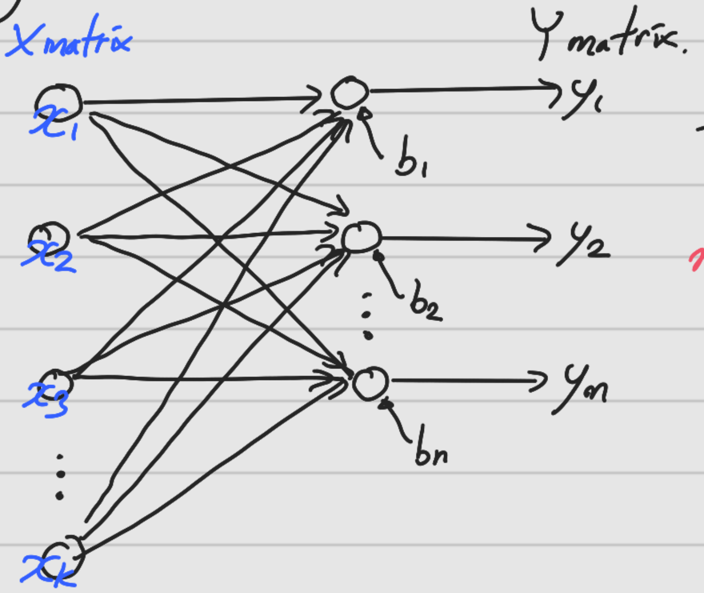
Deep Learning 수학적 증명하기.

1. SLP의 경우

a) Regression의 순전파/역전파 과정의 수학적 증명.



수학적 표현

$$Y = XW + B$$

$m\times n \quad m\times k \quad k\times n$

(순전파)

$Y = XW + B \qquad y_{ij} = x_{i1}w_{1j} + x_{i2}w_{2j} + \cdots + x_{ik}w_{kj} + b_j$

일반적인 수식

$$L = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(\underbrace{label_{ij}}_{\text{given answer}} - \underbrace{y_{ij}}_{\text{predicted value}})^2$$

(역전파)

L is function of $Y_{11}, Y_{12} \cdots Y_{mn}$.

$$L = f(Y_{11}, Y_{12}, \cdots Y_{mn})$$

$$\frac{\partial L}{\partial w_{ij}} = \frac{\partial L}{\partial Y} \cdot \frac{\partial Y}{\partial w_{ij}}$$

w_{ij}가 영향을 주는 모든 Y값들을 고려해야.

$Y_{1j}, Y_{2j} \cdots Y_{mj}$ 총 m개의 Y값.

$$\boxed{= \frac{\partial L}{\partial Y_{1j}}\cdot\frac{\partial Y_{1j}}{\partial w_{ij}} + \frac{\partial L}{\partial Y_{2j}}\cdot\frac{\partial Y_{2j}}{\partial w_{ij}} + \cdots + \frac{\partial L}{\partial Y_{mj}}\cdot\frac{\partial Y_{mj}}{\partial w_{ij}}}$$

$$\frac{\partial L}{\partial Y_{ij}} = \frac{1}{mn} \cdot 2(label_{ij} - Y_{ij}) \cdot (-1)$$

$$\frac{\partial Y_{xj}}{\partial W_{ij}} = X_{xi}$$

$$\frac{\partial L}{\partial W} =$$ this is the matrix to updating weight matrix W $(k \times n)$

$$X^T \cdot \left(\frac{\partial Y}{\partial W}\right) \qquad \left(\frac{\partial L}{\partial Y}\right)$$

$$= \begin{bmatrix} \text{———} \\ \text{———} \\ X_{1i} \; X_{2i} \; X_{3i} \cdots X_{mi} \\ \vdots \\ \text{———} \end{bmatrix} \begin{bmatrix} & diff_{1j} & & \\ & diff_{2j} & & \\ \cdots & diff_{3j} & \cdots & \\ & \vdots & & \\ & diff_{mj} & & \end{bmatrix} \cdot \frac{1}{mn} \cdot (-2)$$

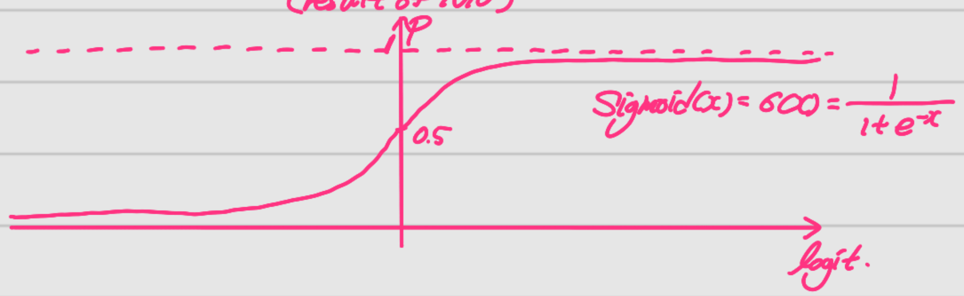
and the component $\frac{\partial Y}{\partial W_{ij}} = \frac{1}{mn} \big[2(label_{1j} - Y_{1j}) \cdot (-1) \cdot X_{1i}$
$+ 2(label_{2j} - Y_{2j}) \cdot (-1) \cdot X_{2i}$

so $\frac{\partial L}{\partial W} = (-2 \times \frac{1}{mn}) \cdot \underbrace{X^T}_{(k \times n)} \cdot \underbrace{(label\ matrix - Y\ matrix)}_{\equiv Diff\ matrix\ (m \times n)}$ $k \times n$ matrix.

b) Binary Classification 수학적 증명

우리는 Binary Classification을 위해 2가지 개념을 새로이 도입한다.

1. Sigmoid function → logit value (result of NN) 을 0과 1사이의 확률로서 재가공하는 비선형 함수



2. Cross-Entropy(교차 Entropy) ; 두 확률분포 P,Q가 주어질 때, 두 확률분포가 유사한 정도를 하나의 숫자로서 표현하기 위한 개념이다.

$$H(P,Q) = \sum_{i=1}^{m} P_i \log \frac{1}{q_i} = \sum_{i=1}^{m} -P_i \log q_i$$

이 교차 Entropy는 두 확률분포가 유사할수록, 작은 값을 지닌다.

Binary Classification에서 P의 의미;

모범답안으로 주어지는 label의 정보는 입력에 따른 조건부 확률로서 P를 재해석하고, 이를 이용하여 cross entropy를 계산하는 방식을 통해, 문제를 해결한다.

ex) P; label data. Q; predicted probability.

at given input data (x_i)

$\langle x_{i1}\ x_{i2} \cdots x_{ik} \rangle$ this is a condition

and q_i is predicted probability when x_i's condition is given.

label data P_i는 0 혹은 1로서 주어진다.

이 P_i는 조건부 확률 (x_i)에 대한 정답이다.

H(P,Q)를 줄이는 방식으로 학습을 진행한다면, 숨어있는 pattern이

≡ 다양한 입출력 쌍에 대한 교차 Entropy를 줄이는 방향으로 학습을 진행한다면,

신경망에 반영되어 각각의 입력에 알맞는 조건부 확률분포를

≡ 각각의 입력에 맞는 알맞는 결과를 주로.

적절히 만들어낼수 있을 것이다.