CHAPTER6

2021년 3월 7일 일요일 오후 8:28

학습목표

- 신경망 학습의 핵심 개념들을 알아보자!

6.1 매개변수 갱신

P.180 참고

신경망 학습 순서
1단계 : 미니배치
2단계 : 기울기 산출
3단계 : 매개변수 갱신
- 가중치 매개변수를 기울기 방향으로 아주 조금 갱신한다.

4단계 : 반복

6.1.2 확률적 경사하강법(SGD)

확률적 경사 하강법(SGD) : 매개변수의 기울기를 구해, 기울어진 방향으로 매개변수 값을 갱신하는 일을 반복하여 최적의 값에 다가간다.

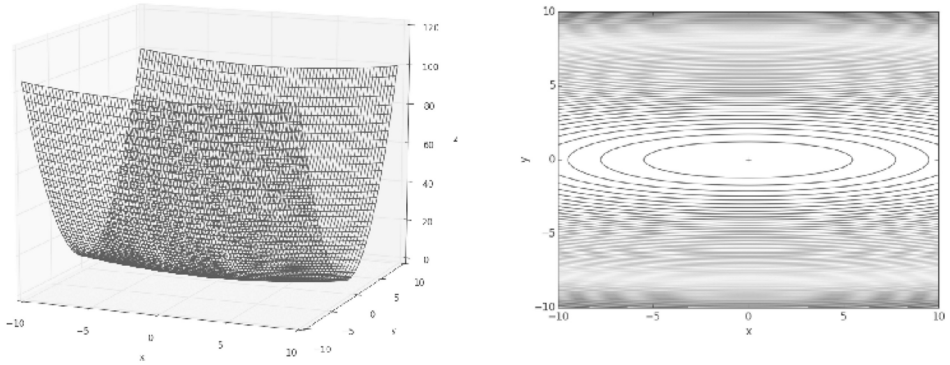
6.1.3 SGD의 단점

SGD는 단순하고 구현하기 쉽지만, 문제에 따라서 비효율적일 때가 있다.

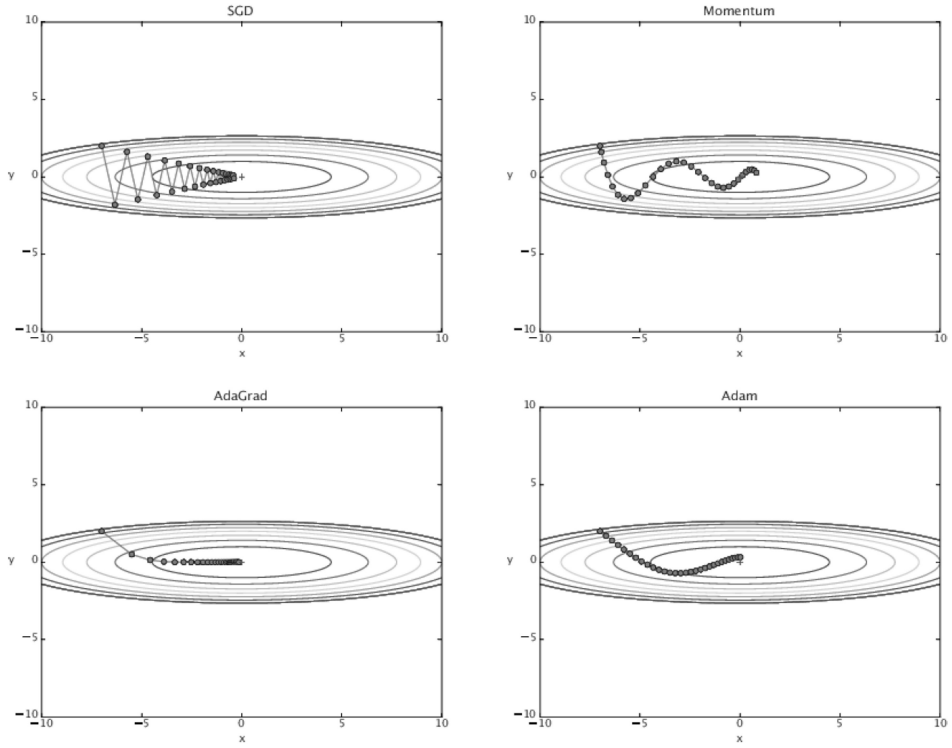
ex)

$$f(x,y) = \frac{1}{20}x^2 + y^2$$

위 함수의 그래프와 등고선



각 기법의 최적화 갱신 경로



비등방성 함수 (방향에 따라 성질(기울기)이 달라지는 함수)에서는 탐색경로가 비효율적이다.

- 기울어진 방향이 본래의 최솟값과 다른 방향을 가리키기 때문 (P.193 그림 6-2 참고)

6.1.4 모멘텀

모멘텀은 운동량이나 가속도를 의미하는 물리학 용어

$$\mathbf{v} \leftarrow \alpha\mathbf{v} - \eta\frac{\partial L}{\partial \mathbf{W}}$$

$$\mathbf{W} \leftarrow \mathbf{W} + \mathbf{v}$$

속도

기울기 방향으로 힘을 받아 물체가 가속된다.

- 위의 최적화 갱신경로 그림을 통해서 SGD보다 x축 방향으로 빠르게 다가가 지그재그 움직임이 줄어드는 것을 확인 할 수 있다.

6.1.5 AdaGrad

개별 매개변수에 적응적으로 학습률을 조정하면서 학습을 진행한다.

학습률이 너무 작으면 학습 시간이 길어지고, 반대로 너무 크면 발산하여 학습이 제대로 이루어 지지 않는 문제가 발생하는데 이 학습률을 정하는 효과적 기술로 '학습률 감소'가 있다.

학습률 감소 : 학습을 진행하면서 학습을 점차 줄여가는 방법

$$\mathbf{h} \leftarrow \mathbf{h} + \frac{\partial L}{\partial \mathbf{W}} \odot \frac{\partial L}{\partial \mathbf{W}}$$

$$\mathbf{W} \leftarrow \mathbf{W} - \eta \frac{1}{\sqrt{\mathbf{h}}} \frac{\partial L}{\partial \mathbf{W}}$$

매개변수의 원소 중에서 많이 움직인 원소는 학습률이 낮아진다. 즉 학습률 감소가 매개변수의 원소마다 다르게 적용된다.

RMSProp : AdaGrad는 과거의 기울기를 제곱하여 계속 더해가기 때문에 학습을 진행할수록 갱신 강도가 약해진다. 이 문제점을 개선한 기법이 바로 RMSProp 이다.

RMSProp는 먼 과거의 기울기는 서서히 잊고 새로운 기울기 정보를 크게 반영하여 과거 기울기의 반영 규모를 기하급수적으로 감소시킨다.

- y축 방향은 기울기가 커서 처음에는 크게 움직이지만, 그 큰 움직임에 비례해 갱신 정도도 큰 폭으로 작아지도록 조정된다. 그래서 y축 방향으로 갱신 강도가 빠르게 약해진다.

6.1.6 Adam

모멘텀과 AdaGrad 기법을 융합한 듯한 방법
(완전히 정확한 설명은 아니다.)

6.1.7 어느 갱신 방법을 이용할 것인가

위의 최적화 기법 비교 그림을 보면 AdaGrad가 가장 좋은 방법 같아 보이지만, 풀어야 할 문제가 무엇이냐에 따라 다르다.

6.1.8 MNIST 데이터셋으로 본 갱신 방법 비교

P.201 그림 6-9 참고

6.2 가중치의 초깃값

가중치의 초깃값을 어떻게 설정에 따라 신경망 학습에 큰 영향을 끼친다.

6.2.1 초깃값을 0으로 하면

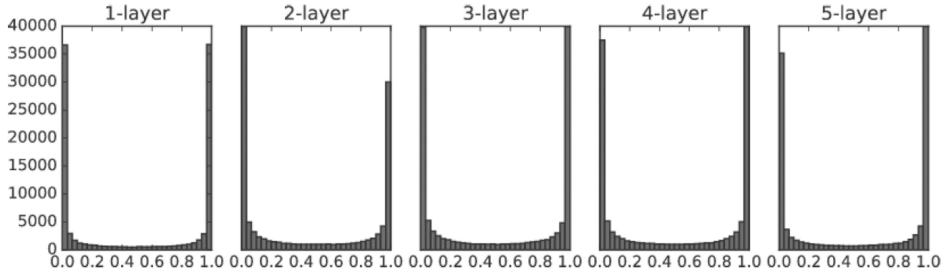
가중치 초깃값을 0으로 하면 습이 올바로 이뤄지지 않는다.
(가중치를 균일한 값으로 설정해서는 안된다.)

가중치가 균일한 값으로 설정되면 초깃값에서 시작하고 갱신을 거쳐도 같은 값을 유지한다.
(가중치를 여러 개 갖는 의미가 없다.)

6.2.2 은닉층의 활성화값 분포

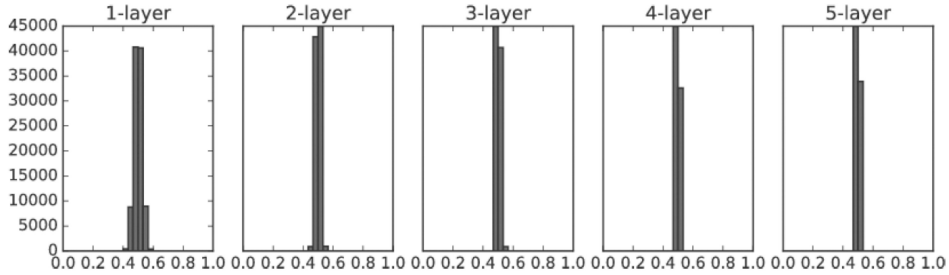
가중치의 초깃값에 따라 은닉층 활성화값들의 변화 확인
(활성화 함수로 시그모이드 함수를 사용하는 5층 신경망에 무작위로 생성한 입력 데이터를 흘리며 각 층의 활성화값 분포를 히스토그램으로 확인)

1. 가중치의 표준편차 1
   (P.204 코드 참고)



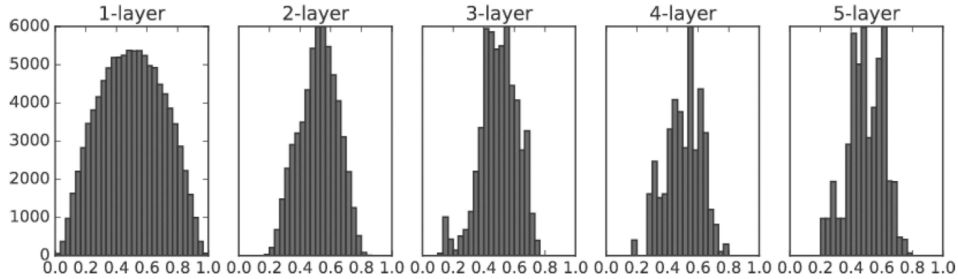
   - 기울기 소실 : 활성화 함수로 시그모이드 함수를 사용하면 여러 층을 거칠수록 출력이 0에 가까워지다가 사라진다.

2. 가중치의 표준편차 0.01
   (P.205 코드 참고)



- 기울기 소실이 일어나지는 않지만 다수의 뉴런이 거의 같은 값을 출력하여 표현력을 제한한다.(여러개의 뉴런이 의미가 없어진다.)

3. Xavier 초깃값
   앞 계층의 노드가 n개라면 표준편차가 √(1/n) 인 분포를 사용
   (P.206 코드 참고)



- 층이 깊어지면서 형태가 다소 일그러지지만, 앞의 방식보다 넓게 분포되는 것을 볼 수 있다.
- tanh함수를 사용하면 위의 일그러짐 현상을 개선 할 수 있다.

6.2.3 ReLU를 사용할 때의 가중치 초깃값

Xavier 초깃값은 활성화 함수가 선형인 것을 전제로 이끈 결과이다.

- sigmoid, tanh

ReLU활성화 함수를 이용할 때는 ReLU에 특화된 He 초깃값을 이용하길 권장한다.

He 초깃값
앞 계층의 노드가 n개라면 표준편차가 √(2/n) 인 분포를 사용
P.208 그림 6-14 참고

가중치 초깃값 정리

활성화 함수로 ReLU를 사용할때는 He 초깃값을, sigmoid나 tanh등의 S자 모양 곡선일 때는 Xavier 초깃값을 사용한다.

6.2.4 MNIST 데이터셋으로 본 가중치 초깃값 비교

P.209 그림 6-15 참고