



부산대학교

PUSAN NATIONAL UNIVERSITY

Gradient Descent

Walking Downhill

Finding Parameters Without a Formula

BX3600087-140 / DS3600087-140 / AB3600087-140

기계학습개론

2026학년도 2학기

오용경

데이터사이언스전공

yongkyung.oh@pusan.ac.kr

Objective Surface

자료 전체의 평균 손실을 매개변수의 함수로 본다

관측별 손실과 목적함수

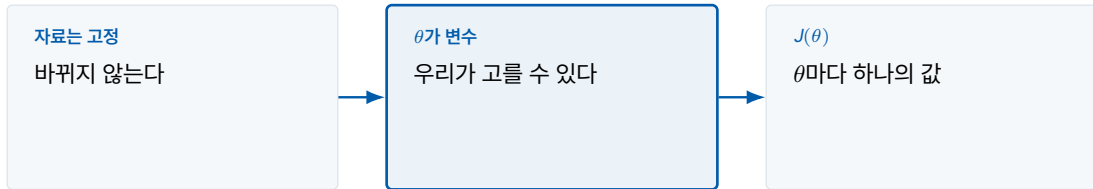
관측별 손실(loss)을 평균하여 자료 전체의 예측 오차를 나타낸다

$$J(\theta) = \frac{1}{n} \sum_{i=1}^n \underbrace{\mathcal{L}(\hat{y}_i(\theta), y_i)}_{\text{한 관측의 손실}}$$

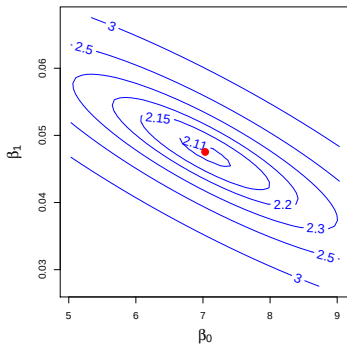
$\mathcal{L}$ 은 관측별 손실을 나타내는 일반 기호다. 선형 회귀에서는 제곱 오차 $(y_i - \hat{y}_i)^2$, 로지스틱 회귀에서는 $\ell(y, p) = -y \log p - (1 - y) \log(1 - p)$ 가 이 자리에 들어간다

학습에서 최소화하는 함수를 목적함수(objective function)라 한다

매개변수를 바꾸면 목적함수 값이 바뀐다



등고선으로 보는 목적함수



두 축은 매개변수 $\theta = (\beta_0, \beta_1)$ 이다. 등고선은 RSS가 같은 점들을 이은 선이며, 안쪽으로 갈수록 RSS가 작아지고 붉은 점에서 최소가 된다. 제곱 손실이면 $J = \text{RSS}/n$ 이라 등고선 모양과 최소점이 같다

Source: James et al. (2023), Figure 3.2.

최소화하는 값과 최솟값을 구분한다

$$\hat{\theta} = \arg \min_{\theta} J(\theta), \quad J(\hat{\theta}) = \min_{\theta} J(\theta)$$

기호	구하는 대상
$\arg \min_{\theta} J(\theta)$	목적함수를 최소화하는 매개변수
$\min_{\theta} J(\theta)$	목적함수의 최솟값

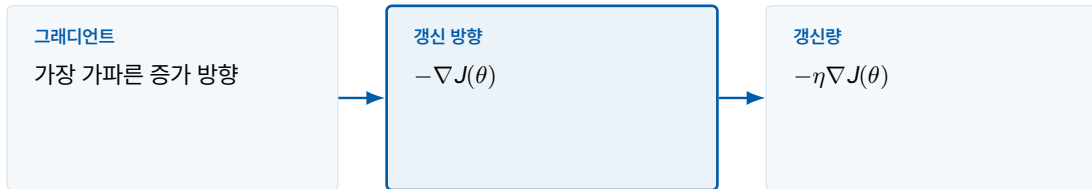
해를 수식으로 구하기 어려우면 반복해서 찾는다

명시적인 수식으로 나타낸 해를 닫힌 해(closed-form solution)라 한다

모델	닫힌 해	계산 방법
최소제곱 선형 회귀	있음	행렬 분해 등으로 계산
로지스틱 회귀	일반적으로 없음	반복 최적화
신경망	일반적으로 없음	반복 최적화

그래디언트의 반대 방향으로 갱신한다

미분 가능한 함수에서 그래디언트(gradient)는 국소적으로 가장 가파르게 증가하는 방향이다



그래디언트가 0이 아닐 때, 유클리드 거리 기준으로 그 반대 방향이 가장 가파른 감소 방향이다. $\eta > 0$ 은 한 번에 움직이는 정도를 정하는 수다

평균 손실에서 한 관측의 비중

평균을 쓰면 관측 수 n 이 달라도 J 의 규모가 비슷하게 유지된다

관측 수 n	관측별 가중치	J 에 더해지는 값
10	$1/10$	$\mathcal{L}(\hat{y}_i, y_i)/10$
1000	$1/1000$	$\mathcal{L}(\hat{y}_i, y_i)/1000$

모든 관측에 같은 가중치 $1/n$ 을 곱하므로, 손실이 큰 관측은 여전히 J 에 크게 기여한다

Checkpoint

Objective Surface

- 목적함수와 한 관측의 손실은 무엇이 다른가
- 제시된 RSS 등고선의 안쪽으로 갈수록 값은 커지는가 작아지는가
- 닫힌 해가 있는 경우와 없는 경우의 예를 각각 들어 보라

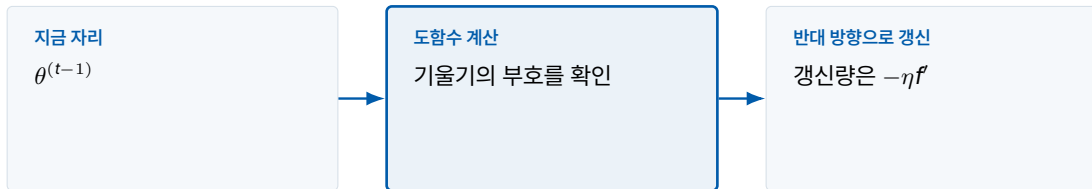
Section 2 • Gradient Descent

One-Dimensional Descent

한 번의 갱신을 직접 계산한다

한 줄로 적는 갱신

$$\theta^{(t)} = \theta^{(t-1)} - \eta f'(\theta^{(t-1)}) \quad (\text{매개변수가 하나인 목적함수 } f(\theta))$$



이 갱신 규칙(update rule)을 학습률(learning rate) η 로 반복 적용하는 절차가 경사하강법(gradient descent)이다

Source: MIT 6.390 (2026), Gradient Descent.

기울기의 반대 방향이 감소 방향이다

지금 자리에서	f 의 부호	갱신 방향 ($-\eta f$)
오른쪽으로 올라간다	양수	왼쪽으로 간다
오른쪽으로 내려간다	음수	오른쪽으로 간다
평평하다	0	움직이지 않는다

첫 갱신을 직접 계산하면

$f(x) = (x - 2)^2$ 에서 $x_0 = 4$, $\eta = 1/2$ 로 시작한다. 앞의 θ 자리에 x 를 쓴다

t	x	$f'(x) = 2(x - 2)$	다음 x
0	4	4	$4 - \frac{1}{2}(4) = 2$
1	2	0	$2 - \frac{1}{2}(0) = 2$

이 η 에서는 한 번 갱신해 최솟점에 도달하며, 이후 갱신량은 0이다

η 를 반으로 줄이면

t	x	$f'(x)$	다음 x
0	4	4	3
1	3	2	2.5
2	2.5	1	2.25
3	2.25	0.5	2.125

$\eta = 1/4$ 이면 네 번 갱신한 뒤 $x = 2.125$ 이며, 반복할수록 2에 가까워진다

경사하강법의 종료 조건

판정량	종료 기준	해석
J 의 변화	정한 허용 오차 이내	목적함수 변화가 작음
그라디언트의 크기	정한 허용 오차 이내	국소적인 변화율이 작음
θ 의 변화	정한 허용 오차 이내	갱신 폭이 작음
반복 횟수	최대 반복 수에 도달	계산 예산을 제한

종료 조건을 미리 정한다. 변화가 작다는 사실만으로 최솟점 도달이 보장되지는 않는다

Checkpoint

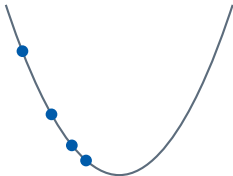
One-Dimensional Descent

- 갱신 식에서 기울기를 더하지 않고 빼는 까닭은 무엇인가
- $f(x) = (x - 2)^2$, $x_0 = 4$, $\eta = 1/2$ 일 때 첫 갱신 뒤의 x 는 얼마인가
- 종료 조건의 예를 들고, 작은 변화량만으로 최솟점 도달을 판단할 수 없는 이유를 설명하라

Learning Rate and Minima

학습률과 목적함수의 형태가 수렴에 미치는 영향을 설명한다

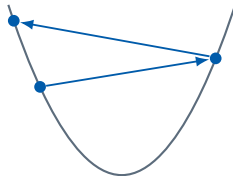
학습률이 갱신 경로를 바꾼다



η 가 작다



안정적으로 감소



η 가 크다

학습률이 작으면 갱신량이 작아 반복 횟수가 늘고, 지나치게 크면 최솟점을 지나쳐 발산할 수 있다. 점은 곡선 위에서 실제로 갱신한 위치다

매개변수와 하이퍼파라미터

구분	정하는 방법
매개변수 θ	학습 자료를 보고 반복해서 갱신
학습률 η 등 설정값	학습 전에 정하거나 검증 결과로 선택

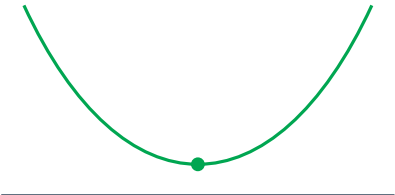
학습 절차의 설정값을 하이퍼파라미터(hyperparameter)라 한다

큰 학습률에서는 발산할 수 있다

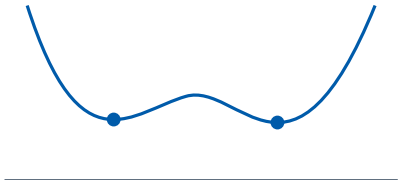
t	x	$f'(x)$	다음 x
0	4	4	$4 - 1.5(4) = -2$
1	-2	-8	$-2 + 1.5(8) = 10$
2	10	16	$10 - 1.5(16) = -14$

Caution $\eta = 1.5$ 에서는 $|x_t - 2|$ 가 갱신마다 커지고 목적함수 값이 발산한다

볼록 함수와 비볼록 함수



볼록 함수의 예



비볼록 함수의 예

볼록 함수(convex function)의 그래프는 임의의 두 점을 잇는 선분보다 위로 올라가지 않는다

국소 최소값과 전역 최소값

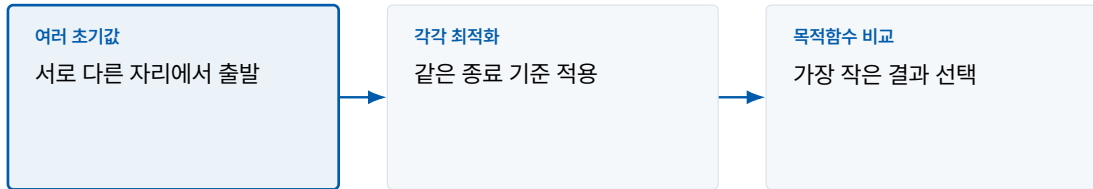
구분	볼록 함수	비볼록 함수
국소 최소값	모두 전역 최소값	전역 최소값과 다를 수 있음
최소점의 수	여러 개일 수도 있음	여러 개일 수도 있음
경사하강법	최소값이 있고 학습률이 충분히 작으면 전역 최소값에 수렴	초기값에 따라 결과가 다를 수 있음

국소 최소값(local minimum)은 주변에서, 전역 최소값(global minimum)은 정의역 전체에서 가장 작은 값이다

Caution 볼록성만으로 최소값의 존재나 최소점의 유일성이 보장되지는 않는다

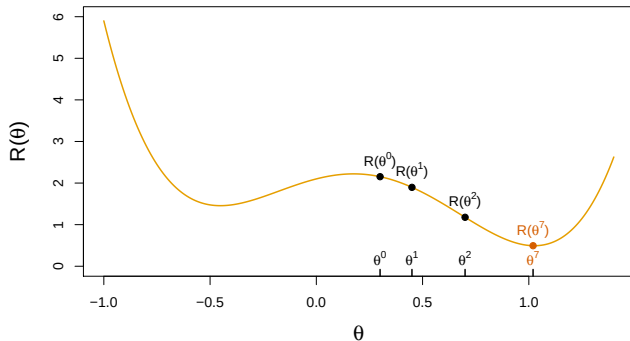
Source: Zhang et al. (2023), Convexity.

비볼록 최적화에서는 초기값을 바꾸어 비교한다



Caution 여러 초기값에서 반복해도 전역 최솟점에 도달한다는 보장은 없다

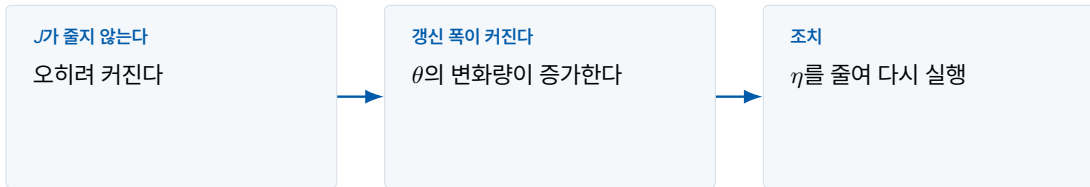
목적함수 위에서 보는 갱신 경로



가로축은 매개변수 θ , 세로축은 목적함수 $R(\theta)$ 이다. 이 초기값에서는 일곱 번 갱신하여 오른쪽 전역 최솟점에 도달한다

Source: James et al. (2023), Figure 10.17.

목적함수 값과 갱신 폭으로 발산을 진단한다



$\eta = 1.5$ 예에서 갱신 폭은 6, 12, 24로, $f(x)$ 는 4, 16, 64, 256으로 커진다

Checkpoint

Learning Rate and Minima

- η 가 너무 크면 x 와 목적함수 값은 어떻게 되는가
- 학습률과 최대 반복 수가 하이퍼파라미터인 이유는 무엇인가
- 볼록 함수의 최솟점은 반드시 하나인가

Section 4 • Gradient Descent

Multivariate Descent

여러 매개변수를 그래디언트로 함께 갱신한다

다변수 함수에서는 그래디언트를 사용한다

$$\theta^{(t)} = \theta^{(t-1)} - \eta \nabla_{\theta} J(\theta^{(t-1)})$$

매개변수의 차원	미분 결과
1차원	f 는 수 하나
p 차원	$\nabla_{\theta} J$ 는 성분이 p 개인 벡터(절편을 두면 $p + 1$ 개)

각 편미분을 모은 그래디언트로 모든 매개변수를 함께 갱신한다

선형 회귀의 그래디언트

$$J(\theta) = \frac{1}{n} \|X\theta - Y\|^2, \quad \nabla_{\theta} J = \frac{2}{n} X^{\top} (X\theta - Y)$$

$$\frac{\partial J}{\partial \theta_j} = \frac{2}{n} \sum_{i=1}^n (\theta^{\top} x_i - y_i) x_{ij}$$

x_i 는 X 의 i 번째 행을 세운 열벡터, x_{ij} 는 그 j 번째 성분이다. 이 성분을 $j = 1$ 부터 p 까지 모으면 $\frac{2}{n} X^{\top} (X\theta - Y)$ 가 된다

배열 모양으로 그래디언트 식을 점검한다

기호	배열 모양	역할
X	$n \times p$	관측이 행, 특성이 열
Y	$n \times 1$	관측값을 세로로 쌓은 열벡터
θ	$p \times 1$	특성마다 하나의 매개변수
$X\theta - Y$	$n \times 1$	관측마다의 오차
$X^T(X\theta - Y)$	$p \times 1$	θ 와 같은 모양

절편을 두려면 1로 채운 열을 X 앞에 붙인다. 그러면 X 는 $n \times (p + 1)$, θ 와 그래디언트는 $(p + 1) \times 1$ 이다

$X\theta$ 와 Y 를 모두 $n \times 1$ 열벡터로 맞추면 그래디언트가 θ 와 같은 모양이 된다

코드로 옮기면

```
1 Y = np.asarray(Y).reshape(-1, 1)
2 theta = np.zeros((X.shape[1], 1))
3 for t in range(T):
4     grad = 2 / n * X.T @ (X @ theta - Y)
5     theta = theta - eta * grad
```

Y를 열벡터로 맞춘 뒤 그래디언트를 계산하고 매개변수를 갱신한다

로지스틱 회귀에도 같은 갱신을 쓴다

단한 해가 없는 로지스틱 회귀도 같은 갱신 규칙으로 계수를 찾는다

$$J(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \sigma(x_i^\top \theta)), \quad \frac{\partial J}{\partial \theta_j} = \frac{1}{n} \sum_{i=1}^n (\sigma(x_i^\top \theta) - y_i) x_{ij}$$

모델	관측별 손실	그래디언트 (θ 와 같은 모양)
선형 회귀	제곱 오차	$\frac{2}{n} X^\top (X\theta - Y)$
로지스틱 회귀	로지스틱 손실 ℓ	$\frac{1}{n} X^\top (\sigma(X\theta) - Y)$

로지스틱 손실(logistic loss)은 이진 교차 엔트로피(binary cross-entropy)라고도 부르며, 관측 하나의 음의 로그 우도다

Source: MIT 6.390 (2026), Classification. James et al. (2023), Section 4.3.2.

로지스틱 회귀를 코드로 옮기면

Default 자료에서 절편 열과 잔액(천 달러)을 X 의 두 열로 두고, $\eta = 1$ 과 영벡터에서 시작한다

```
1 Y = np.asarray(Y).reshape(-1, 1)
2 theta = np.zeros((X.shape[1], 1))
3 for t in range(T):
4     prob = 1 / (1 + np.exp(-(X @ theta)))
5     grad = X.T @ (prob - Y) / n
6     theta = theta - eta * grad
```

반복 수 t	0	1	10	100	1,000	10,000
$J(\theta^{(t)})$	0.693	0.421	0.202	0.133	0.083	0.080

10,000번 뒤 $\theta = (\theta_0, \theta_1) = (-10.640, 5.492)$ 이고, 최대우도 추정값은 $(-10.651, 5.499)$ 다

반복할수록 평균 로지스틱 손실이 줄어 최솟값 0.0798에 다가간다

최소제곱해와 경사하강법의 결과

X의 열들이 선형 독립이면 최소제곱해는 유일하다

방법	결과	계산
직접 풀이	최소제곱해 계산	QR 분해 등 사용
경사하강법	조건을 만족하면 같은 해로 수렴	그래디언트 계산과 갱신 반복

예: $x = (1, 2, 3)$, $y = (1, 2, 2)$ 의 최소제곱해는 $(\beta_0, \beta_1) = (0.667, 0.5)$ 다. $\eta = 0.1$, 영벡터에서 시작하면 100번 뒤 (0.634, 0.514), 500번 뒤 (0.667, 0.500)이다

수렴하려면 충분히 작은 양의 학습률이 필요하며, 유한 번의 반복에서는 일반적으로 근사해를 얻는다

달힌 해가 있어도 반복 최적화를 쓰는 이유

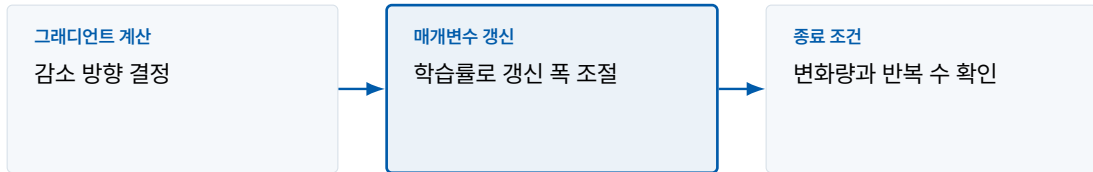
상황	직접 풀이의 고려점	반복 최적화의 특징
특성이 아주 많을 때	행렬 분해 비용과 메모리	행렬-벡터 곱으로 갱신
근사해로 충분할 때	전체 풀이 비용 발생	허용 오차에 따라 조기 종료

문제 크기와 필요한 정확도에 따라 계산 방법을 선택한다

경사하강법의 주요 용어

용어	정의
목적함수	학습에서 최소화하는 함수
닫힌 해	명시적인 수식으로 나타낸 해
갱신 규칙	현재 매개변수에서 다음 매개변수를 구하는 식
학습률	그래디언트에 곱하여 갱신 폭을 조절하는 η
하이퍼파라미터	학습 절차의 설정값
볼록 함수	그래프가 두 점을 잇는 선분보다 높지 않은 함수

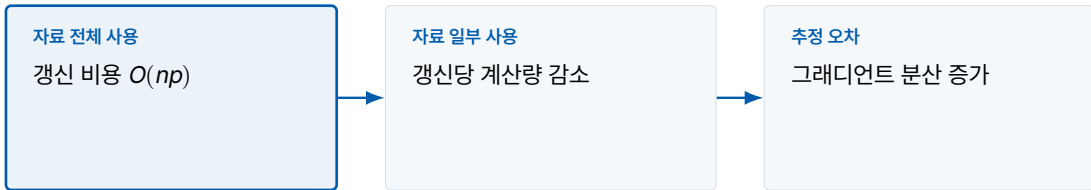
경사하강법의 반복 절차



종료 조건을 만족할 때까지 세 단계를 되풀이한다

전체 그래디언트는 자료 전체를 사용한다

한 번 갱신할 때 모든 관측의 그래디언트를 평균한다



한 번의 갱신에 드는 계산량

계산 단계	선형 회귀의 계산량	계산량을 결정하는 값
그래디언트 계산	$O(np)$	관측 수 n 과 특성 수 p
θ 갱신	$O(p)$	매개변수 수

$O(np)$ 는 계산량이 n 과 p 의 곱에 비례해 늘어난다는 뜻이다

선형 회귀에서는 한 번의 갱신 비용 대부분이 전체 자료의 그래디언트 계산에서 나온다

Checkpoint

Multivariate Descent

- 그래디언트의 배열 모양은 무엇과 같아야 하는가
- 닫힌 해가 있는데도 반복을 쓰는 까닭을 두 가지 들어 보라
- 코드에서 갱신에 해당하는 줄은 어느 것인가

What If You Only Look at Part of the Data?

Next Lecture: Stochastic Gradient Descent

- 한 번의 갱신에 자료 일부만 쓰기
- 전체 배치와 미니배치
- 무작위 추출에 따른 그래디언트의 변동