4. 최적화 알고리즘

날짜	@2023년 11월 13일

▼ 목차

미니 배치 경사하강법

1️⃣ Batch vs. Mini-batch

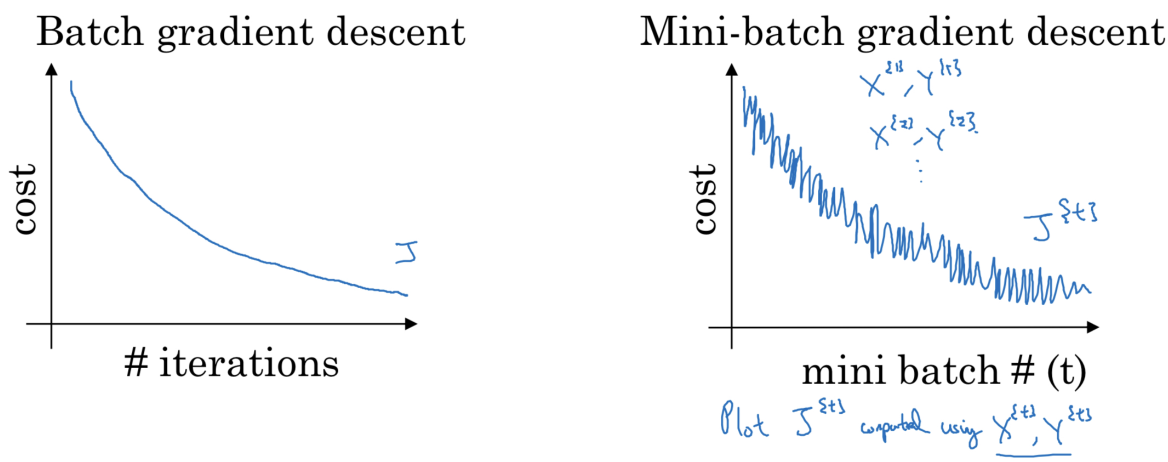
- 배치 경사 하강법
  - 전체 훈련 샘플에 대해 훈련 후 경사 하강 진행

- 미니배치 경사 하강법
    - 전체 훈련 샘플을 작은 훈련 세트인 미니배치로 나눈 후, 미니배치 훈련 후 경사 하강 진행
    - 배치 경사 하강법은 큰 데이터 세트를 훈련하는 데 많은 시간이 들어 경사 하강을 진행하기까지 오랜 시간이 걸리므로, 이를 해결하기 위해 작은 훈련 세트인 미니배치로 나누어 훈련 후 경사 하강을 진행
    - 예를 들어 전체 훈련 세트 크기가 5,000,000이라고 할 때 이를 사이즈가 1,000인 미니배치 5,000개로 나누어 훈련 및 경사 하강법을 진행

2️⃣ Mini-batch Gradient Descent

- 표기법
    - i번째 훈련 세트: $x^{(i)}$
    - l번째 신경망의 값: $z^{[l]}$
    - t번째 미니배치: $X^{\{t\}}, Y^{\{t\}}$

3️⃣ Understanding Mini-batch Gradient Descent



- 배치 경사 하강법에서는 한 번의 반복을 돌 때마다 비용 함수의 값은 계속 작아져야 함
- 미니배치 경사 하강법에서는 전체적으로는 비용 함수가 감소하는 경향을 보이지만 많은 노이즈가 발생함
- 학습 속도를 조절하기 위해 미니배치 사이즈의 최적값을 찾아내는 것이 중요함
    - 만약 훈련 세트가 작다면(2,000개 이하) 모든 훈련 세트를 한 번에 학습시키는 배치 경사 하강을 진행함

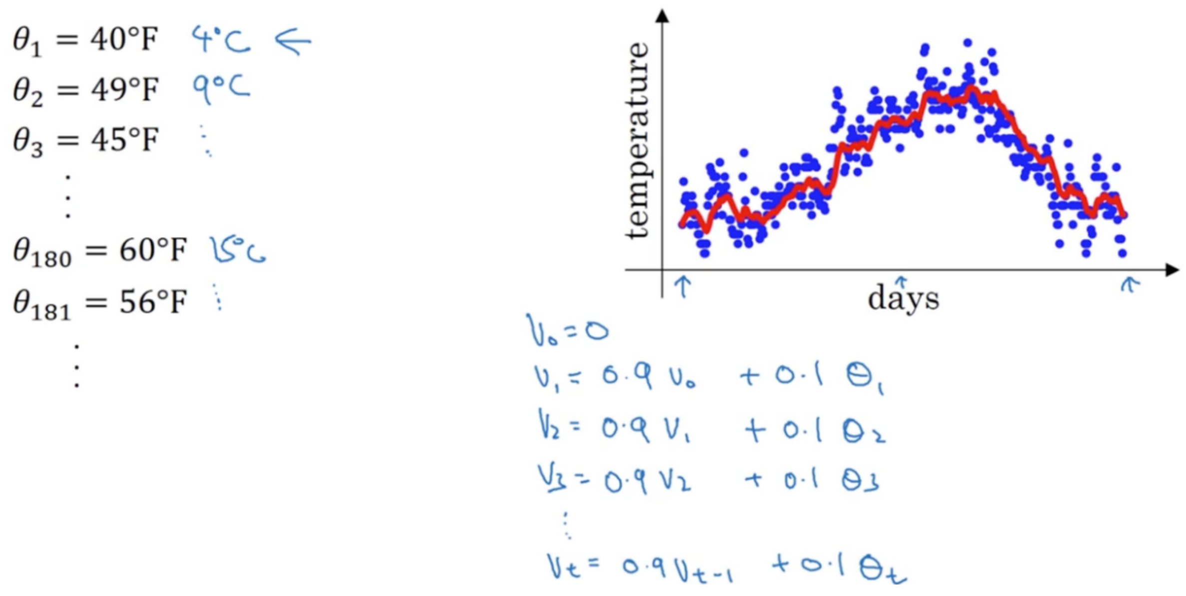
- 훈련 세트가 2,000개보다 클 경우 일반적으로 64, 128, 256, 512와 같은 2의 제곱수로 미니배치 사이즈를 설정함

지수 가중 이동 평균

1️⃣ Exponentially Weighted Averages

- 최근의 데이터에 더 많은 영향을 받는 데이터들의 평균 흐름을 계산하기 위한 알고리즘으로, 최근 데이터 지점에 더 높은 가중치를 줌
- 구현 시 아주 적은 메모리를 사용한다는 장점이 있음

Temperature in London Example



- θ_t: t번째 날의 기온
- 지수 가중 이동 평균(v_t): $v_t = \beta v_{t-1} + (1-\beta)\theta_t$
    - $\frac{1}{1-\beta}$ 기간 동안 기온의 평균을 의미
        - β = 0.9일 때 10일의 기온 평균
        - β = 0.5일 때 2일의 기온 평균
- β: 하이퍼파라미터로 최적의 값을 찾아야 하며, 일반적으로 0.9

2️⃣ Implementing Exponentially Weighted Averages

- β = 0.9일 때 특정 시점에서의 지수 가중 이동 평균 식
  - $v_{100} = 0.1\theta_{100} + 0.1 \times 0.9\theta_{99} + 0.1 \times (0.9)^2\theta_{98} + \cdots$
- 이를 그림으로 표현하면 v_{100}을 기준으로 보았을 때 지수적으로 감소하는 그래프가 나타남
  - v_{100}이 각각의 요소에 지수적으로 감소하는 요소$(0.1 \times (0.9)^n)$를 곱하여 더한 것이기 때문
- 얼마의 기간이 이동하면서 평균이 구해졌는지?
  - $\beta = (1-\varepsilon)$라고 정의할 때, $(1-\varepsilon)^n = \frac{1}{e}$을 만족하는 n이 그 기간이 되며 일반적으로 $\frac{1}{\varepsilon}$로 구할 수 있음

지수 가중 이동 평균의 편향보정

1️⃣ Bias Correction

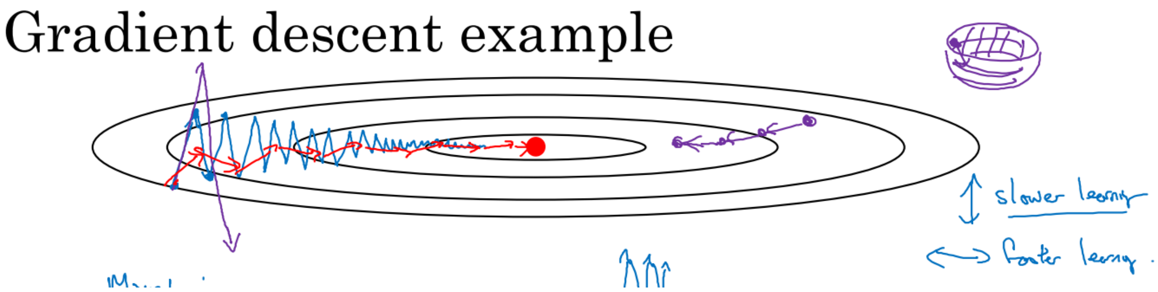
🔹 편향 보정으로 평균을 더 정확하게 계산할 수 있음

- 지수 가중 이동 평균식에 따르면, $t = 1$일 때 $(1-\beta)$를 곱한 값이 첫 번째 값이 되며 이는 우리가 원하는 실제 v_1값과 차이가 나게 됨
- 따라서 $v_t/(1-\beta^t)$를 취하여 초깃값에서 실제값과 비슷해지도록 함
- 머신러닝에서는 시간이 지남에 따라 $(1-\beta^t)$가 1에 가까워져서 결국 원하는 값과 일치하게 되므로 일반적으로 구현하지 않음

Momentum 최적화 알고리즘

1️⃣ Gradient Descent

- Momentum을 이용하면 매 단계의 경사 하강 정도를 부드럽게 만들어줄 수 있음



알고리즘은 다음과 같음

- $V_{dW} = \beta_1 V_{dW} + (1-\beta_1)dW$
- $w := w - \alpha V_{dW}$

2️⃣ Implementation Details

- 일반적으로 Momentum 알고리즘에서는 편향 추정을 실행하지 않는데, 그 이유는 step이 10단계 이상을 넘어가면 이동평균이 준비되어 편향 추정이 더 이상 일어나지 않기 때문임

RMSProp 최적화 알고리즘

1️⃣ RMSProp

알고리즘은 다음과 같음

- $S_{dW} = \beta_2 S_{dW} + (1-\beta_2)dW^2$
  - dW^2: 요소별 제곱
- $w := w - \alpha \frac{dW}{\sqrt{S_{dW} + \epsilon}}$

- 미분값이 큰 곳에서는 업데이트 시 큰 값으로 나눠주기 때문에, 기존 학습률보다 작은 값으로 업데이트되어 진동을 줄이는 데 도움이 됨
- 반면 미분값이 작은 곳에서는 업데이트 시 작은 값으로 나눠주기 때문에, 기존 학습률보다 큰 값으로 업데이트되어 더 빠르게 수렴하는 효과를 줌

Adam 최적화 알고리즘

1️⃣ Adam Optimization Algorithm

- Adam은 Momentum과 RMSProp을 섞은 알고리즘
    - Adaptive moment estimation의 약자

💠 알고리즘은 다음과 같음

- $V_{dW} = 0, S_{dW} = 0$로 초기화
- Momentum 항: $V_{dW} = \beta_1 V_{dW} + (1 - \beta_1) dW$
- RMSProp 항: $S_{dW} = \beta_2 S_{dW} + (1 - \beta_2) dW^2$
- Bias correction: $V_{dW}^{correct} = \frac{V_{dW}}{1-\beta_1^t}, S_{dW}^{correct} = \frac{S_{dW}}{1-\beta_2^t}$
- $w := w - \alpha \frac{V_{dW}^{correct}}{\sqrt{S_{dW}^{correct} + \epsilon}}$

학습률 감쇠

1️⃣ Learning Rate Decay

- 작은 미니배치일수록 잡음이 심하므로 일정한 학습률로는 최적값에 수렴하기 어려움
- 학습률 감쇠를 통해 점점 학습률을 작게 줌으로써 최적값을 더 빨리 찾을 수 있음

2️⃣ Learning Rate Decay Methods

- 1 epoch = 전체 데이터를 1번 훑고 지나가는 횟수
- step별로 α를 다르게 설정

Different Methods

- $\alpha = \frac{1}{1 + decay\ rate\ \times epoch\ num} \alpha_0$
- $\alpha = 0.95^{epoch\ num} \alpha_0$ (exponential decay)
- $\alpha = \frac{k}{\sqrt{epoch\ num}} \alpha_0$

- $\alpha = \frac{k}{\sqrt{batch\ num}}\alpha_0$

지역 최적값 문제

1️⃣ Local Optima in Neural Networks

- 일반적으로 고차원 비용함수에서 경사가 0인 경우 지역 최적값이 아닌 안장점일 가능성이 높음
- 안정지대는 안장점으로 향하는 구간으로, 미분값이 아주 오랫동안 0에 가깝게 유지되는 지역
- 대개 충분히 큰 Network 학습 시 지역 최적값에 갇히는 일은 거의 없음

2️⃣ Problem of Plateaus

- 안정지대는 경사가 거의 0에 가깝기 때문에 학습속도가 느려짐
- 다른 쪽으로 방향 변환이 없다면 안정지대에서 벗어나기 어려움
  - Adam, RMSprop 등의 알고리즘을 활용하여 해결할 수 있음

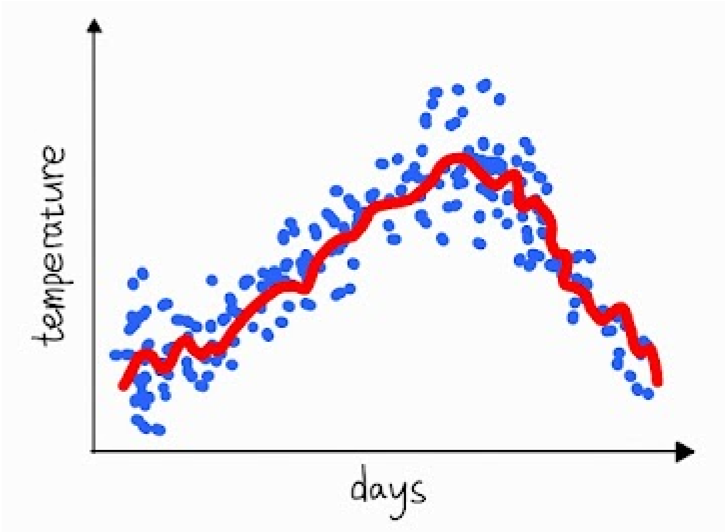
출석퀴즈 오답노트

▼ 1. 8번째 미니배치, 7번째 입력, 3번째 레이어의 활성화 표기

a^(7){8}[3] (X)

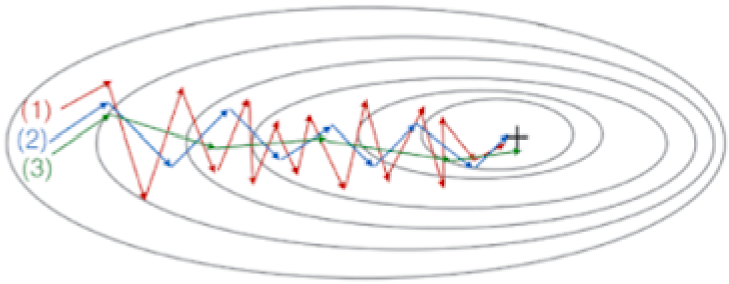
a^[3]{8}(7) (O)

▼ 7. 지수 가중 평균에서 β값의 변화에 따른 그래프 변화 양상



- β를 감소시키면 빨간 선 내에서 더 많은 진동 발생
- β를 증가시키면 빨간 곡선이 약간 오른쪽으로 이동

▼ 8. 그래프를 보고 경사 하강법 구분



(1) 경사 하강법

(2) 모멘텀을 적용한 경사 하강법 (작은 β)

(3) 모멘텀을 적용한 경사 하강법 (큰 β).

▼ 9. 심층 신경망에서 배치 경사 하강법을 이용하자 비용함수를 작게 하는 매개변수 값을 찾는 데 너무 오랜 시간이 걸리는 경우, 다음과 같은 방법을 시도할 수 있음

- Adam 사용해보기
- 가중치에 대한 더 나은 무작위 초기화 시도해보기
- 학습률 α 조정해 보기
- 미니배치 경사 하강법 시도해 보기

▼ 10. Adam에 관해 적절하지 않은 설명

- 각 매개변수마다 학습률을 조정하는 방식으로 작동한다. (O)
- Adam은 미니배치보다 배치 경사 계산에 사용해야 한다. (X)