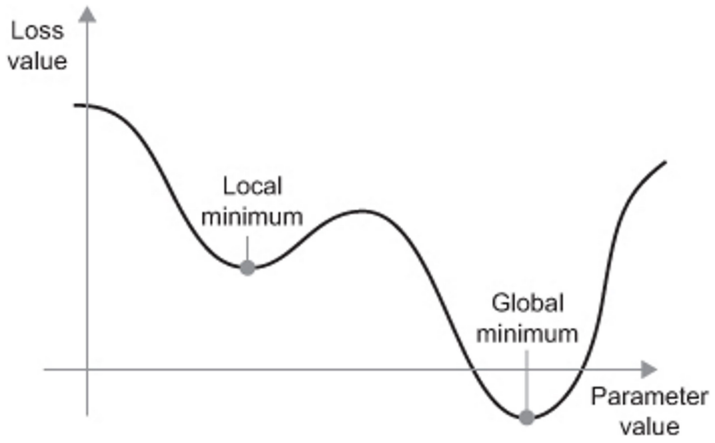
05. 최적화 알고리즘

2020년 12월 30일 수요일 오후 9:45

● 경사 하강법은 기본적으로 기울기를 바탕으로 가중치와 bias를 조금씩 조정하여 오차가 최소화 되도록 신경망을 최적화

● 최적화 알고리즘 개요



Global Minimum에 도착하기 위한 구체적인 전략

1. 확률적 경사 하강법(Stochastic Gradient Descent, SGD)

● 가중치와 bias를 수정하기 위해 반복학습을 할 때, 전체 샘플 데이터를 사용하지 않고 무작위로 일부 샘플 데이터를 선택하여 수행

● 가중치와 bias를 수정하는 식은 학습률을 이용하는 방법 그대로

$$w = w - \eta \frac{\partial \mathrm{E}}{\partial w}$$

$$b = b - \eta \frac{\partial \mathrm{E}}{\partial b}$$

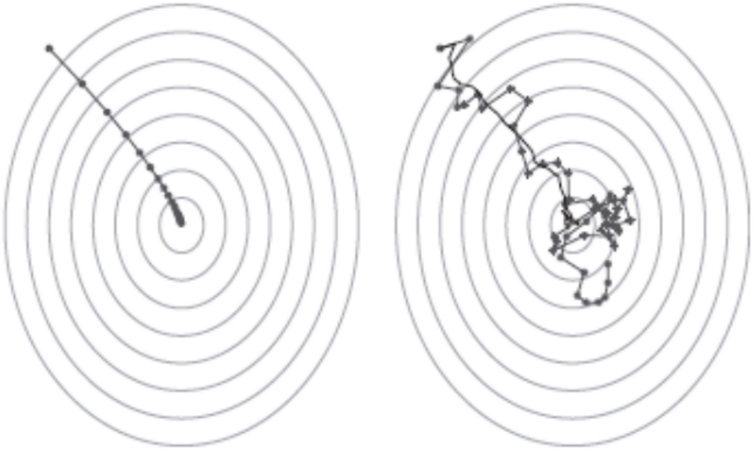
1) SGD의 장점

● 전체 샘플 대신 일부 샘플을 이용하기 때문에 전체 학습시간이 단축됨

● 무작위로 샘플을 선택하기 때문에 local minima 에 잘 빠지지 않음

2) SGD의 단점

● 학습의 진행과정에 따른 수정량을 유연하게 조정할 수 없음



2. 모멘텀(Momentum)
   ● SGD 에 물리학적 관성 기법을 적용한 알고리즘

$$w = w - \eta\frac{\partial \mathrm{E}}{\partial w} + \alpha\Delta \mathrm{w}$$
$$b = b - \eta\frac{\partial \mathrm{E}}{\partial b} + \alpha\Delta b$$

   ● α 는 모멘텀 상수(일반적으로 0.9 사용)

1) 장점
   ● local minima에 잘 빠지지 않음

2) 조정해야 하는 상수가 늘어남

3. 아다그라드 (Adaptive Gradient : AdaGrad)
   ● 학습이 진행되면서 알고리즘에 의해 학습률이 조금씩 감소함

$$h = h + (\frac{\partial \mathrm{E}}{\partial w})^2$$

$$w = w - \eta\frac{1}{\sqrt{h}}\frac{\partial \mathrm{E}}{\partial w}$$

   bias도 위와 동일

   ● 첫 번째 식에서 h는 학습이 진행되면서 계속 증가
   ● 두 번째 식은 h가 분모에 있으므로 감소
   ● h는 가중치별로 계산되므로, 그 때 까지의 총 수정량이 적은 가중치는 새로운 수정량이 증가, 또는 vice versa
   ● 이런 로직 때문에 처음에는 넓은 영역에서 탐색을 시작해 점차 탐색범위를 좁혀가는 효율적인 탐색이 가능

1) 장점

● 조정해야할 하이퍼 파라미터가 η 밖에 없음

2) 단점

● 수정량이 계속해서 감소하기 때문에 도중에 수정량이 0이 되어버려 더는 최적화가 진행되지 않을 수 있음

4. RMSProp

● AdaGrad는 최솟값에 도달하기 전에 학습률이 0에 수렴할 수 있음

● AdaGrad는 간단한 convex(볼록) 함수에서는 잘 동작하지만, 복잡한 다차원 곡면 함수에서는 성능이 떨어짐. 즉, 기울기의 단순한 누적만으로는 부족함

$$h = \rho h + (1 - \rho)\left(\frac{\partial \mathrm{E}}{\partial w}\right)^2$$

$$w = w - \eta \frac{1}{\sqrt{h}} \frac{\partial \mathrm{E}}{\partial w}$$

** 일반적으로 ρ(로) $= 0.9$ 사용

5. Adam (ADAptive Moment estimation)

● Momentum 과 AdaGrad를 융합한 방법

● 식이 굉장히 복잡

● 뭘 써야 할 지 모르겠으면 걍 Adam 써라

● 일반적으로 가장 성능이 좋음(지금까지는)

** 각 알고리즘별 성능 시각화 https://seamless.tistory.com/38