기초미분과 최적화

윤 정 훈

들어가며

- 대표적인 분류기 SVM에서는 여백이라는 개념을 정의하고 여백을 **최대**화하는 결정 직선을 찾는 것이 목적이다.

- 신경망에서는 모델을 추정하기 위하여, 목적함수 또는 손실함수를 **최소**로 하는 모수를 찾는 것이 목적이다.

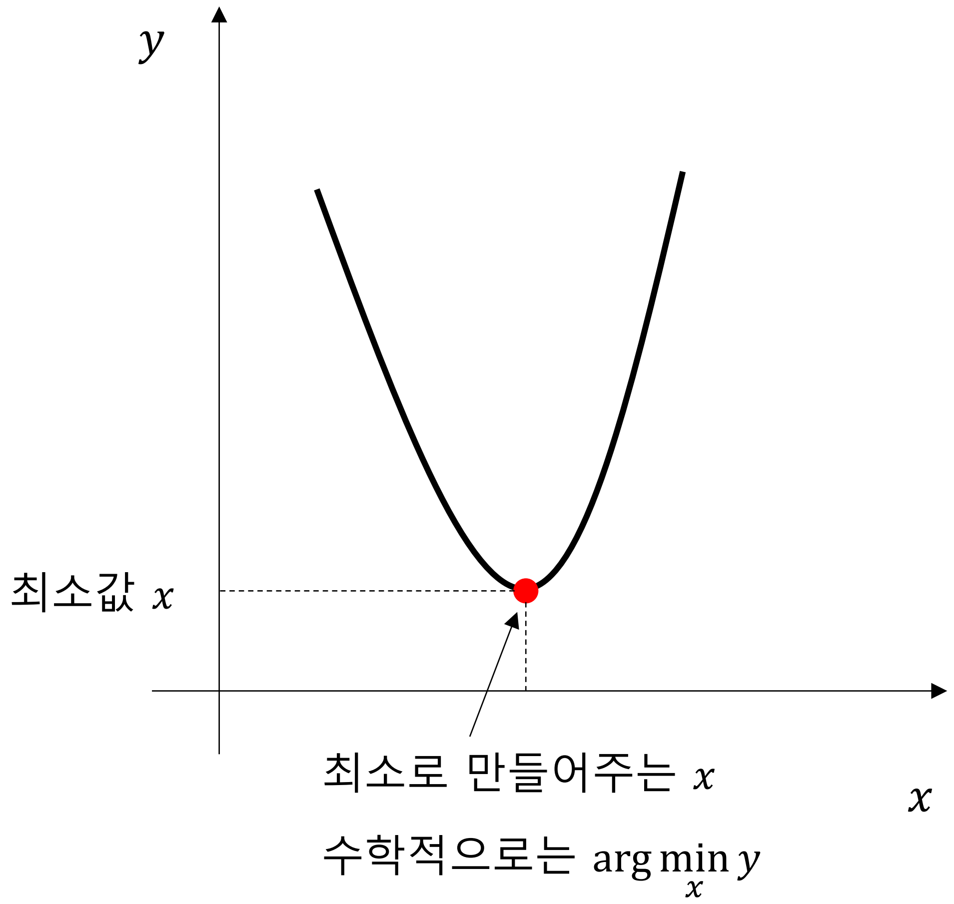
- 베이시언 분류기에서는 분류 오류를 **최소**화 하는 분류기를 찾는 것이 목적이다.

최적화 문제

$y = (x-2)^2 + 2$ 이 주어졌다.

이 함수가 최소값을 가지도록 하는

x의 값을 찾고 싶다.



최적화 문제

최소값을 만들어주는 x 는 왜 찾을까?

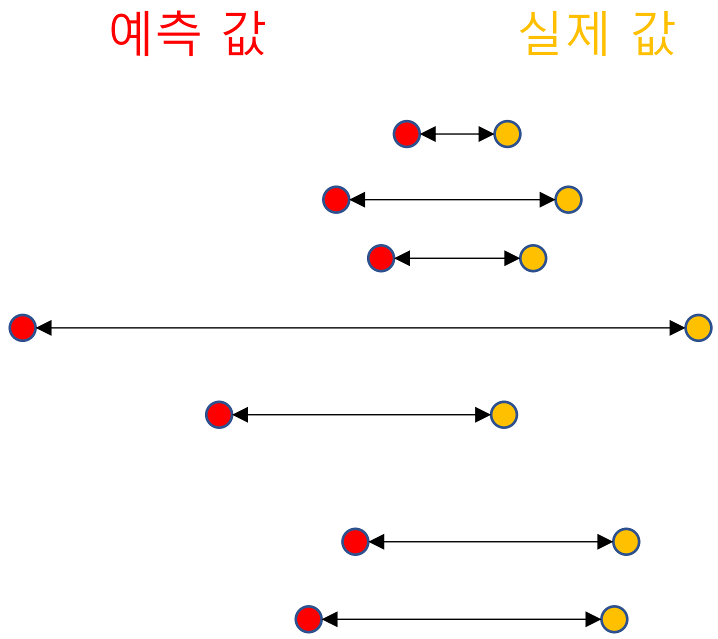
두 가지 경우를 비교해보자.

오른쪽의 경우가,
예측 모델의 성능이 더 좋다.

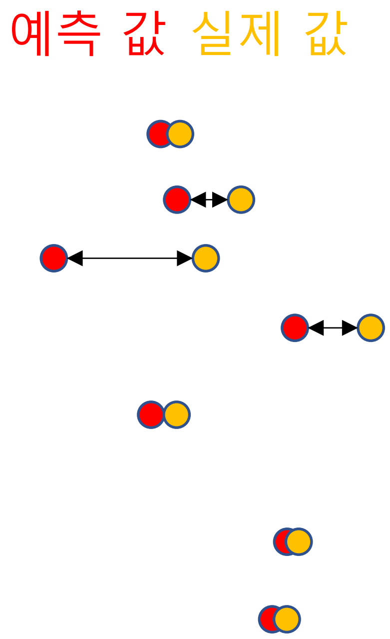
성능이 좋다는 근거는 무엇인가?

예측 값과 실제 값의 거리가 가깝고,
차이가 적기 때문이다.

한마디로 차이가 작으면 작을수록
더 좋은 모델이라고 할 수 있다.



화살표는 두 점 사이의 거리를 의미한다. 두 집단 간 거리가 멀다. (entry 개별 비교시)



두 집단 간 거리가 가깝다.

최적화 문제

- 좋은 예측 모델을 만드는 방법 (직관적으로 생각해 볼 때)
  - 실제 값과, 실제 값을 예측한 값의 차이를 측정할 수 있는(measure) 수치를 만든다.
    - 두 값의 거리 (dissimilarity)
    - 예측 값이 실제 값과 일치 하지 않는 경우 (error rate)
  - 이 수치를 가장 작게 만드는 모델이 좋은 모델이다.

이 수치를 목표로 삼고, 수치의 값을 가장 작게 만들도록 노력한다. → 머신러닝의 목표

최적화 문제

✓ $J(\theta)$: 목적 함수(target function) 또는 비용 함수(cost function)이며, 우리가 최대화 하거나 최소화 하려는 함수이다.

✓ θ : 매개 변수

- 최대화 문제

  $J(\theta)$를 최대로 하는 $\hat{\theta}$를 찾아라. 즉 $\hat{\theta} = \arg\max_{\theta} J(\theta)$

- 최소화 문제

  $J(\theta)$를 최소로 하는 $\hat{\theta}$를 찾아라. 즉 $\hat{\theta} = \arg\min_{\theta} J(\theta)$

분석적 방법 vs 수치해석적 방법

- 분석적 방법
  - 입력 : 목적 함수 $J(\theta)$
  - 출력 : $\hat{\theta}$ (최고 점 또는 최저 점)
  - 알고리즘

① $J(\theta)$를 θ로 미분한다.

② 방정식 $\frac{\partial J(\theta)}{\partial \theta} = 0$을 만족하는 $\hat{\theta}$를 구한다.

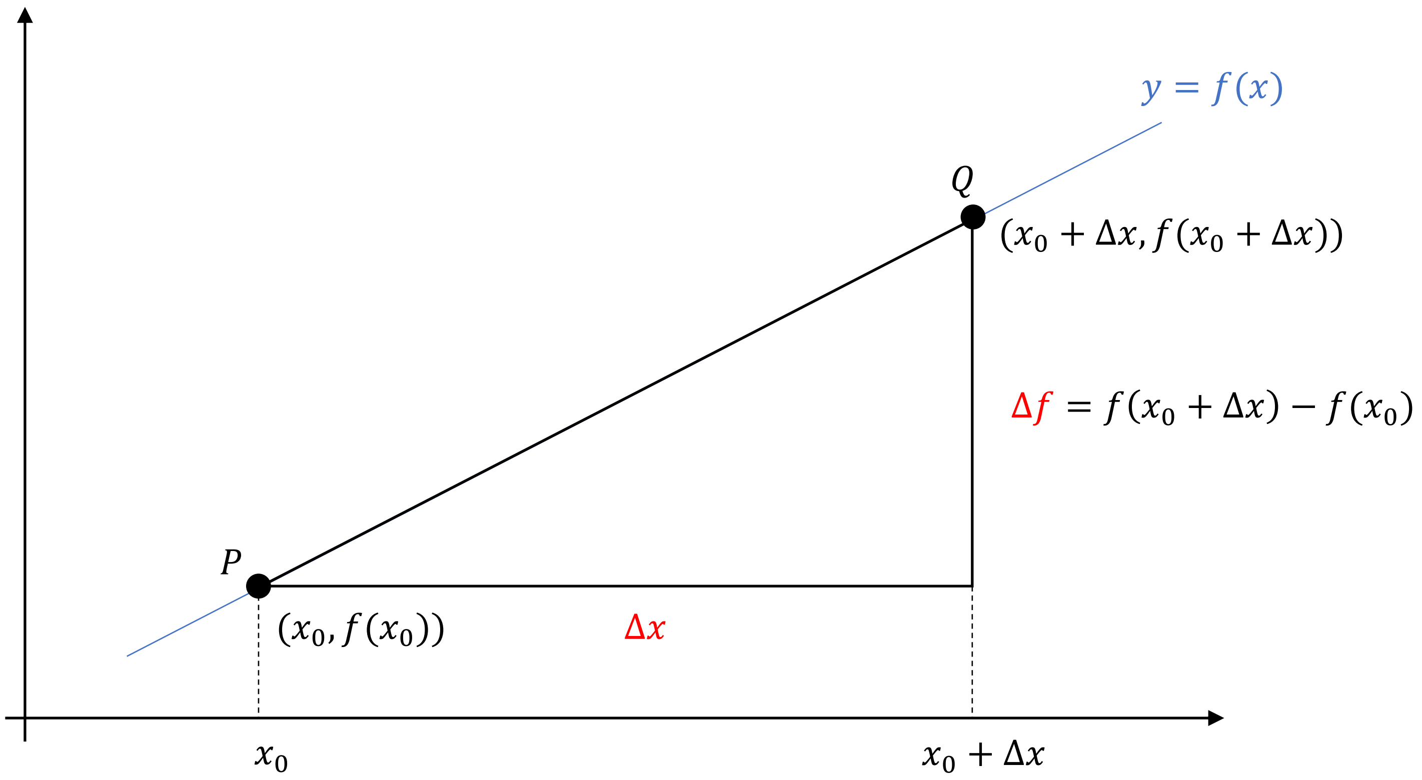
③ $\hat{\theta}$를 리턴한다.

분석적 방법 vs 수치해석적 방법

- 수치해석적 방법
  - 선형회귀모형, 로지스틱회귀모형에서 모수의 값을 추정하기 위하여, 정규방정식을 풀어야 한다. 정규방정식을 푸는 방법 중의 하나는 방정식의 초기해를 설정한 후, 초기해를 지속적으로 업데이트 하여, 우리가 구하고자 하는 해 즉 최대값 또는 최소값에 가장 근접하도록 하는 방법이다. 이러한 방법을 수치해석적으로 해를 구하는 방법이라고 한다.
  - 대표적인 이론이 경사하강법(Gradient Descent)이다. 이 이론을 정확하게 이해하기 위해서는 기본적인 미분, 편미분에 대한 개념을 정확하게 이해해야 한다.

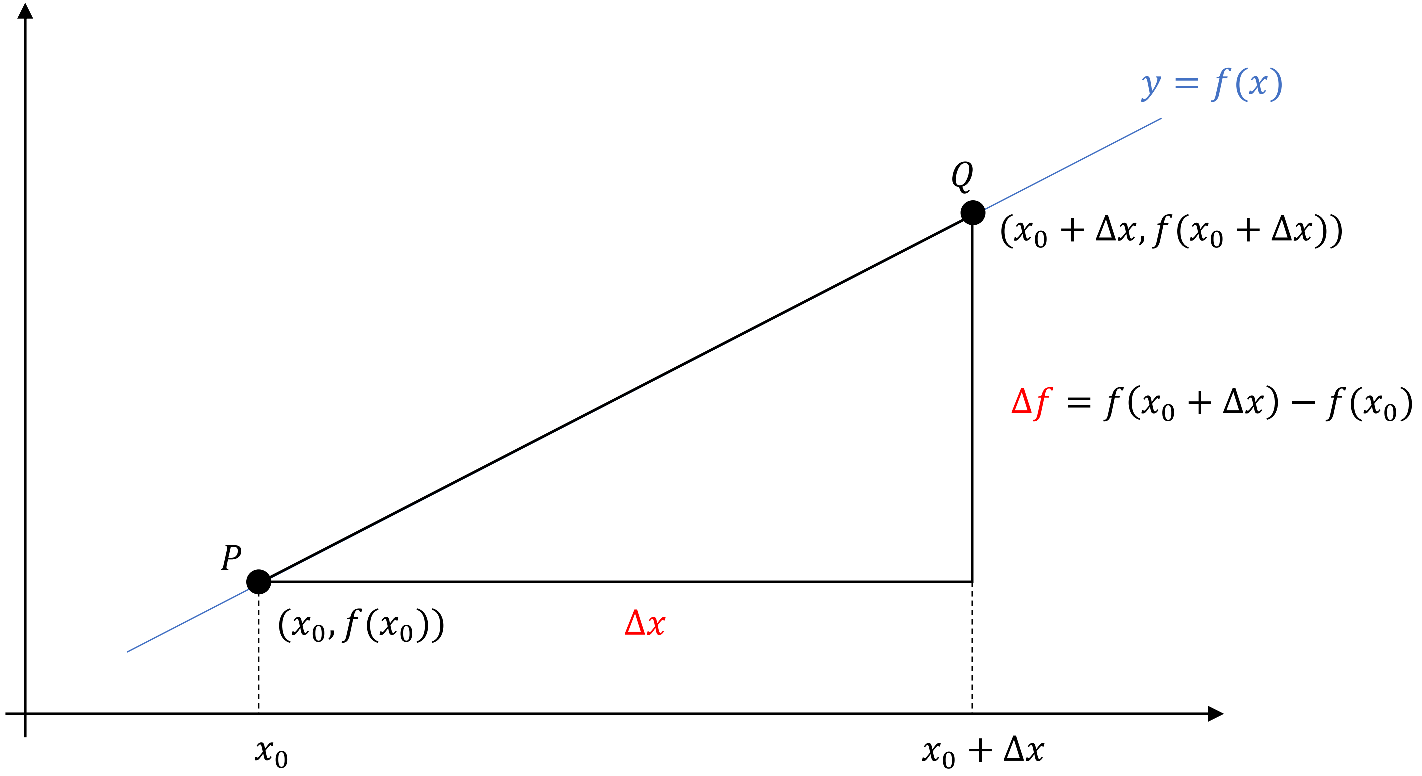
증분

- x의 증분(Δx) : x의 증가량
- y의 증분(Δy) : y의 증가량

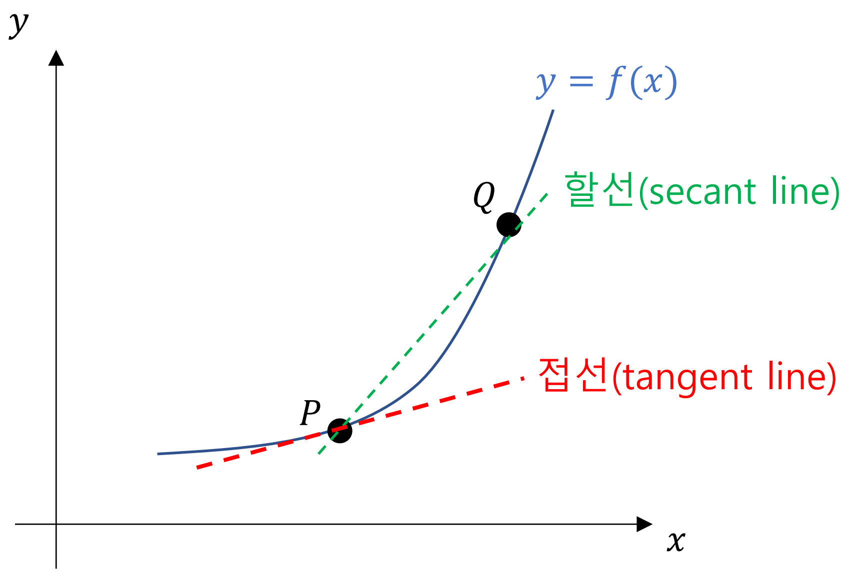


기울기(slope)

- 기울기 : $\frac{y\text{의 증분}}{x\text{의 증분}} = \frac{\Delta y}{\Delta x} = \frac{f(x_0+\Delta x)-f(x_0)}{\Delta x}$



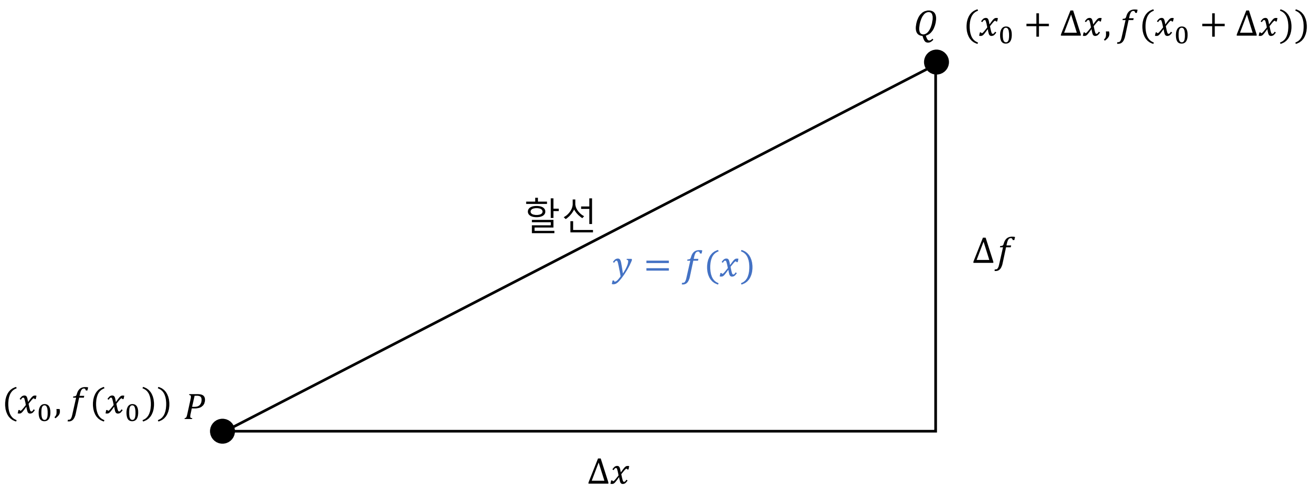
도함수(Derivatives)



도함수는 그래프 $f(x)$의 접선의 기울기이다. 그러면 접선은 무엇인가?

- 그래프의 한 지점(point)에서 만나는 선이 절대 아니다.
- 접선은 두 지점 사이의 할선이며, 두 지점 사이의 거리가 0으로 갈 때의 할선의 극값이다.

도함수(Derivatives)



- 도함수 : $P \longrightarrow Q$ 일 때, 할선 PQ 기울기의 극한 값

- $\lim_{\Delta x \to 0} \frac{\Delta f}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x_0+\Delta x)-f(x_0)}{\Delta x} = f'(x_0) = \frac{df(x_0)}{dx} = \frac{dy}{dx}$

여러가지 함수의 도함수 Advanced

- $y = \{f(x)\}^n \Longrightarrow \frac{dy}{dx} = n \cdot f'(x) \cdot \{f(x)\}^{n-1}$
- $y = \log_a f(x) \Longrightarrow \frac{dy}{dx} = \ln a \frac{f'(x)}{f(x)}$
- $y = e^{f(x)} \Longrightarrow \frac{dy}{dx} = f'(x) \cdot e^{f(x)}$
- $y = f(x) \cdot g(x) \Longrightarrow \frac{dy}{dx} = f'(x) \cdot g(x) + f(x) \cdot g'(x)$
- $y = \frac{f(x)}{g(x)} = \frac{f'(x) \cdot g(x) - f(x) g'(x)}{\{g(x)\}^2}$

도함수의 예제 Advanced

- $y = \log_e x \Longrightarrow y' = ?$
- $y = \log_e(x^3) \Longrightarrow y' = ?$
- $y = e^{4x} \Longrightarrow y' = ?$
- $y = xe^{3x} \Longrightarrow y' = ?$
- $y = \frac{e^x}{1+e^x} \Longrightarrow y' = ?$

Chain Rule

- 합성함수 $y = f(g(t))$의 도함수 $\frac{dy}{dt}$
  - $\frac{dy}{dt} = \frac{dy}{dx}\frac{dx}{dt}$
- 합성함수 $y = (f \circ g)(x)$의 도함수 $\frac{d}{dx}f\big(g(x)\big)$
  - $\frac{d}{dx}f\big(g(x)\big) = f'\big(g(x)\big)g'(x)$
- 예를 들어, $y = \sin x$ 이고 $x = t^2$ 일 때, $\frac{dy}{dt}$를 구해보자.
  - $\frac{dx}{dt} = 2t$, $\frac{dy}{dx} = \cos x$
  - $\frac{d}{dt}(\sin(t^2)) = \left(\frac{dy}{dx}\right)\left(\frac{dx}{dt}\right) = (\cos x)(2t) = 2t \cdot \cos(t^2)$

고차원의 도함수 Advanced

- 고차원의 도함수는, 도함수의 도함수들을 의미한다.
  - $f'(x) = \frac{df}{dx} = Df$
  - $f''(x) = \frac{d^2f}{dx^2} = D^2f$
  - $f'''(x) = \frac{d^3f}{dx^3} = D^3f$
  - $f^{(n)}(x) = \frac{d^nf}{dx^n} = D^nf$

다변수함수의 미분

- 이전까지는 독립변수가 하나인 함수를 대상으로 도함수를 구하였다. 하지만 다양한 분야에서 사용되는 많은 이론은 여러 변수들에 의해 함수값이 결정되는 다변수함수의 형태를 띠고 있다.

- 예를 들면 경제학에서 사용하는 수요함수는 자신의 가격(P_1)뿐만 아니라 타 재화의 가격 $(P_2, \dots, P_n)$, 소득(M) , 기호(T)에 의해서 결정된다. 따라서 수요함수는 $D = f(P_1, P_2, \dots, P_n, M, T)$로 나타낼 수 있다. 이 때 이들 독립변수가 변하면 수요의 변화는 어떻게 될까?

다변수함수의 미분의 종류

- 편도함수

- 전미분

1차편도함수

2변수함수 $z = f(x, y)$에서 y변수가 특정한 값에 고정되어 변하지 않는다고 가정하면(y변수를 상수로 가정하면) 2변수함수는 실질적으로 독립변수가 하나인 1변수함수가 된다. 이 함수를 x로 미분하면 도함수가 구해지는데 이것을 x의 편도함수(partial derivative)라 한다. 마찬가지로 x변수가 특정한 값에 고정되어 변하지 않는다고 가정하면 y의 편도함수를 구할 수 있다.

1차편도함수

- 편도함수는 다음과 같이 정의된다.

함수 $f(x, y)$가 모든 점에서 미분 가능할 때 x와 y에 대한 각각의 편도함수는 다음과 같이 정의된다.

$$\frac{\partial f}{\partial x} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x, y) - f(x, y)}{\Delta x}$$

$$\frac{\partial f}{\partial y} = \lim_{\Delta x \to 0} \frac{f(x, y + \Delta y) - f(x, y)}{\Delta y}$$

1차편도함수

- 편도함수의 정의는 한 변수의 평균변화율을 나타내는 차원에서 실질적으로 동일한 의미를 갖는다. 편도함수를 나타내기 위해 f_x , $\frac{\partial f}{\partial x}$, $D_x f$ 등의 기호로 표시한다.

- 어떤 특정한 점 (a, b)에서 x의 편미분계수는 $\frac{\partial f}{\partial x}|_{(a,b)}$ 또는 $f_x(a, b)$로, y의 편미분계수는 $\frac{\partial f}{\partial y}|_{(a,b)}$ 또는 $f_y(a, b)$로 표시한다. 점 (a, b)에서 편미분계수 $f_x(a, b)$는 $y = b$로 고정되어 있는 상태에서 $x = a$에서 x단위 변화에 대한 z의 변화율을 나타낸다. 다변수함수의 경우에도 관심대상의 변수 외의 모든 변수는 상수 취급하므로 실질적으로 1변수함수가 된다. 이와 같이 편도함수를 구하는 것을 "편미분한다" 라고 한다.

1차편도함수 예제1

- $f(x,y) = xy^2 + x^2y$의 x와 y의 편도함수를 구해보자.

  x의 편도함수는 y를 상수로 하고 x에 대해서만 미분하므로 $f_x(x,y) = y^2 + 2xy$가 구해진다.

  y의 편도함수는 x를 상수로 하고 y에 대해서만 미분하므로 $f_y(x,y) = 2xy + x^2$이 구해진다.

  점 $(1,2)$에서,

  x의 편미분계수를 구하면, $f_x(1,2) = 2^2 + 2\times1\times2 = 8$ 이고

  y의 편미분계수를 구하면, $f_y(1,2) = 2\times1\times2 + 1^2 = 5$ 이다.

1차편도함수 예제2

- $f(x,y) = \ln(x^2 + 2xy - y^2)$에서 f_x , f_y를 구하고 점 $(1,1)$에서 각 변수의 편미분계수를 구하라.

  $x^2 + 2xy - y^2 = u$라고 놓으면, $f(x,y) = \ln u$가 된다. 연쇄법칙에 의하여

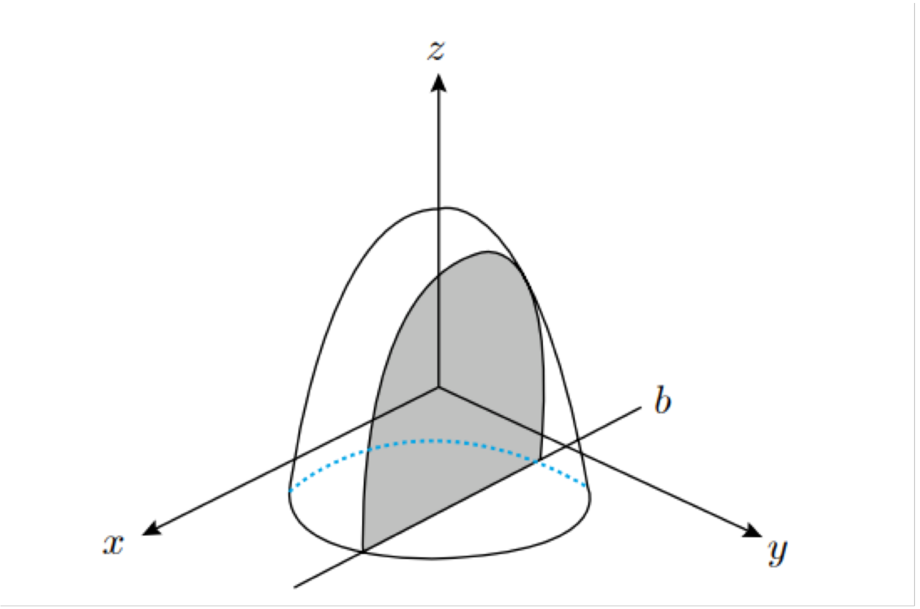
  x의 편도함수는 $f_x(x,y) = \frac{\partial z}{\partial x} = \frac{\partial z}{\partial u} \cdot \frac{\partial u}{\partial x} = \frac{1}{u} \cdot \frac{\partial}{\partial x}(x^2 + 2xy - y^2) = \frac{2x+2y}{x^2+2xy-y^2}$,

  y의 편도함수는 $f_y(x,y) = \frac{\partial z}{\partial y} = \frac{\partial z}{\partial u} \cdot \frac{\partial u}{\partial y} = \frac{1}{u} \cdot \frac{\partial}{\partial y}(x^2 + 2xy - y^2) = \frac{2x-2y}{x^2+2xy-y^2}$ 가 성립한다.

  x의 편미분계수 $f_x(1,1) = \frac{4}{2} = 2$이고,

  x의 편미분계수 $f_x(1,1) = \frac{0}{2} = 0$이다.

1차편도함수의 기하학적 의미



$z = f(x, y)$의 그래프

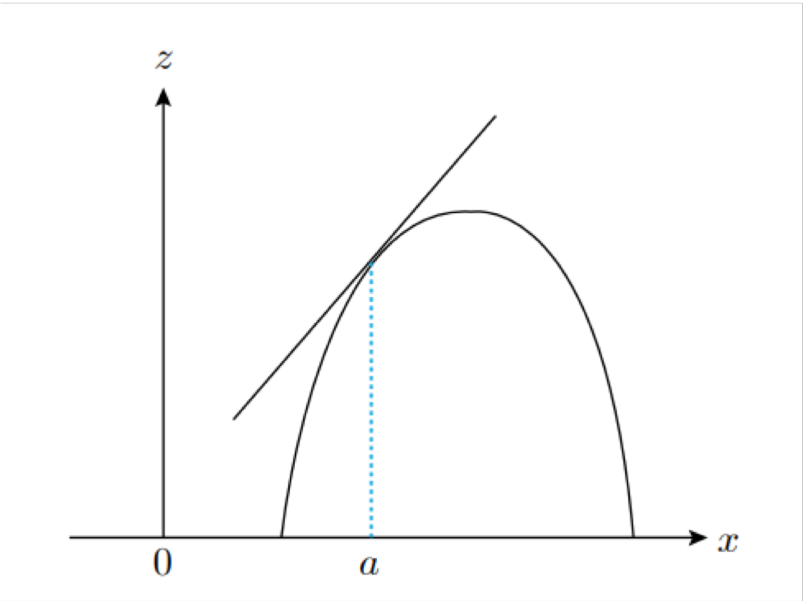
$y = b$는 기하학적으로

xz평면에 평행하고 $y = b$선을 지나는 평면이다.

이 평면으로 $z = f(x, y)$의 곡면을 절단하면

단면에 나타나는 함수는 $z = f(x, b)$이다.

1차편도함수의 기하학적 의미



x의 편미분계수 : $f_x(x, y)$

$f_x(a, b)$는 함수

$z = f(x, b)$의 $x = a$에서 그은 접선의 기울기와 일치한다.

즉, y방향으로 전혀 움직이지 않고

$x = a$에서 x축 방향으로 단위 변화할 때 z의 변화율을 나타낸다.

y의 편도함수 $f_y(a, b)$의 경우는,

$x = a$ 평면으로 $z = f(x, y)$ 공간을 자르고 나타난

함수 $z = f(a, y)$의 $y = b$에서 그은 접선의 기울기가 된다.

다변수함수의 편미분

- 다변수함수 $z = f(x_1, x_2 \dots, x_n)$에서 x_i를 제외한 모든 변수가 고정되었다고 하면(x_i를 제외한 모든 변수가 상수라 하면) x_i의 편도함수는 다음과 같이 나타낸다.

$$\frac{\partial f}{\partial x_i} = f_i(x_1, x_2, \dots, x_n) \text{ , } (i = 1, \dots, n)$$

ex) $f(x, y, z) = xy + yz + xz$의 1차 편도함수를 구하라.

x의 1차편도함수는 y , z를 상수로 하고 x에 대해서만 미분하면

$f_x = y + z$

$f_y = x + z$

$f_z = x + y$

가 된다.

벡터의 편미분

- 다변수함수 $z = f(x_1, x_2 \dots, x_n)$라고 하고, 벡터 $\mathbb{x} = (x_1, x_2, \dots, x_n)$이라고 하자. 즉, $z = f(\mathbb{x})$이다. 이 때, 벡터 $\mathbb{x}$의 편미분 $\frac{\partial z}{\partial \mathbb{x}}$는 다음과 같이 나타낸다.

$$\frac{\partial z}{\partial \mathbb{x}} = \left(\frac{\partial z}{\partial x_1}, \frac{\partial z}{\partial x_2}, \dots, \frac{\partial z}{\partial x_n}\right)$$

ex) $\mathbb{x} = (x_1, x_2, x_3)$에 대하여, $f(\mathbb{x}) = f(x_1, x_2, x_3) = x_1x_2 + x_2x_3 + x_3x_1$의 편미분 $\frac{\partial z}{\partial \mathbb{x}}$을 구하라.

$$\frac{\partial f}{\partial x_1} = x_2 + x_3$$
$$\frac{\partial f}{\partial x_2} = x_1 + x_3$$
$$\frac{\partial f}{\partial x_3} = x_2 + x_1$$

이므로 $\frac{\partial z}{\partial \mathbb{x}} = (x_2 + x_3, x_1 + x_3, x_2 + x_1)$ 이다.

2차편도함수 Advanced

- 함수 $z = f(x,y)$의 1차편도함수 $f_x(x,y)$, $f_y(x,y)$는 x, y의 형태를 띤다. 따라서 1차편도함수가 미분 가능하면 편도함수 정의에 의해서 1차편도함수를 가지고 2차편도함수를 구할 수 있다.

  - $f_{xx} = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial x^2}$

  - $f_{xy} = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial x \partial y}$

  - $f_{yx} = \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) = \frac{\partial^2 f}{\partial y \partial x}$

  - $f_{yy} = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right) = \frac{\partial^2 f}{\partial y^2}$

2차편도함수 예제 Advanced

- 함수 $f(x, y) = x^3 + 2xy^3$의 2차편도함수를 구하라.

  x, y에 대한 편미분하면 1차편도함수는

  - $f_x = 3x^2 + 2y^3$
  - $f_y = 6xy^2$

  2차편도함수는

  - $f_{xx} = 6x$
  - $f_{xy} = 6y^2 = f_{yx}$
  - $f_{yy} = 12xy$

1차/2차 편도함수 예제 Advanced

- 1차 편도함수를 구하여라.
  - $f(x, y) = (x - 6y)(2x + 3y^2)$
  - $g(t_1, t_2) = \frac{2t_2 + 4t_1}{t_2^2 - 3t_1}$
  - $f(x, y, z) = (3x^2 + z^2)(z - y)$
- 2차 편도함수를 구하여라.
  - $f(x_1, x_2, x_3) = 2x_1^3 x_2 + 5x_3^4 x_2$
  - $f(x, y, z) = x - \sqrt{y^2 + z^2}$
  - $f(x, y) = x \ln y$

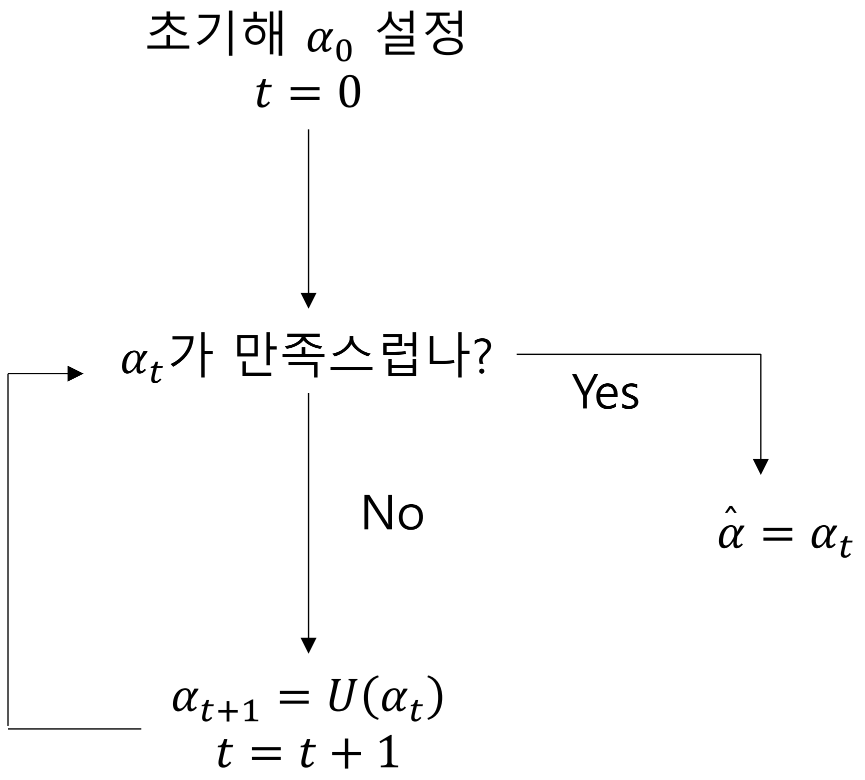
경사하강법 들어가기

- machine learning에서는 매개 변수(parameter, 선형회귀에서는 θ_0, θ_1)가 수십~수백 차원의 벡터인 경우가 대부분이다. 또한 목적 함수(선형회귀에서는 $\Sigma\epsilon_i^2$)가모든 구간에서 미분 가능하다는 보장이 항상 있는 것도 아니다.

- 따라서 한 번의 수식 전개로 해를 구할 수 없는 상황이 적지 않게 있다.

- 이런 경우에는 초기 해에서 시작하여 해를 반복적으로 개선해 나가는 수치적 방법을 사용한다. (미분이 사용 됨)

경사하강법의 개념


초기해 α_0 설정
$t = 0$
α_t가 만족스럽나?
Yes
No
$\hat{\alpha} = \alpha_t$
$\alpha_{t+1} = U(\alpha_t)$
$t = t + 1$

경사하강법의 정의

- **Gradient Descent**

  현재 위치에서 경사가 가장 급하게 하강하는 방향을 찾고, 그 방향으로 약간 이동하여 새로운 위치를 잡는다. 이러한 과정을 반복함으로써 가장 낮은 지점(즉 최저 점)을 찾아 간다.

- **Gradient Ascent**

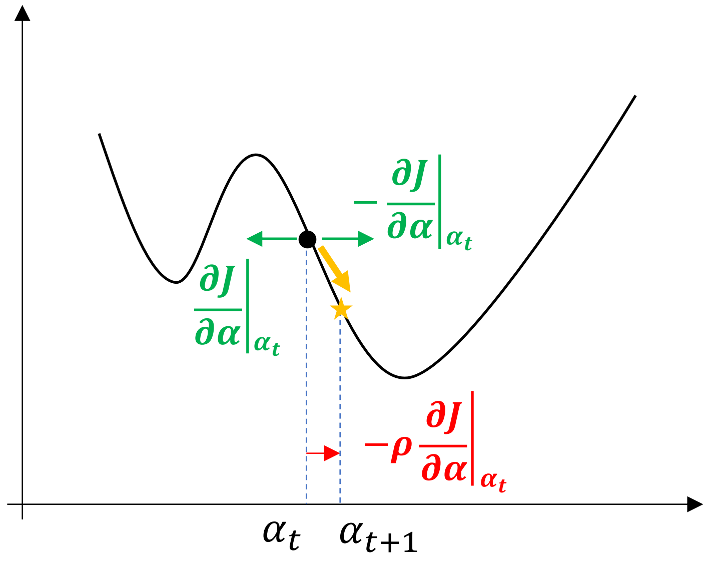
  - 현재 위치에서 경사가 가장 급하게 상승하는 방향을 찾고, 그 방향으로 약간 이동하여 새로운 위치를 잡는다. 이러한 과정을 반복함으로써 가장 높은 지점(즉 최대 점)을 찾아 간다.

경사하강법 알고리즘

J = 목적함수

$\left.\frac{\partial J}{\partial \alpha}\right|_{\alpha_t}$: α_t에서의 도함수 $\frac{\partial J}{\partial \alpha}$의 값

$$\alpha_{t+1} = \alpha_t - \rho \left.\frac{\partial J}{\partial \alpha}\right|_{\alpha_t}$$



α_t에서의 미분값은 음수이다.

그래서 $\left.\frac{\partial J}{\partial \alpha}\right|_{\alpha_t}$ 를 더하게 되면
왼쪽으로 이동하게 된다.
그러면 목적함수의 값이 증가하는
방향으로 이동하게 된다.

따라서 $\left.\frac{\partial J}{\partial \alpha}\right|_{\alpha_t}$ 를 빼준다.
그리고 적당한 ρ(스텝크기, 학습률)를 곱해주어서
조금만 이동하게 한다.

경사하강법 알고리즘

$J =$ 목적함수

$\left.\frac{\partial J}{\partial \alpha}\right|_{\alpha_t}$: α_t에서의 도함수 $\frac{\partial J}{\partial \alpha}$의 값

Gradient Descent

$$\alpha_{t+1} = \alpha_t - \rho \left.\frac{\partial J}{\partial \alpha}\right|_{\alpha_t}$$

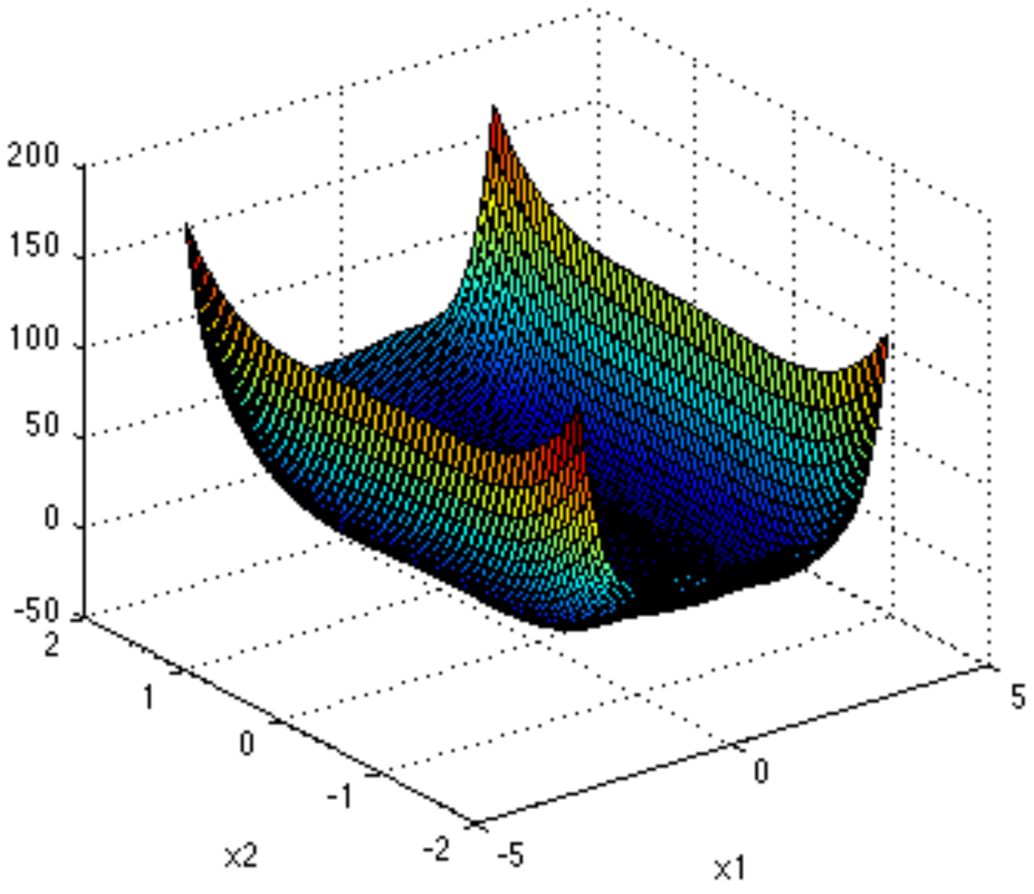
Gradient Ascent

$$\alpha_{t+1} = \alpha_t + \rho \left.\frac{\partial J}{\partial \alpha}\right|_{\alpha_t}$$

Gradient Descent, Gradient Ascent는 전형적인 Greedy algorithm이다.
<u>과거 또는 미래를 고려하지 않고 현재 상황에서 가장 유리한 다음 위치를 찾아</u>
<u>Local optimal point로 끝날 가능성</u>을 가진 알고리즘이다.

경사하강법 예제1

- 낙타 등 함수(six-hump camelback function)



경사하강법 예제1

- 낙타 등 함수(six-hump camelback function)
  - ✓ $J(\Theta) = \left(4 - 2.1\theta_1^2 + \frac{\theta_1^4}{3}\right)\theta_1^2 + \theta_1\theta_2 + (-4 + 4\theta_2^2)\theta_2^2$
  - ✓ 초기값을 $\Theta_0 = (-0.5,0.5)^T$로 하고 학습률을 $\rho = 0.01$로 하자.
  - ✓ $J'(\Theta) = \frac{\partial J}{\partial \Theta} = \left(\frac{\partial J}{\partial \theta_1}, \frac{\partial J}{\partial \theta_2}\right)^T = \left(2\theta_1^5 - 8.4\theta_1^3 + 8\theta_1 + \theta_2, 16\theta_2^3 - 8\theta_2 + \theta_1\right)^T$

- Θ_1을 구해보자
  ① $\frac{\partial J}{\partial \Theta}|_{\Theta_0} = (-2.5125, -2.5)^T$
  ② $\Theta_1 = \Theta_0 - 0.01\times\frac{\partial J}{\partial \Theta}|_{\Theta_0} = (-0.5,0.5)^T - 0.01\times(-2.5125, -2.5)^T = (-0.4748,0.525)^T$

경사하강법 예제1

- 낙타 등 함수(six-hump camelback function)
  - ✓ $J(\Theta) = \left(4 - 2.1\theta_1^2 + \frac{\theta_1^4}{3}\right)\theta_1^2 + \theta_1\theta_2 + (-4 + 4\theta_2^2)\theta_2^2$
  - ✓ 초기값을 $\Theta_0 = (-0.5, 0.5)^T$로 하고 학습률을 $\rho = 0.01$로 하자.
  - ✓ $J'(\Theta) = \frac{\partial J}{\partial \Theta} = \left(\frac{\partial J}{\partial \theta_1}, \frac{\partial J}{\partial \theta_2}\right)^T = \left(2\theta_1^5 - 8.4\theta_1^3 + 8\theta_1 + \theta_2, 16\theta_2^3 - 8\theta_2 + \theta_1\right)^T$

- Θ_2을 구해보자!

① $\frac{\partial J}{\partial \Theta}|_{\Theta_1} = (-2.4228, -2.3596)^T$

② $\Theta_2 = \Theta_1 - 0.01 \times \frac{\partial J}{\partial \Theta}|_{\Theta_1} = (-0.4748, 0.525)^T - 0.01 \times (-2.4228, -2.3596)^T = (-0.4506, 0.5486)^T$

경사하강법 예제1

$\Theta_0 = (-0.5, 0.5)^T$

$\Theta_1 = (-0.4748, 0.525)^T$

$\Theta_2 = (-0.4506, 0.5486)^T$

위의 값을 대입하여 $J(\Theta)$를 계산하면 아래와 같다.

$J(\Theta_0) = -0.12604$

$J(\Theta_1) = -0.24906$

$J(\Theta_2) = -0.36036$

학습률의 영향

- Gradient에서 알 수 있는 것은 함수값이 가장 빠르게 증가하는 방향이다. 그 방향으로 얼만큼을 가야하는지는 알려주지 않는다. 얼만큼 가야하는지를 의미하는 학습률(learning rate ρ)는 Gradient를 사용하는 모델을 학습시킬 때 있어 가장 중요한 hyperparameter이다.

- 학습률을 너무 크게 하면 :

  최저 점을 중심으로 좌우를 왔다갔다하는 진자 현상이 발생한다.

- 학습률을 너무 작게 하면 :

  수렴 속도가 느려진다.

학습 데이터 수에 따른 경사하강법 종류

- Mini-batch gradient descent - MGD
  - 우리가 구하고자 하는 모델의 파라미터를 한 번 업데이트하려고 학습데이터 전체를 계산에 사용하는 것은 낭비가 될 수 있다. 학습데이터의 전체가 아닌 배치(batches)만 이용해서 gradient를 계산하는 것이다.
  - 예를들어 120만개 중에 256개짜리 배치만을 이용하여 gradient를 구하고 파라미터를 업데이트한다.
  - 학습데이터가 서로 상관관계가 있기 때문에 전체 데이터를 보지 않고 배치만 이용하여도 이 방법이 효과적임

학습 데이터 수에 따른 경사하강법 종류

- Stochastic gradient descent - SGD
  - 온라인 그라디언트 하강이라고도 한다.
  - Mini-batch gradient descent의 배치 크기가 데이터 한 개 일때 이다. 즉 모델의 파라미터를 계산할 때, 데이터 하나에 대하여 모수를 업데이트한다.
  - 모델의 파라미터를 계산할 때, 행렬 및 벡터의 연산이기 때문에, 한 예제에서 100번 계산하는 것보다, 100개의 예제에서 1번 계산하는게 더 빠르다.
  - 엄밀하게는 데이터 한 개에 대하여, 계산한 후 파라미터를 업데이트 하는 것이 SGD이나, 많은 사람들이 MGD를 의미하면서 SGD라고 부르기도 한다.