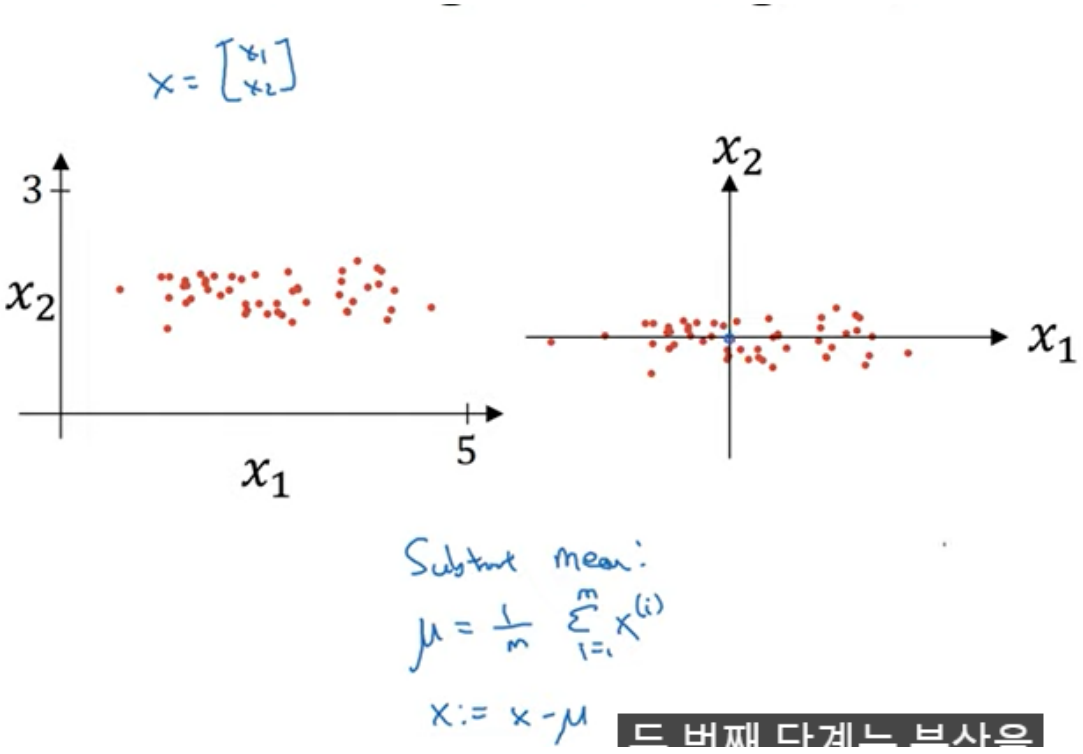
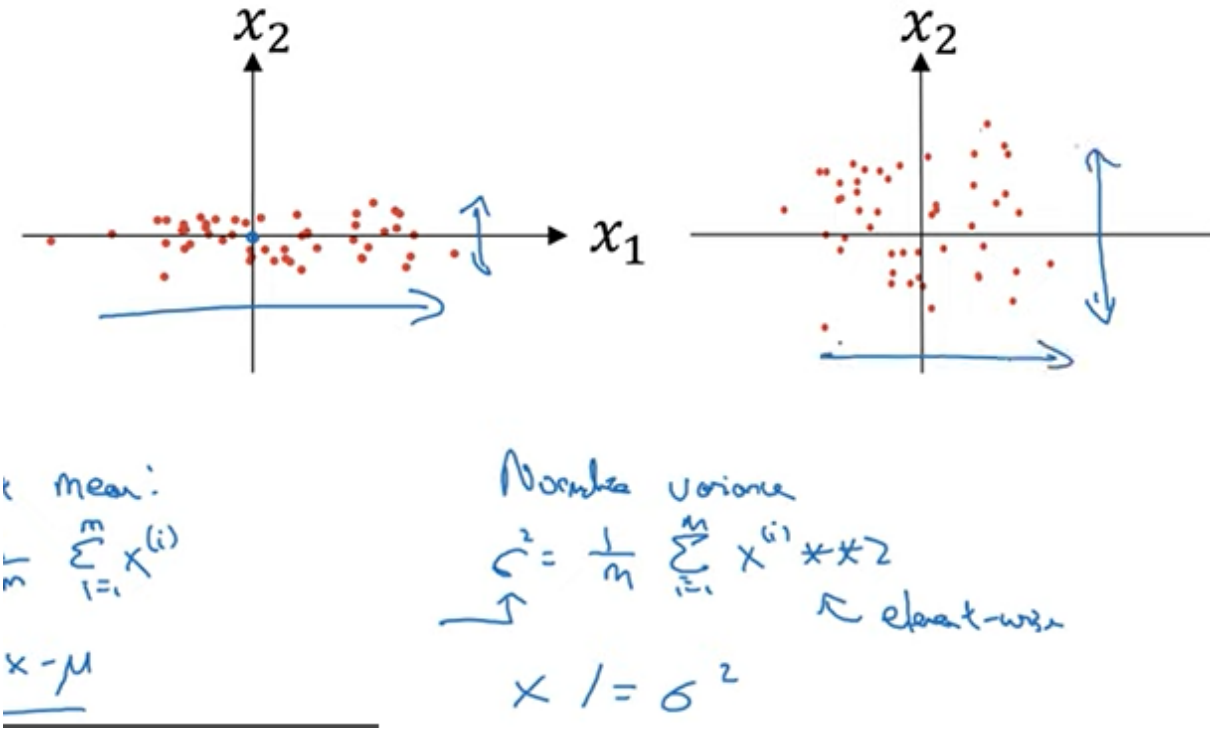
유런 9주차 정리

입력값의 정규화



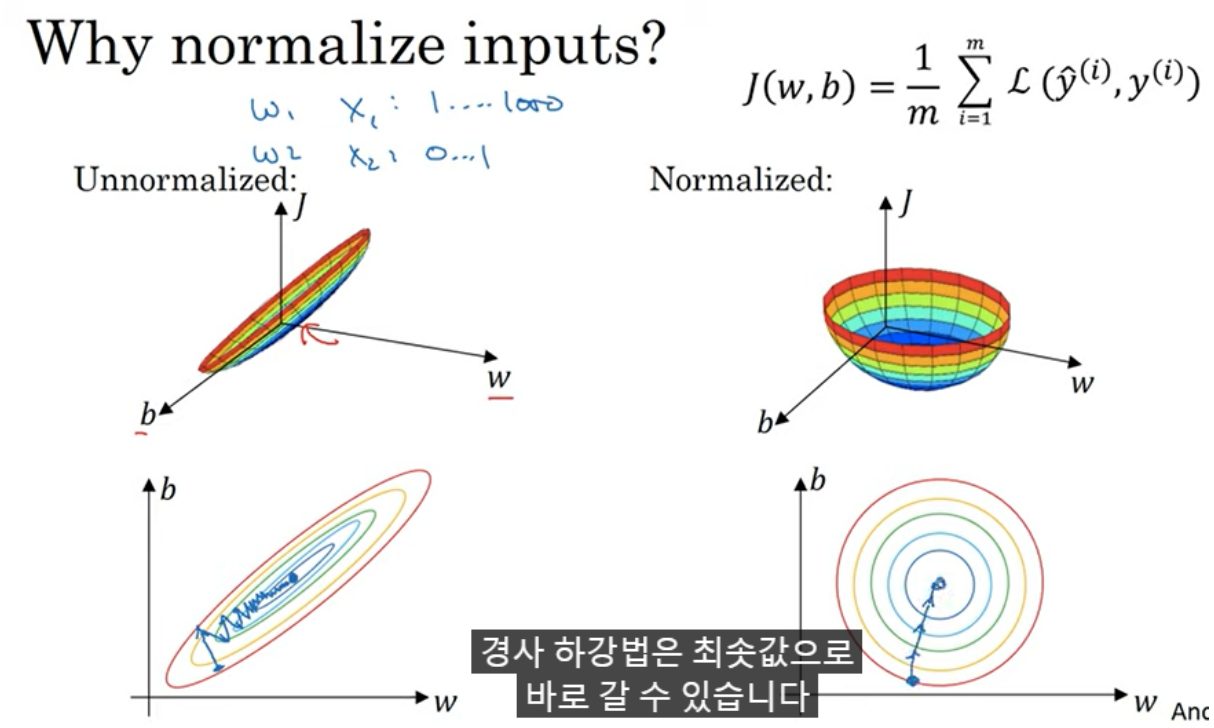
1. 평균을 뺀다.



2. 분산을 정규화한다.
    a. 두번째 그래프의 x1은 x2보다 더 큰 분산을 갖고있다.

테스트 세트를 정규화할 때 훈련 데이터에 사용한 μ 와 σ 를 사용해야한다.

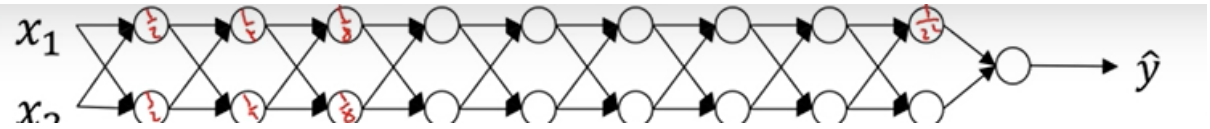
왜 입력 특성을 정규화해야하는가?



- 정규화를 통해 비용함수의 모양은 더 둥글고 최적화하기 쉬운 모습이 된다. 그로 인해 학습 알고리즘이 빨리 실행된다.
- 어떤 것은 0부터 1, 어떤 것은 1부터 1000같이 입력 특성이 다르다면, 그 특성을 정규화하는 것이 중요하다. 이런 정규화는 해를 가하진 않는다.

경사소실, 경사폭발

매우 깊은 신경망을 훈련시킬 때, 미분값 혹은 기울기가 아주 작아지거나 커질 수 있다.



활성화 함수가 선형함수인 경우에, 은닉층 노드 갯수가 2개이고 layer 개수가 많다면 $\hat{y}$ 은 각 layer 의 weight matrix를 다 곱하고 입력벡터를 곱한 값이 될 것이다.

1. weight matrix가 모두 [[1.5, 0],[0, 1.5]] 라면, $\hat{y}$ = $[1.5^L x_1, 1.5^L x_2]$ 이 될 것이다. 단위행렬 보다 큰 값을 계속 곱해줬으므로 y의 예측값은 매우 커질 것이다.

   ⇒ y 의 값은 폭발할 것이다.

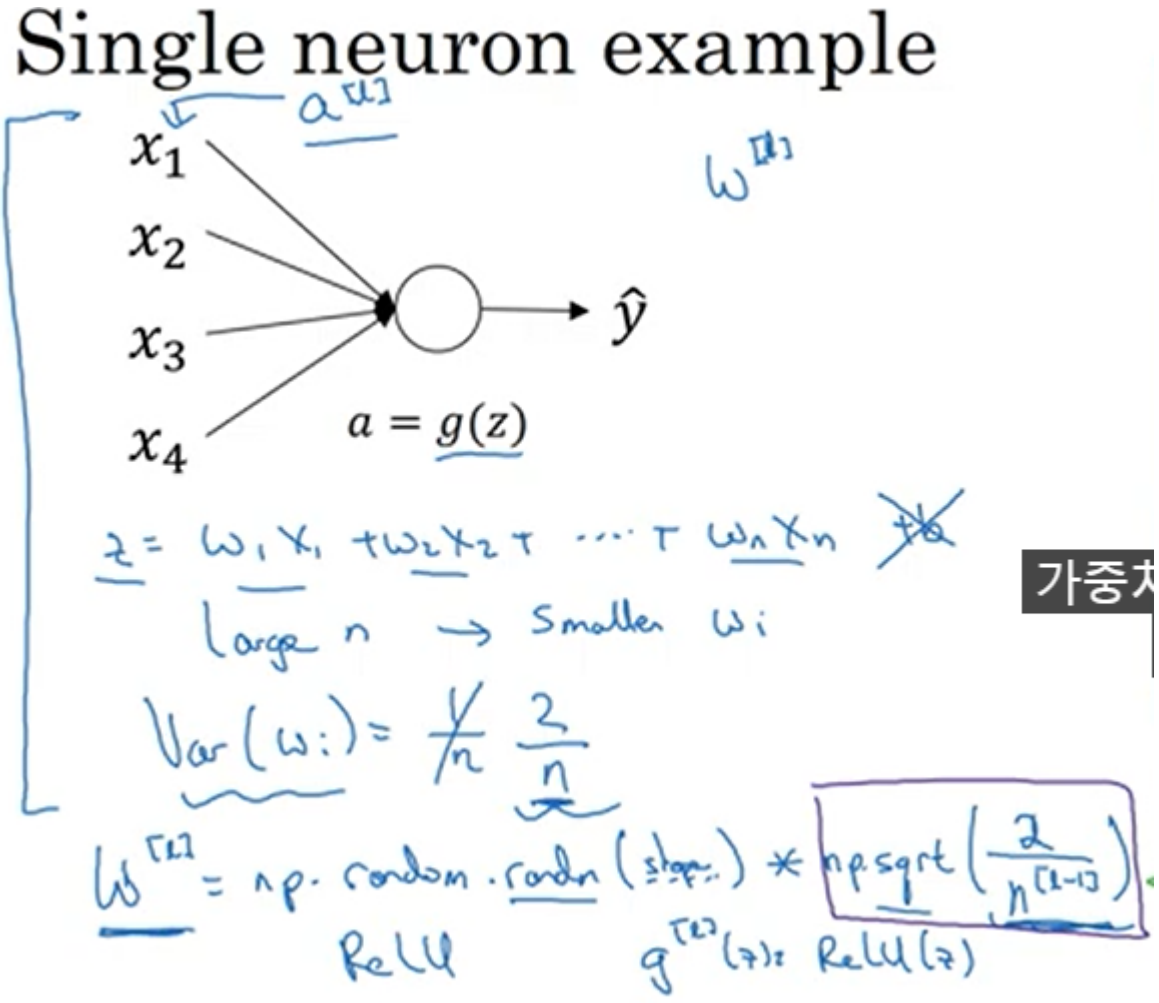
2. weight matrix가 모두 [[0.5, 0],[0, 0.5]] 라면, $\hat{y}$ = $[1.5^L x_1, 0.5^L x_2]$ 이 될 것이다. 단위행렬 보다 작은 값을 계속 곱해줬으므로 y의 예측값(활성값)은 매우 작아질 것이다.

⇒ y 의 값은 소실될 것이다.

현대 신경망은 보통 150개의 신경망을 갖는다.

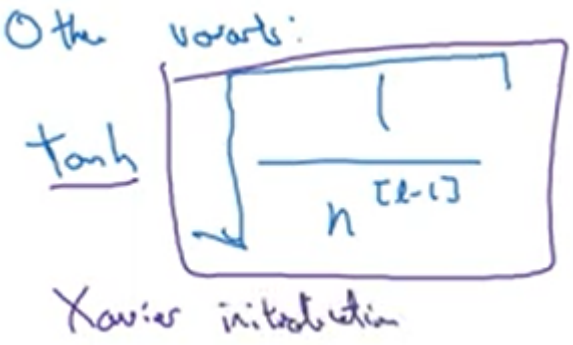
신경망이 깊어질수록 경사가 기하급수적으로 작거나 큰 경우에는 훈련을 시키는 것이 어려워진다. 따라서 가중치 초기화 값을 신중하게 해야한다.

심층 신경망의 가중치 초기화



- 가중치 초기화 방법
  - w_i 의 분산을 $1/n$ 으로 설정한다(n : 입력 특성의 개수)
  - ReLU 활성화 함수를 사용하는 경우 w_i 의 분산을 $2/n^{[l-1]}$ 으로 설정한다.
  - tanh 활성화 함수를 사용하는 경우, w_i 의 분산을 $1/n^{[l-1]}$ 또는 $2/(n^{[l-1]} + n^{[l]})$ 으로 설정한다.

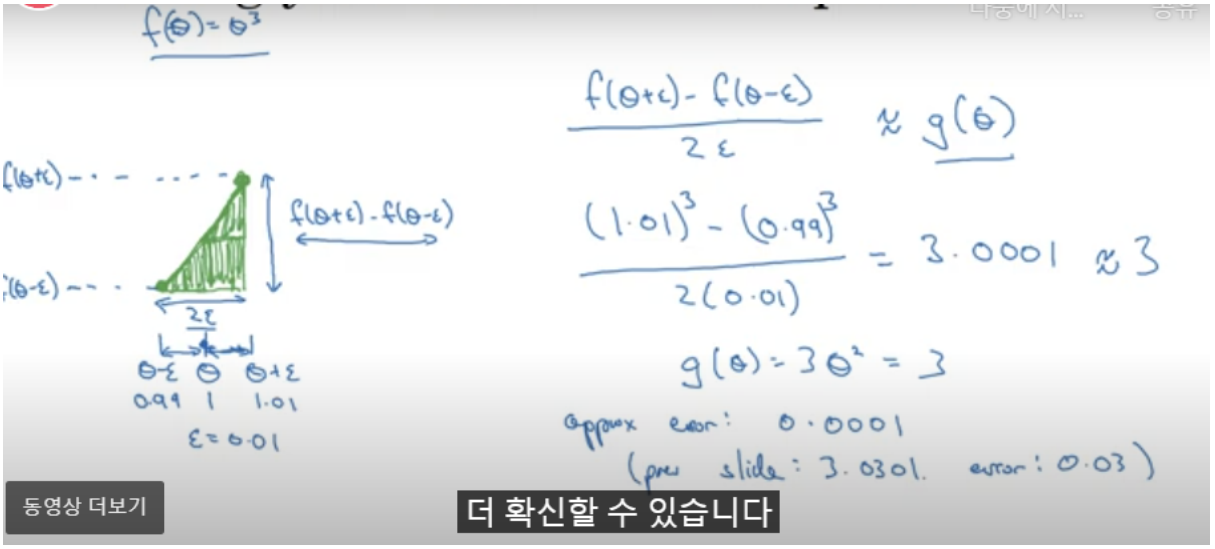
위 수식은 완전히 해결하지는 못하지만 경사 소실과 폭발 문제에 확실히 도움을 줄 수 있다. 가중치 행렬 w를 1보다 너무 커지거나 너무 작아지지 않게 해서 너무 빨리 폭발하거나 소실되지 않게 한다.



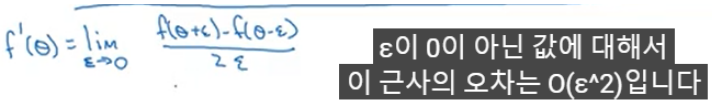
tanh 활성화 함수를 사용한다면 상수 2대신 상수 1을 사용하라는 말도 있다.

기울기의 수치 근사

역전파를 맞게 구현했는지 확인할 수 있는 방법이다.



도함수의 정의



분모가 ϵ 이 된다면 오차는 $O(\epsilon)$ 이므로 오차가 더 크다.

경사 검사

디버그 하는데 도움을 준다.

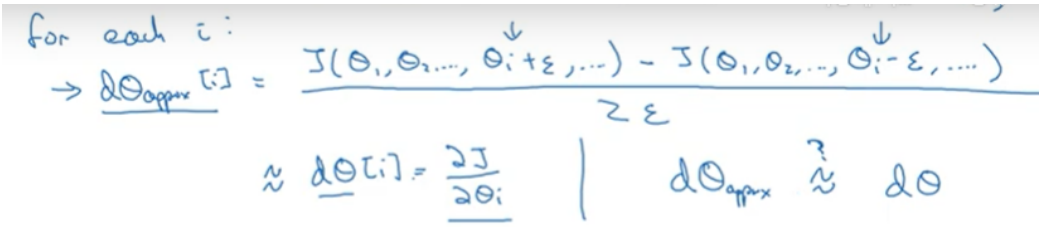
Take $W^{[1]}, b^{[1]}, \dots, W^{[L]}, b^{[L]}$ and reshape into a big vector θ.

1. 경사 검사를 위한 첫 번째는 이 매개 변수들을 하나의 큰 벡터 θ 로 바꾸는 것이다.
2. 그러면 J 함수는 w, b 에 대한 함수 대신에 θ 의 함수가 된다.

Take $dW^{[1]}, db^{[1]}, \dots, dW^{[L]}, db^{[L]}$ and reshape into a big vector $d\theta$.

3. 미분값도 마찬가지로 하나의 벡터로 만든다.

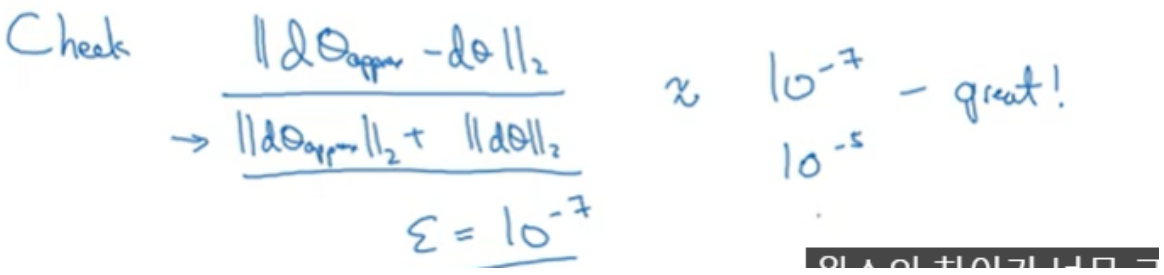
⇒ $d\theta$ 가 비용함수 $J(\theta)$의 기울기랑 같은가?



$J(\theta)$ 를 이용하여 근사 벡터를 만들고 $d\theta$ 와 가까운지 구한다.

두 벡터가 꽤 가까운지 어떻게 정의할 수 있는가?

- 두 벡터의 유클리드 거리를 계산한다(L_2 norm 을 이용한다, 벡터의 길이로 정규화한다.)



10^{-5} 보다 크게 나온다면 벡터의 원소를 살펴보고 너무 큰 원소가 있는지 살펴본다.

보통 거리가 10^{-7} 보다 작으면 잘 계산되었다고 판단한다.

경사 검사 시 주의할 점

- 속도가 굉장히 느리기 때문에 훈련시에는 절대 사용 하지 않고 디버깅 할때만 사용한다. 모든 i 에 대해 계산하는 것이 시간이 많이 소요되기 때문이다.
- 알고리즘이 경사 검사에 실패 했다면, 어느 원소 부분에서 실패했는지 찾아본다. 특정 부분에서 계속 실패했다면, 그 경사가 계산된 층에서 문제가 생긴것을 확인 할 수 있다.
- $d\theta$ 는 θ 에 대응하는 J의 정규화 항도 포함하기 때문에 경사 검사 계산시 같이 포함해야 한다.
- 드롭아웃에서는 무작위로 노드를 삭제하기 때문에 적용하기 쉽지 않다. 따라서 통상은 드롭아웃을 끄고(keep_prop 을 1로 설정) 검사한 다음, 다시 드롭아웃을 켠다.
- 마지막으로 거의 일어나지 않지만 가끔 무작위 초기화를 해도 초기에 경사 검사가 잘 되는 경우가 있다. 이 때는 훈련을 조금 시킨 다음에 경사 검사를 다시 해보는 방법이 있다.