< Part 2 > 데이터 사이언스 필수 수학

[Ch 01.] 딥러닝 네트워크의 연산

[CH0 .] 01, 02, 03

Lecture. 4

Loss functions

<Lecture. 4 Loss Functions>

CH04_01.[이론강의]_Mean_Squared_Error.

<Cartesian Product for Predictions / Labels>

$\hat{y}$ → 실수이면 대부분 Regression 이라고 봄.

$\hat{y}$ → 0 → 강아지
1 → 고양이
2 → ··
⋮

⇒ discrete → classification

← class끼리 대소관계가 없음.

$\hat{y}$
↑
Model
↑
$(\vec{x})^T$

- 목표 $\hat{y}$(예측) ⟷ y(실제값)

$\hat{y}$
↑
Model
↑
$(\vec{x})^T$

Dataset

공부시간	성적
5	100
4	8
x값	y값

→ 비슷한값 $\hat{y}$ 만들기 : Model

⇒ Loss 필요성.

- **Regression**
  ↓
  $\hat{y} \in \mathbb{R}$
  ↓
  MSE

- **Classification**
  ↓
  $\hat{y} \in C$(class index)
  $\{0, 1, 2, \cdots, k\}$
  - Binomial → Binary Cross Entropy
  - Multi → Categorical Cross Entropy

- 집합

R 실수집합

$B = \{0, 1\}$ Binary 집합

$C = \{c_1, c_2, \cdots, c_k\}$ 멀티클래스집합

$P = \{x \mid 0 \le x \le 1\}$ 확률집합

$B_i^n = \{(b_1, b_2, \cdots, b_n)^T \mid \forall b \in B, \sum_{i=1}^{n} b_i = 1\}$: on-hot vector 집합.

강아지 $= (0, 0, 1)$ 고양이 $= (0, 1, 0)$ 호랑이 $= (1, 0, 0)$

<Mean Squared Error>

- Dataset for Regression

1번학생: $(\vec{x}^{(1)}) = (x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, \dots, x_{l_I}^{(1)}) \quad y^{(1)} \in \mathbb{R}$

공부시간 거주지 학년

$\vdots$

N번학생 $(\vec{x}^{(N)}) = (x_1^{(N)}, x_2^{(N)}, x_3^{(N)}, \dots, x_{l_I}^{(N)}) \quad y^{(N)} \in \mathbb{R}$

$$\longrightarrow \quad X^T = \begin{pmatrix} \leftarrow \vec{x}^{(1)} \rightarrow \\ \leftarrow \vec{x}^{(2)} \rightarrow \\ \vdots \\ \leftarrow \vec{x}^{(N)} \rightarrow \end{pmatrix} \in \mathbb{R}^{N \times l_I}$$

$$Y = \begin{pmatrix} y^{(1)} \\ \vdots \\ y^{(N)} \end{pmatrix} \in \mathbb{R}^{N \times 1}$$

- 단일데이터의 case

Squared Error

$\uparrow J \in \mathbb{R}$

$J = (y - \hat{y})^2 \quad y \in \mathbb{R}$

$\uparrow \hat{y} \in \mathbb{R}$

$\cancel{L^{[0]} \ \vdots \ L^{[2]} \ L^{[1]}}$

$\uparrow$

$(\vec{x})^T$

$J = (y - \hat{y})^2$

no activation

affine $\quad z = (\vec{a}^{[0-1]})^T \vec{w}^{[0]} + b^{[0]}$

$\uparrow$

$\mathbb{R}$

- 여러데이터 → Matrix

Squared Error

$\uparrow J \in \mathbb{R}$

$J = \frac{1}{N}\sum_{i=1}^{N} (y^{(i)} - \hat{y}^{(i)})^2 \quad Y \in \mathbb{R}^{N \times 1}$

$\uparrow \hat{Y} \in \mathbb{R}^{N \times 1}$

$\cancel{L^{[0]} \ \vdots \ L^{[2]} \ L^{[1]}}$

$\uparrow$

X^T

$$J = \frac{1}{N}\sum_{i=1}^{N} (y^{(i)} - \hat{y}^{(i)})^2 : \text{MSE}$$

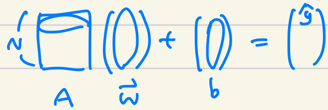
$$Y = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{pmatrix} \in \mathbb{R}^{N \times 1}$$

$$\hat{Y} = \begin{pmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \vdots \\ \hat{y}^{(N)} \end{pmatrix} \in \mathbb{R}^{N \times 1}$$

affine $\quad z = (A^{[0-1]})^T \cdot \vec{w}^{[0]} + b^{[0]}$

$\downarrow$

minibatch가 들어와도 그대로 한다.



CH 04_02. [이론강의]_Binary_Cross_Entropy.

- Dataset for Binary Classification

$(\vec{x}^{(1)})^T = (x_1^{(1)} \quad x_2^{(1)} \quad x_3^{(1)} \dots x_{\ell_I}^{(1)}) \quad y^{(1)} \in \mathbb{B}$ 1 합격

공부시간 사교육비 몸무게 · · ·

$(\vec{x}^{(2)})^T = (x_1^{(2)} \quad x_2^{(2)} \quad x_3^{(2)} \dots x_{\ell_I}^{(2)}) \quad y^{(2)} \in \mathbb{B}$ 0 불합격

$(\vec{x}^{(N)})^T = (x_1^{(N)} \quad x_2^{(N)} \quad x_3^{(N)} \dots x_{\ell_I}^{(N)}) \quad y^{(N)} \in \mathbb{B}$ 1 합격

$$X^T = \begin{pmatrix} \leftarrow (\vec{x}^{(1)})^T \rightarrow \\ \leftarrow (\vec{x}^{(2)})^T \rightarrow \\ \vdots \\ \leftarrow (\vec{x}^{(N)})^T \rightarrow \end{pmatrix} \in \mathbb{R}^{N \times \ell_I}$$

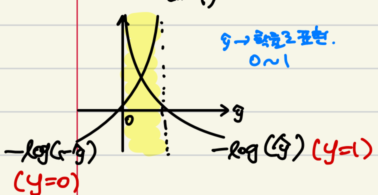
$$Y = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{pmatrix} \in \mathbb{B}^{N \times 1}$$

$$\Rightarrow \underset{\hat{y}}{\mathbb{R}} \sim \underset{y}{\mathbb{B}}$$

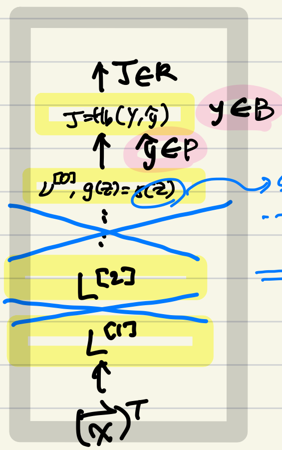
<Binary Cross Entropy>

$$\begin{cases} y=1 \rightarrow \mathcal{L} = -\log(\hat{y}) \\ y=0 \rightarrow \mathcal{L} = -\log(1-\hat{y}) \end{cases}$$

$\Rightarrow$ 합치기 : $\underset{\text{Entropy}}{H}_{\overset{\text{binary}}{b}}(y, \hat{y}) = -[y\log(\hat{y}) + (1-y)\log(1-\hat{y})]$

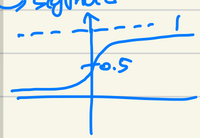


$\hat{y}$ → 확률로 표현. 0~1

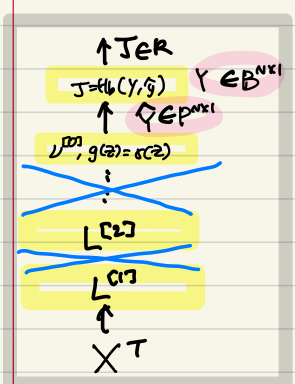


$\nu^{[0]} = \frac{\text{sigmoid}}{\text{affine}}$

sigmoid



- Minibatch



$$Y = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{pmatrix} \in \mathbb{B}^{N \times 1}$$

$$\hat{Y} = \begin{pmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \vdots \\ \hat{y}^{(N)} \end{pmatrix} \in \mathbb{P}^{N \times 1}$$

$$J = H_b(Y, \hat{Y})$$

$$= \frac{1}{N} \sum_{i=1}^{N} [y^{(i)} \log(\hat{y}^{(i)}) + (1-y^{(i)}) \log(1-\hat{y}^{(i)})]$$

CH04_03. [이론정리]_Categorical_Cross_Entropy

- Dataset for Multi-class Classification

$\mathbb{C} = \{ C_1, C_2, \cdots, C_K \}$ 강아지 고양이

$(\vec{x}^{(1)}) = (x_1^{(1)}\ x_2^{(1)} \cdots x_{l_I}^{(1)})$ $y^{(1)} \in \mathbb{C}$

$(\vec{x}^{(2)}) = (x_1^{(2)}\ x_2^{(2)} \cdots x_{l_I}^{(2)})$ $y^{(2)} \in \mathbb{C}$

$\vdots$

$(\vec{x}^{(N)}) = (x_1^{(N)}\ x_2^{(N)} \cdots x_{l_I}^{(N)})$ $y^{(N)} \in \mathbb{C}$

$$\longrightarrow \quad X^T = \begin{pmatrix} \leftarrow \vec{x}^{(1)} \rightarrow \\ \leftarrow \vec{x}^{(2)} \rightarrow \\ \vdots \\ \leftarrow \vec{x}^{(N)} \rightarrow \end{pmatrix} \in \mathbb{R}^{N \times l_I}$$

$$Y = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{pmatrix} \in \mathbb{C}^{N \times 1}$$

- One-hot Encoding

$(\vec{y}^{(2)})^T \in \mathbb{B}_1^{1 \times K}$ $\longleftarrow$

여기서는 one-hot encoding 이라는 의미로 사용

$(\vec{y}^{(1)})^T = (1\ 0 \ldots 0),\ \alpha = 1$ 강아지

$(\vec{y}^{(2)})^T = (0\ 1 \ldots 0),\ \alpha = 2$ 고양이

$\vdots$

$(\vec{y}^{(K)})^T = (0\ 0 \ldots 1),\ \alpha = K$ 토끼

$(\vec{y}^{(1)})^T \in \mathbb{B}_1^{1 \times K}$

$(\vec{y}^{(2)})^T \in \mathbb{B}_1^{1 \times K}$

$\vdots$

$(\vec{y}^{(N)})^T \in \mathbb{B}_1^{1 \times K}$

$$\longrightarrow \quad Y^T = \begin{pmatrix} \leftarrow (\vec{y}^{(1)})^T \rightarrow \\ \leftarrow (\vec{y}^{(2)})^T \rightarrow \\ \vdots \\ \leftarrow (\vec{y}^{(N)})^T \rightarrow \end{pmatrix} \in \mathbb{B}_1^{N \times K}$$

- 계산과정

$$H(\vec{y}, \hat{\vec{y}}) = -\sum_{i=1}^{K} y_i \log(\hat{y}_i)$$

$\uparrow J$

$$\boxed{J = H(\vec{y}, \hat{\vec{y}})}$$

$(\vec{y})^T = (y_1\ y_2 \cdots y_K)$

$y_i \in \{0, 1\}$

근가

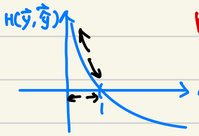
output layer에서 softmax를 거쳐서 0~1 사이의 값으로 변함.

$(\hat{\vec{y}})^T = (P(C_1), P(C_2), \cdots, P(C_K))$

$= (\hat{y}_1\ \hat{y}_2 \cdots \hat{y}_K)$

$0 \le \hat{y}_i \le 1$

ex) $\vec{y} = (1\ 0\ 0\ 0)$ $\quad \hat{\vec{y}} = (0.8\ 0.1\ 0.1\ 0)$



$$H(\vec{y}, \hat{\vec{y}}) = -\log(\hat{y}_1)$$

$$= -\log(P(C_1))$$

∴ $P(C_1)$가 0에 가까워질수록 loss값이 엄청 커짐.

↔ 1에 가까워질수록 0에 가까워짐. → 목표!

- Model : 하나의 input vector 경우

Cross Entropy

$\uparrow J \in \mathbb{R}$

$J = H(\vec{y}, \hat{\vec{y}})$ $\quad (\vec{y})^T \in B_1^{1\times K}$, $(\hat{\vec{y}})^T \in P^{1\times K}$

Softmax

$L^{[O]}$ → No activation

$\vdots$

$L^{(2)}$

$L^{(1)}$

$(\vec{x})^T$

Model : minibatch의 경우

Cross Entropy

$J \in \mathbb{R}$

$J = H(Y^T, \hat{Y}^T)$ $\quad Y^T \in B_1^{N\times K}$, $\hat{Y}^T \in P^{N\times K}$

softmax

$L^{(O)}$

$\vdots$

$L^{(2)}$

$L^{(1)}$

$X^T \in \mathbb{R}^{N\times \ell_I}$

$$H(Y, \hat{Y}) = \frac{1}{N}\sum_{i=1}^{N} H(\vec{y}^{(i)}, \hat{\vec{y}}^{(i)})$$

$$= -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} y_j^{(i)} \log(\hat{y}_j^{(i)})$$

↑ 전체평균 .