째 노드의 오차
$a-j^{l-1}$: $l-1$계층 j번째 노드의 출력

b^l_k: l계층 k번째 노드의 바이어스
y_k: k번째 노드의 정답 레이블
$\varsigma(x^l_k)$: 활성화 함수
m: $l+1$계층의 노드 개수(시그모이드 함수 등)

† 학습시에는 배치 사이즈의 개수만큼 순전파를 진행, 경사하강법과 오차역전파법을 사용해서 가중치와 바이어스를 갱신. 2와 3을 반복하여 w와 b의 근삿값을 찾음

모델 평가

홀드아웃 holdout 평가는 정답률을 보고 결정함. 분석된 결과를 카테고리별로 분류한 후 그 중에서 몇%가 정답인지 확인함.

드랍아웃 dropout 무작위로 뉴런(노드)을 제거하여 정보의 전달을 막아 학습 데이터의 노이즈나 특징에 영향을 덜 받게 만드는 방법


That's all Folks!

참고문헌

이시카와 아키히코 저/신상재, 이진희 역, "인공지능을 위한 수학 꼭 필요한 것만 골라 배우는 인공지능 맞춤 수학," 프리렉, 2018년 11월 22일, ISBN: 9788965402282