



부산대학교
PUSAN NATIONAL UNIVERSITY

Stochastic Gradient Descent

Fewer Examples per Step

Noisy Gradients, Mini-batches, and Epochs

BX3600087-140 / DS3600087-140 / AB3600087-140

기계학습개론

2026학년도 2학기

오용경

데이터사이언스전공

yongkyung.oh@pusan.ac.kr

Full-Gradient Cost

한 번의 갱신에 자료 전체를 쓸 때의 비용을 계산한다

한 번 갱신할 때마다 전체 그래디언트를 계산한다

∇J 는 각 관측의 그래디언트(gradient)를 평균한 것이다

$$\nabla J(\theta) = \frac{1}{n} \sum_{i=1}^n \nabla \mathcal{L}(\hat{y}_i(\theta), y_i)$$

$\hat{y}_i(\theta)$ 는 매개변수 θ 로 계산한 관측 i 의 예측값이다

DEFINITION

배치 경사하강법(batch gradient descent)

한 번의 갱신에 자료 전체의 평균 그래디언트를 사용하는 방법

$$\nabla J(\theta) = \frac{1}{n} \sum_{i=1}^n \nabla \mathcal{L}_i(\theta)$$

$\mathcal{L}_i(\theta) = \mathcal{L}(\hat{y}_i(\theta), y_i)$ 는 관측 i 의 손실이다. 배치(batch)는 한 번의 갱신에 쓰는 관측 묶음이며, 자료 전체일 수도 일부일 수도 있다

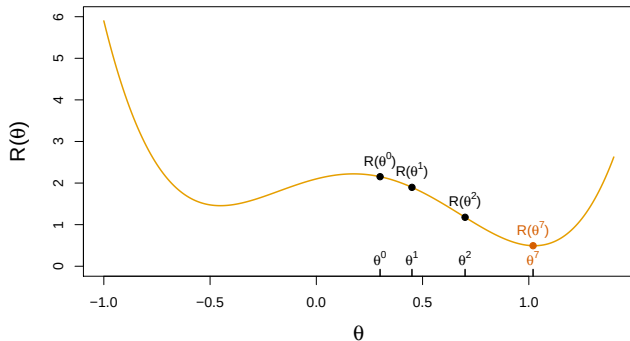
n 이 커지면

특성 수 p 가 같은 선형 모델의 계산량을 비교한다

관측 수 n	계산량의 규모 np	갱신 횟수
1,000	$1,000 \times p$	1
60,000	$60,000 \times p$	1

Caution 선형 모델의 그래디언트 계산은 $O(np)$ 이다. p 가 같을 때 관측 수가 60배면 계산량도 약 60배다

전체 그래디언트로 갱신하는 경로



같은 목적함수와 현재 매개변수에서 전체 그래디언트는 하나로 정해진다. 가로축은 반복 횟수가 아니라 매개변수다

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

Checkpoint

Full-Gradient Cost

- $\nabla J(\theta)$ 를 구하는 데 자료를 몇 개 보아야 하는가?
- 자료가 커지면 한 번의 갱신 비용은 어떻게 변하는가?
- 배치는 무엇을 뜻하는가?

One Example at a Time

자료 일부의 그래디언트로 전체 방향을 추정한다

관측 하나로 그래디언트를 추정한다

관측 n 개 중 하나를 같은 확률 $1/n$ 로 골라 갱신한다

$$\theta \leftarrow \theta - \eta \nabla \mathcal{L}(\hat{y}_i(\theta), y_i)$$

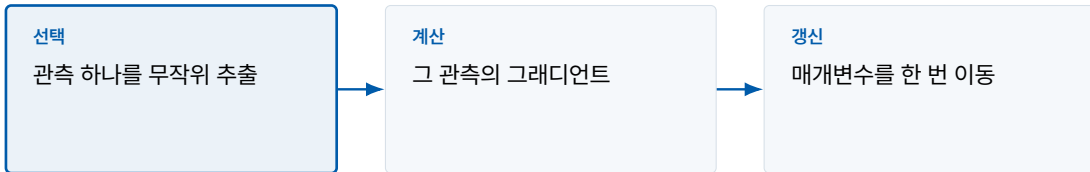
η 는 학습률(learning rate)

특성 수 p 가 같으면 한 번의 갱신 비용은 관측 수 n 과 관계없다

DEFINITION

확률적 경사하강법(stochastic gradient descent, SGD)

무작위로 고른 관측의 그래디언트로 매개변수를 갱신하는 방법



균등 추출의 기댓값은 전체 그래디언트와 같다

현재 매개변수 θ 를 고정하고 관측 하나를 같은 확률 $1/n$ 로 뽑는다

$$\mathbb{E}_i[\nabla \mathcal{L}_i(\theta)] = \frac{1}{n} \sum_{i=1}^n \nabla \mathcal{L}_i(\theta) = \nabla J(\theta)$$

관측 하나의 그래디언트는 매번 다를 수 있지만, 균등 추출한 기댓값은 전체 그래디언트와 같다

그래디언트의 한 성분을 직접 평균하면

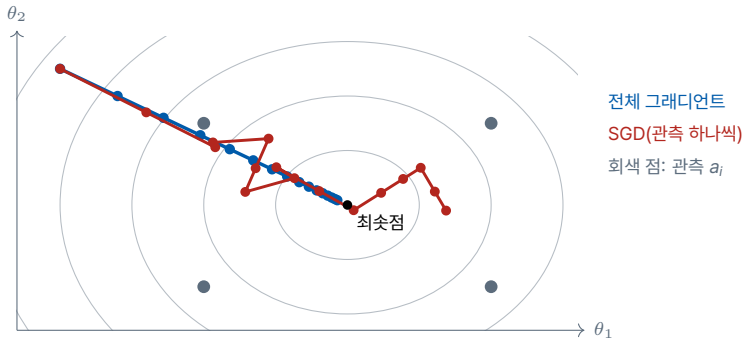
고른 관측	같은 θ 에서의 성분값
1	$g_1 = -2$
2	$g_2 = 1$
3	$g_3 = 4$

같은 확률로 뽑으면 $\mathbb{E}[g_i] = (-2 + 1 + 4)/3 = 1$ 이다

계산량과 그래디언트의 변동

비교 항목	배치 경사하강법	SGD
선형 모델의 갱신 비용	$O(np)$	$O(p)$
사용하는 방향	전체 그래디언트	관측 하나의 그래디언트
같은 θ 에서의 변동	없음	고른 관측에 따라 다름

같은 출발점에서 두 방법의 경로가 다르다



$J(\theta) = \frac{1}{4} \sum_i \|\theta - a_i\|_2^2$ 에서 출발점 $(-1, 5)$, $\eta = 0.1$ 로 15번 갱신했다. 전체 그래디언트는 최솟점 $(3, 2.5)$ 로 곧게 가고, SGD는 매번 고른 관측 쪽으로 끌려 흔들린다

섞고, 자르고, 갱신한다

에포크(자료 전체를 한 번 훑는 단위)마다 순서를 섞은 뒤 `batch_size`개씩 묶은 미니배치로 갱신한다

```
1 for epoch in range(num_epochs):
2     rng.shuffle(order)
3     for start in range(0, n, batch_size):
4         idx = order[start:start + batch_size]
5         grad = grad_loss(X[idx], y[idx], theta).mean(axis=0)
6         theta -= learning_rate * grad
```

Adapted from Zhang et al. (2023), Minibatch Stochastic Gradient Descent.

한 번에 쓰는 관측 수에 따른 계산량을 비교하면

관측 수 $n = 60,000$, 특성 수 $p = 784$ 인 선형 모델을 가정한다

방식	갱신 한 번의 계산량 Bp	전체 배치 대비
배치 경사하강법	47,040,000	1
미니배치 (32)	25,088	1/1875
SGD (1)	784	1/60000

B 는 한 번에 쓰는 관측 수다. Bp 는 비용 규모이며 실제 곱셈 횟수나 실행 시간은 아니다. 한 에포크의 계산량은 세 방식 모두 $np = 47,040,000$ 으로 같다

Checkpoint

One Example at a Time

- SGD의 갱신식은 배치 경사하강법과 어디가 다른가?
- 균등 추출한 그래디언트의 기댓값이 전체 그래디언트와 같은 이유는 무엇인가?
- 관측 하나로 갱신하면 갱신 한 번의 계산 비용은 배치 경사하강법과 어떻게 달라지는가?

Mini-Batches and Epochs

자료 일부를 묶어 그래디언트 분산과 계산량을 조절한다

미니배치는 자료 일부를 묶어 사용한다

DEFINITION

미니배치(mini-batch)

한 번의 갱신에 함께 쓰는 자료의 일부. 관측 수 B 를 미니배치 크기(batch size)라고 한다

관측 수	사용하는 그래디언트
$B = 1$	관측 하나의 그래디언트
$1 < B < n$	미니배치의 평균 그래디언트
$B = n$	자료 전체의 평균 그래디언트

미니배치 크기가 커지면 그래디언트 분산이 감소한다

$$g_B = \frac{1}{B} \sum_{j=1}^B g_{ij}, \quad \text{Var}(g_B) = \frac{\sigma^2}{B}$$

같은 θ 에서 독립적으로 추출한 한 성분의 분산을 σ^2 라 한다. B 가 네 배면 평균의 분산은 사분의 일이다

반복과 에포크

단위	뜻	세는 것
반복(iteration)	매개변수 갱신 한 번	미니배치 하나
에포크(epoch)	자료 전체를 한 번 봄	전체 자료

Caution 반복 수와 에포크 수는 서로 다르므로 학습량을 기록할 때 단위를 명시한다

에포크당 갱신 횟수

자료 60,000개를 미니배치 크기별로 나누어 본다

미니배치 크기	에포크당 갱신 횟수	한 번에 쓰는 관측 수
1	60,000	1
32	1,875	32
256	235	256 (마지막은 96)
60,000	1	60,000

마지막 작은 묶음도 사용하면, 한 에포크에 모든 관측을 한 번씩 사용한다

미니배치는 벡터화 연산을 활용한다

처리 단위	연산 호출	실행 특성
자료 한 건씩	작은 벡터 연산 B 회	호출 비용이 반복됨
B 건을 한 묶음으로	행렬 연산 1회	병렬 계산과 캐시 활용

미니배치는 관측 하나보다 그래디언트 변동이 작고, 벡터화된 계산도 활용한다

자료를 섞고 미니배치로 나눈다

단계	하는 일	이유
1	에포크마다 자료를 섞는다	특정 순서의 영향을 줄이기 위해
2	앞에서부터 B 개씩 자른다	정한 크기로 묶어 계산
3	마지막 묶음은 작을 수 있다	n 이 B 로 나누어떨어지지 않으면

섞지 않으면 에포크마다 저장된 순서대로 같은 미니배치가 만들어진다

미니배치 크기를 정할 때 확인할 것

고려할 항목	확인할 내용
메모리	미니배치가 클수록 많은 자료를 동시에 저장
그래디언트 변동	같은 조건에서 미니배치가 클수록 평균의 분산 감소
실행 효율	벡터화와 연산 장치의 활용도 측정
설정 예	32, 64, 128 등을 비교하되 보편적인 최적값은 없음

미니배치 크기는 계산 환경과 검증 결과를 보고 선택하는 하이퍼파라미터(hyperparameter)다

한 번에 쓰는 관측 수 B 로 구분하는 세 가지 갱신

방식	B	계산량	그라디언트
배치 경사하강법	n	$O(np)$	전체 평균
미니배치 SGD	$1 < B < n$	$O(Bp)$	묶음의 평균
SGD	1	$O(p)$	관측 하나

계산량은 특성 수가 p 인 선형 모델 기준이다. SGD라는 이름은 흔히 미니배치 SGD까지 함께 가리킨다

순서를 섞어 배치 구성의 편중을 줄인다

범주별로 정렬된 자료를 그대로 자르면 한 배치가 같은 범주로 채워질 수 있다

구성 방법	배치의 특성	학습에 미치는 영향
정렬 순서 유지	특정 범주에 편중될 수 있음	연속 갱신이 자료 순서에 좌우됨
순서를 섞은 뒤 분할	편중될 위험을 줄임	고정 순서의 영향을 완화

Caution 고정된 θ 에서 전체 그래디언트는 자료 순서와 무관하다. 섞으면 연속된 갱신이 자료 순서에 덜 좌우된다

Checkpoint

Mini-Batches and Epochs

- 반복과 에포크는 각각 무엇을 세는 단위인가?
- $n = 60,000$, $B = 32$ 이면 한 에포크에 몇 번 갱신하는가?
- 미니배치를 쓰는 두 가지 이유는 무엇인가?

Convergence Under Noise

그래디언트 추정값이 변동할 때 손실이 줄어드는지 평가한다

수렴 상태는 손실 곡선으로 판단한다

DEFINITION

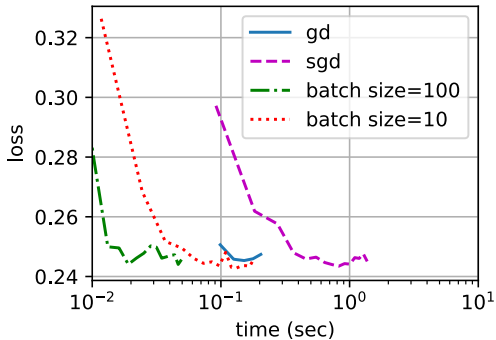
수렴(convergence)

반복이 진행되며 매개변수나 목적함수 값이 일정한 값에 가까워지는 성질

손실 곡선을 읽을 때는 가로축이 반복 횟수인지 에포크인지 먼저 구분한다.

손실 곡선이 평평해졌다는 사실만으로 최솟값 도달을 보장하지는 않는다

실행 시간에 따른 손실 감소



가로축은 실행 시간(로그 눈금), 세로축은 손실이다. 미니배치 크기마다 학습률을 다르게 둔 실험이므로, 모든 문제에서의 속도 순위를 뜻하지 않는다

Source: Zhang et al. (2023), Minibatch Stochastic Gradient Descent.

최솟점에서도 갱신이 계속될 수 있다

상태	결과
미분 가능한 내부 최솟점	전체 그래디언트는 0
개별 관측의 그래디언트	0이 아닐 수 있음
개별 그래디언트가 0이 아니면	양의 학습률에서 매개변수가 다시 이동

예: $\mathcal{L}_i(\theta) = (\theta - a_i)^2$, $a = (1, 2, 3)$ 이면 $\theta = 2$ 에서 전체 그래디언트는 0이지만 개별 그래디언트는 2, 0, -2다. $\eta = 0.25$ 로 관측 하나를 뽑아 갱신하면 θ 는 1.5, 2, 2.5 중 하나가 된다

Caution 전체 그래디언트가 0이어도 SGD의 개별 갱신량은 0이 아닐 수 있다

학습률을 줄이면 갱신 폭도 줄어든다

같은 그래디언트에서 갱신량의 크기는 학습률에 비례한다

$$\Delta\theta = -\eta g, \quad \|\Delta\theta\| = \eta \|g\| \quad (\eta > 0)$$

학습률 설정	갱신에 미치는 영향
일정하게 유지	개별 그래디언트의 변동이 갱신량에 계속 반영
점차 감소	같은 그래디언트에서 갱신량과 그 변동 폭이 감소

학습률과 미니배치 크기의 선택

설정값	큰 값의 영향	작은 값의 영향
η	발산할 수 있다	수렴이 느리다
미니배치 크기	평균의 분산 감소, 메모리 증가	평균의 분산 증가, 메모리 감소

학습률과 미니배치 크기는 학습 상태와 검증 결과를 보고 선택한다

비교 기준에 따라 가로축을 선택한다

가로축	고정되는 비교 단위	방식마다 달라지는 양
반복 횟수	매개변수 갱신 수	사용한 관측 수와 계산량
에포크	전체 자료를 사용한 횟수	갱신 수와 실행 시간
실행 시간	사용한 시간	관측 수와 갱신 수

반복 횟수, 에포크, 실행 시간은 서로 다른 비교 단위이므로 함께 기록한다

검증 손실로 종료 시점을 정한다

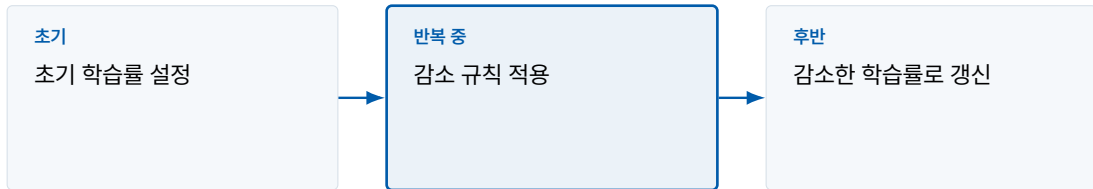
관찰값	대응
검증 손실의 개선이 멈춤	미리 정한 기준에 따라 학습 중단 검토
학습 손실의 변동이 큼	여러 갱신의 추세와 이동 평균 확인
최대 에포크 수에 도달	정한 계산 예산에 따라 종료

학습 손실은 학습 자료, 검증 손실은 검증 자료에서 계산한 목적함수 값이다

종료 시점 선택에는 검증 자료를 쓰며, 최종 테스트 자료를 반복해서 사용하지 않는다

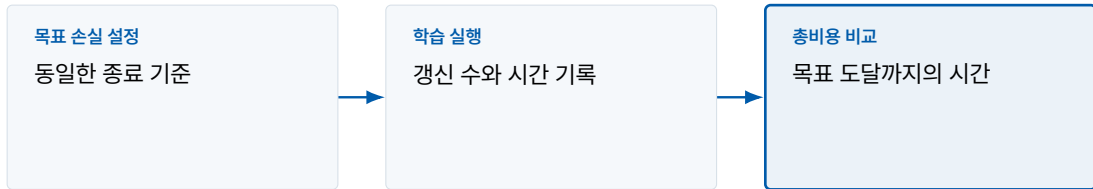
학습률을 정해 둔 규칙에 따라 줄인다

학습이 진행될수록 학습률을 줄이는 방법을 사용할 수 있다



학습률을 너무 빨리 줄이면 충분히 학습하기 전에 갱신 폭이 지나치게 작아질 수 있다

갱신당 계산량만으로 총 실행 시간을 판단할 수 없다



미니배치 구현의 점검 항목

점검 항목	확인할 내용
관측 선택	한 에포크를 분할할 때 빠지거나 중복된 관측이 없는지 확인
그래디언트 평균	마지막 묶음도 실제 관측 수로 나눔
배열 모양	그래디언트와 매개변수의 배열 모양이 일치
학습 기록	갱신 수, 에포크, 실행 시간을 구분

Checkpoint

Convergence Under Noise

- 손실 곡선의 가로축을 반복으로 둘 때와 에포크로 둘 때 무엇이 달라지는가?
- 전체 그래디언트가 0이어도 SGD의 갱신이 계속될 수 있는 이유는 무엇인가?
- 미니배치 크기는 그래디언트의 분산과 계산 비용에 어떤 영향을 주는가?

Can One Mistake Move the Boundary?

Next Lecture: The Perceptron

- 선형 분류기의 예측과 오분류
- 잘못 분류한 관측을 이용한 가중치 갱신
- 퍼셉트론의 학습 규칙