유런 12주차

다중 클래스 분류

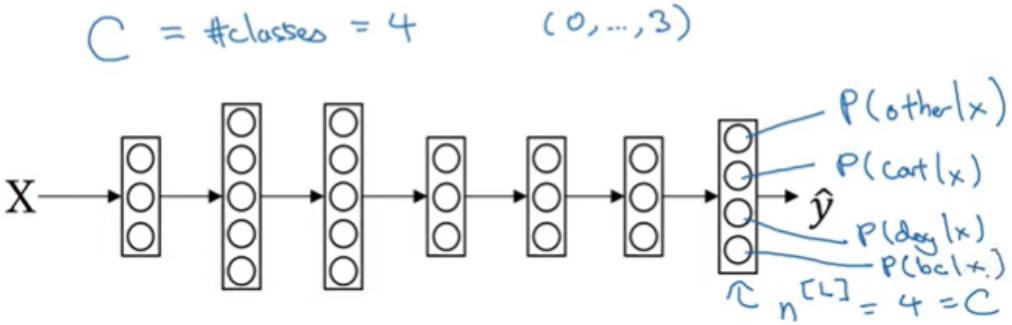
softmax regression

로지스틱 회귀를 일반화한 형태

이미지와 속해 있는 클래스들

C=#classes=4(0~3)

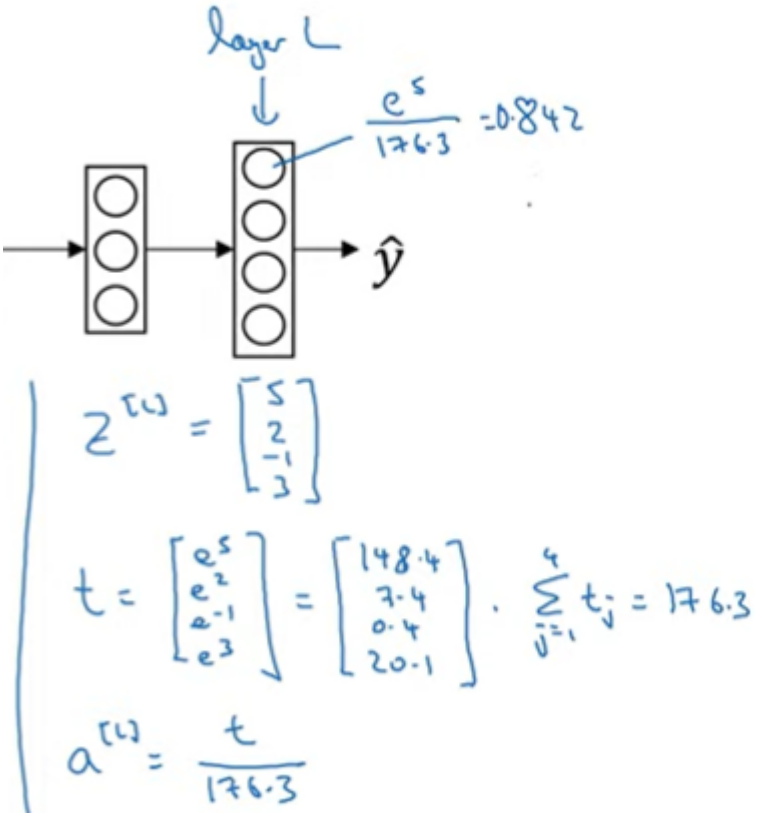
출력층의 노드 갯수는 C가 된다.



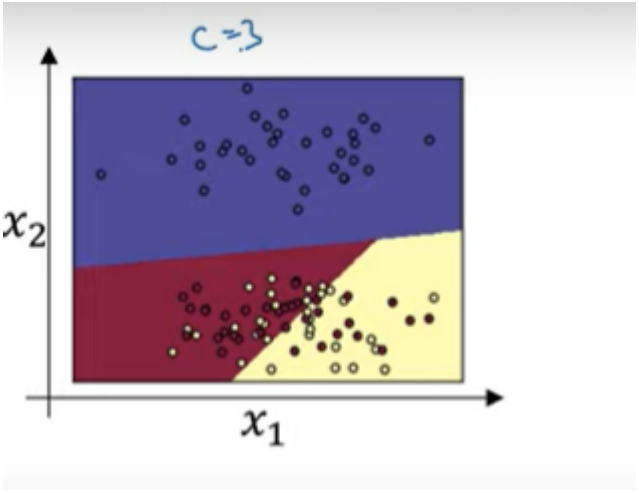
출력층 각각의 노드는 p(노드 담당 클래스|x) 이고 차원은 (C,1) 차원이 된다.

그리고 $\hat{y}$ 의 합은 1이 될 것이다. 이를 위해 소프트맥스 층을 만든다.

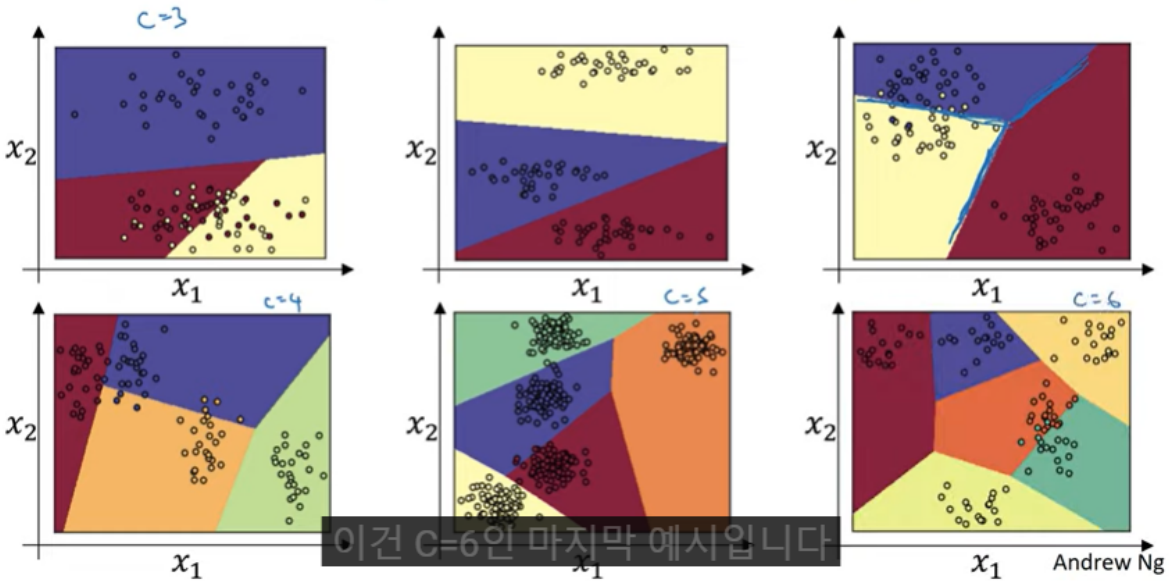
1. $\hat{y}$ 의 모든 원소의 밑에 e 를 둔다.
2. 분모에 1에서 구한 값을 다 더하고
3. 분자에 각 노드에 해당하는 1에서 구한 값을 둔다.



소프트맥스 g 의 특이한 점은 (4,1) 벡터를 받아서 (4,1) 벡터를 통과시키는 것이다. 이전에는 활성 함수가 하나의 실수값을 받은 반면, 소프트 맥스는 벡터를 받는다.



색깔은 소프트맥스 분류 함수에 따라 출력값을 나타낸 것이고 입력값은 가장 높은 확률의 출력값에 따라 색을 입혔다. 선형 기준을 갖고 있는 로지스틱 회귀의 일반적인 형태이다.



Softmax 분류기 훈련시키기

C=4

$$(4,1)$$

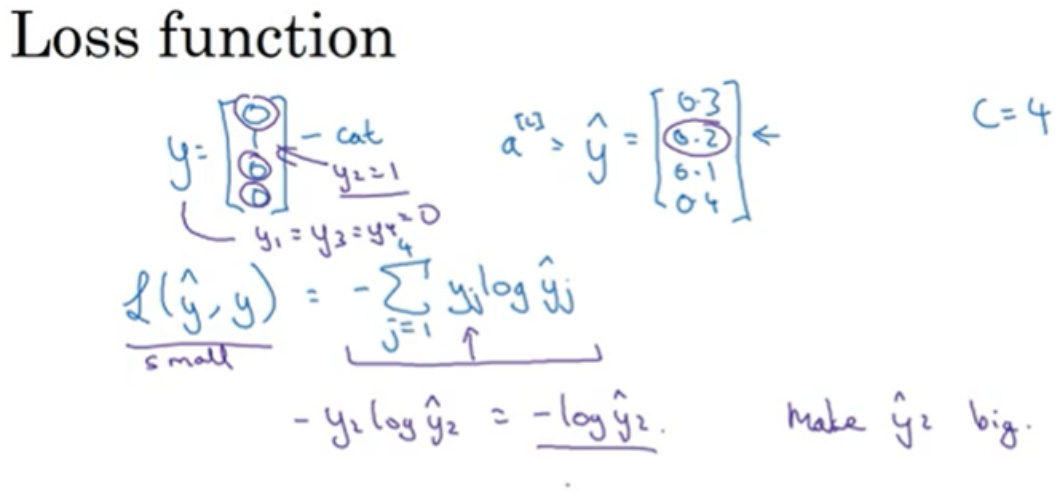
$$z^{[L]} = \begin{bmatrix} 5 \\ 2 \\ -1 \\ 3 \end{bmatrix} \qquad t = \begin{bmatrix} e^5 \\ e^2 \\ e^{-1} \\ e^3 \end{bmatrix}$$

$$a^{[L]} = g^{[L]}(z^{[L]}) = \begin{bmatrix} e^5/(e^5+e^2+e^{-1}+e^3) \\ e^2/(e^5+e^2+e^{-1}+e^3) \\ e^{-1}/(e^5+e^2+e^{-1}+e^3) \\ e^3/(e^5+e^2+e^{-1}+e^3) \end{bmatrix} = \begin{bmatrix} 0.842 \\ 0.042 \\ 0.002 \\ 0.114 \end{bmatrix}$$

hard max : 가장 큰 값에 1을 주고 나머지에 0을 준다

Loss function

정답 클래스의 log likelihood 만 남는다.

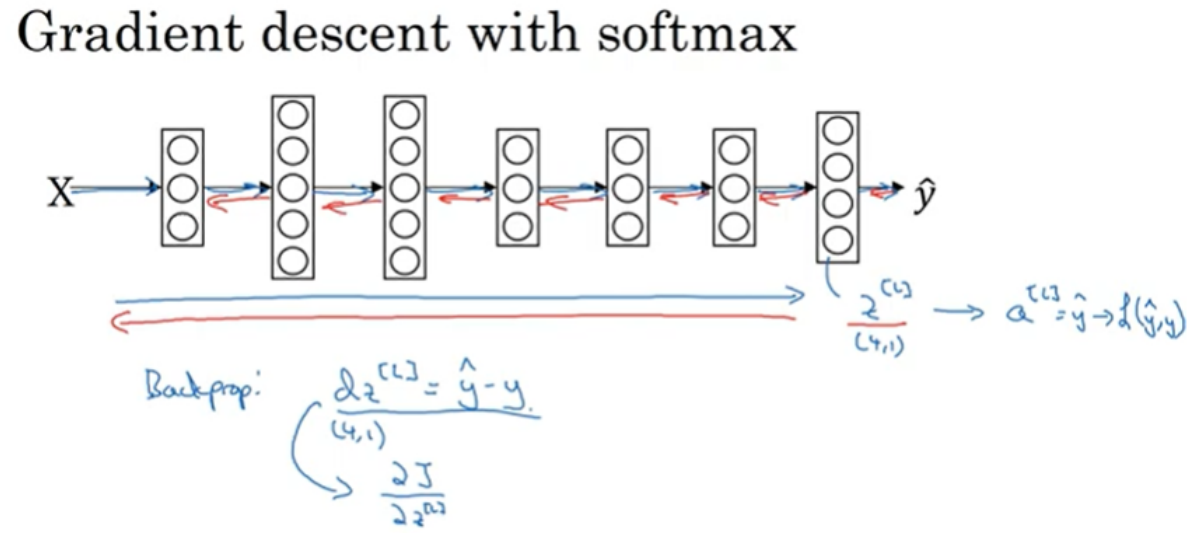


여기서 정답에 해당하는 값인 0.2 를 최대한 많이 키워야한다.

Cost function (모든 데이터셋에 대한 loss 함수의 평균)

$$J(w^{[1]}, b^{[1]}, \ldots) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(\hat{y}^{(i)}, y^{(i)})$$

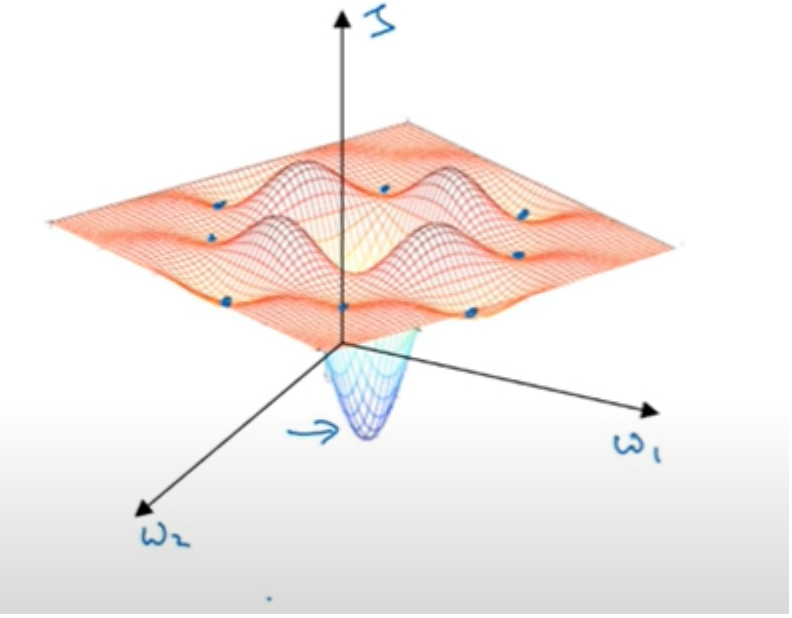
경사하강법을 어떻게 할 것인지?



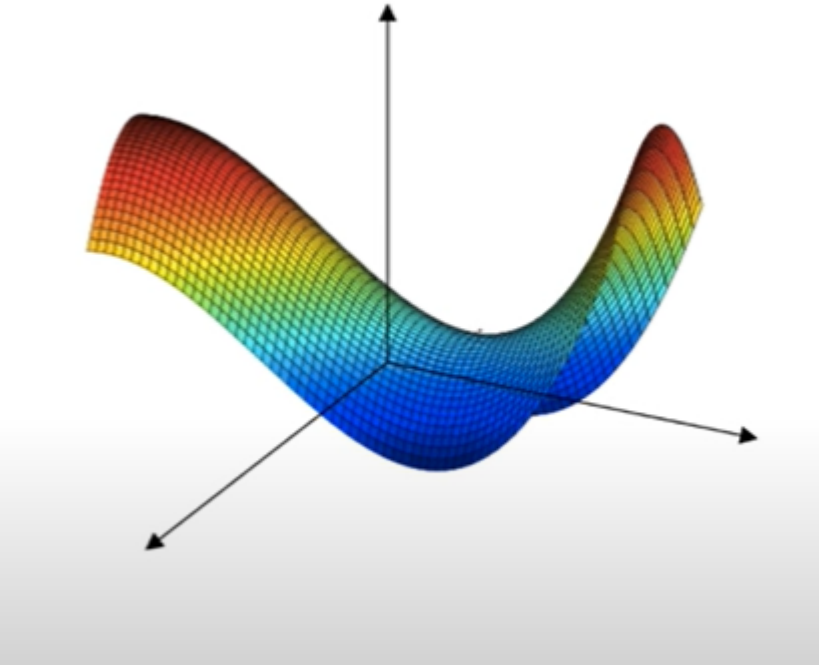
프로그래밍 프레임워크 소개

local optima 문제

global optima 에 가기 전에 local optima 에 갇히는 문제를 말한다.



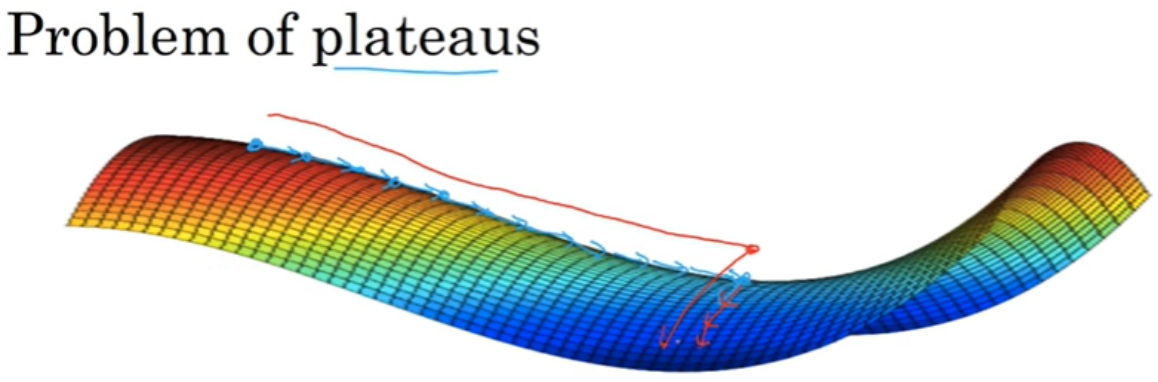
비용함수의 경사가 0인 경우에 위의 그림 경우 보다는 대개 아래와 같이 안장점(saddle point)이다.



왜냐하면 변수가 20000개 일 때, J는 20000 차원의 벡터에 대한 함수일 테고 이러면 local optima 가 아니라 saddle point 가 더 많이 발생한다.

saddle point 가 왜 문제인가?

안정 지대(plateaus) 미분값이 아주 오랫동안 0에 가깝게 유지되는 지역을 말하는데 이 경우, 학습을 매우 지연시킬 수 있다.



빨간 점에 도달하기 위해 매우 오랜 시간이 걸린다. 충분히 큰 신경망을 학습시킨다면 local optima 에 갇힐 일이 잘 없다. 따라서 모멘텀이나 Adam 같은 최적화 알고리즘이 안정지대 내에서 움직이거나 벗어나는 속도를 올릴 수 있다.