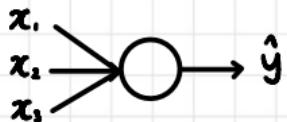


# Batch Normalization

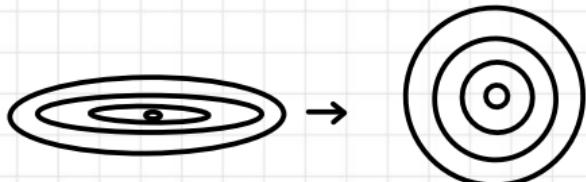


$$\mu = \frac{1}{m} \sum_i x^{(i)}$$

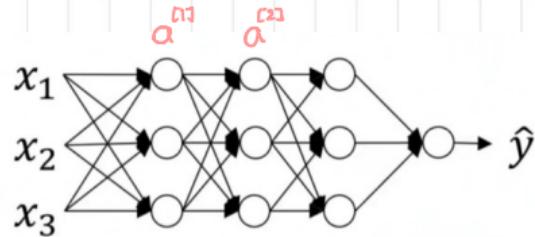
$$x = x - \mu$$

$$\sigma^2 = \frac{1}{m} \sum_i x^{(i)2}$$
element-wise

$$x = x / \sigma$$



→ 로지스틱 회귀에서 입력 변수들을 정규화하면 학습이 빨라짐



$z^{(1)}$ 은 layer 2의 입력값

$z^{(2)}$ 은 layer 3의 입력값

⋮

$z^{(l)}$ 은 layer  $l+1$ 의 입력값

노드의 입력값인  $z$ 를 정규화한다면 ?

"Batch Normalization"

# Batch Normalization

- 하이퍼파라미터 흐름을 쉽게 만들어주며, 신경망과 하이퍼파라미터의 상관관계를 줄여줌
- 보통, 활성화 함수 이전에 사용 ( $Z$ 에 적용)

$$\mu = \frac{1}{m} \sum_i Z^{(i)}$$

평균

$$\sigma^2 = \frac{1}{m} \sum_i (Z^{(i)} - \mu)^2$$

분산

$$Z_{\text{norm}}^{(i)} = \frac{Z^{(i)} - \mu}{\sqrt{\sigma^2 + \epsilon}}$$

$\rightarrow$  정규화  
분산 +  $\epsilon$

$$\tilde{Z}^{(i)} = \gamma Z_{\text{norm}}^{(i)} + \beta$$

learnable parameters

정규화 이후 선형변환을 하는 이유는

항상 같은 분포값을 갖지 않기 위함

## Batch Normalization 효과

① 이전 layer의 가중치 영향을 줄여줌

$Z$ 의 분포를 제한하기 때문

② 드롭아웃과 유사한 효과

mini-batch로 계산한 평균과 분산은 잡음이 존재  $\rightarrow$  정규화 효과

$\rightarrow$  batch size가 커지면 정규화 효과 ↓

# Batch Norm at test time

$$\mu = \frac{1}{m} \sum z^{(i)}$$

평균

$$\sigma^2 = \frac{1}{m} \sum (z^{(i)} - \mu)^2$$

분산

$$z_{\text{norm}}^{(i)} = \frac{z^{(i)} - \mu}{\sqrt{\sigma^2 + \epsilon}}$$

→ 정규화 b 소셜  
(z - z 평균)

$$\tilde{z}^{(i)} = \gamma z_{\text{norm}}^{(i)} + \beta$$

learnable parameter

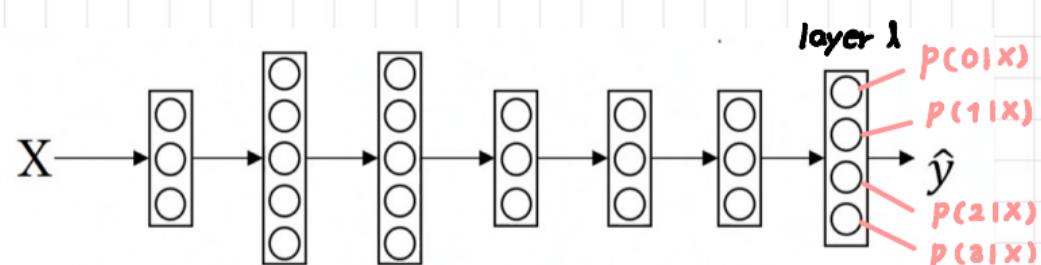
test 과정에서는 batch가 1이기 때문에

평균과 분산을 계산할 수 X

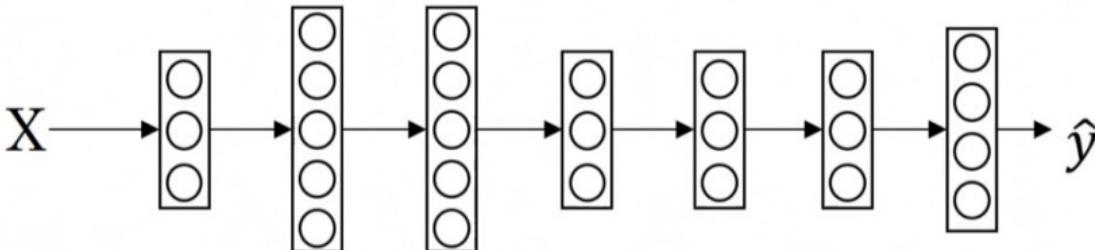
→ train 과정에서 사용된 mini-batch들의  
지수 가중 이동 평균을 추정치로 사용

# Softmax Regression

여러개의 클래스 분류시 사용  $C = \# \text{classes} = 4 \ (0, 1, 2, 3)$



# Softmax layer



$$z^{(l)} = W^{(l)} \alpha^{(l-1)} + b^{(l)} \quad (4.1)$$

Activation function:

$$t = e^{(z^{(l)})}$$

$$\alpha_i^{(l)} = \frac{e^{z_i^{(l)}}}{\sum_{j=1}^4 t_j}, \quad \alpha_i^{(l)} = \frac{t_i}{\sum_{j=1}^4 t_j} \quad (4.1)$$

$$z^{(l)} = \begin{bmatrix} 5 \\ 2 \\ -1 \\ 3 \end{bmatrix} \Rightarrow t = \begin{bmatrix} e^5 \\ e^2 \\ e^{-1} \\ e^3 \end{bmatrix} = \begin{bmatrix} 148.4 \\ 7.4 \\ 0.4 \\ 20.1 \end{bmatrix}$$

$$\sum_{i=1}^4 t_i = 176.3$$

$$\alpha_i^{(l)} = g(z^{(l)}) = \frac{t_i}{176.3} = \begin{bmatrix} 0.842 \\ 0.042 \\ 0.002 \\ 0.114 \end{bmatrix}$$

Class가 0일 확률  
Class가 1일 확률  
Class가 2일 확률  
Class가 3일 확률

전체 합 = 1

# Training Softmax Classifier

$$L(\hat{y}, y) = -\sum_{j=1}^C y_j \log \hat{y}_j$$

$C = 4$

$$y = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \hat{y} = \begin{bmatrix} 0.3 \\ 0.2 \\ 0.1 \\ 0.4 \end{bmatrix}$$

y가 0인 경우는 무시  
→ 1일 때만 고려

$$L(\hat{y}, y) = -y_2 \log \hat{y}_2 = -\log \hat{y}_2$$

make  $\hat{y}_2$  big !