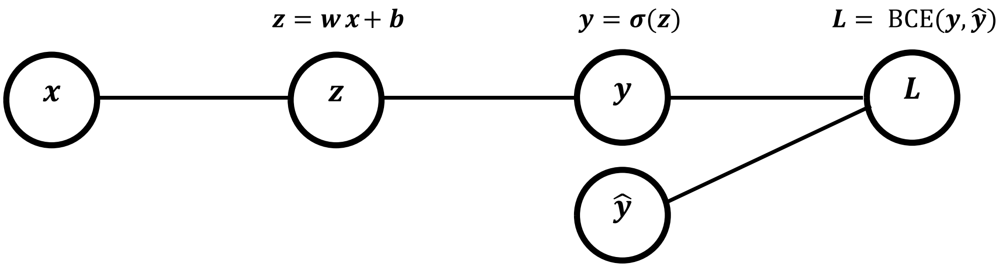
Backpropagation 실습 - Gradient 계산 과정 보충 설명



Model 구조: 단일 linear layer non-linear activation function으로 sigmoid 함수 채택

$\boldsymbol{x}$: 모델의 입력 값 (x_i)

$\boldsymbol{w}, \boldsymbol{b}$: Linear layer의 weight와 bias (둘 다 scalar)

$\boldsymbol{\sigma}$: Non-linear activation function인 sigmoid 함수 $\left(\boldsymbol{\sigma}(\boldsymbol{x}) = \frac{1}{1+e^{-x}}\right)$

$\boldsymbol{y}$: 모델의 예측 값 (y_pred), $\hat{\boldsymbol{y}}$: 실제 값 (y_i)

$\boldsymbol{L}$: loss function (binary cross entropy (BCE), $\boldsymbol{L} = \mathrm{BCE}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = -\sum\{\hat{\boldsymbol{y}} \log \boldsymbol{y} + (1-\hat{\boldsymbol{y}}) \log(1-\boldsymbol{y})\}$

Chain rule을 활용하여 모델의 optimization (gradient descent)에 필요한 gradient의 값들을 구하자.

$$\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{f_1}} = \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{y}} = \frac{\partial}{\partial \boldsymbol{y}}\left[-\sum\{\hat{\boldsymbol{y}} \log \boldsymbol{y} + (1-\hat{\boldsymbol{y}}) \log(1-\boldsymbol{y})\}\right] \quad \text{(단일 } \boldsymbol{y} \text{ 값이므로 } \Sigma \text{ 기호 무시 가능)}$$

$$= -\left\{\frac{\hat{\boldsymbol{y}}}{\boldsymbol{y}} \times \frac{\partial}{\partial \boldsymbol{y}}(\boldsymbol{y}) + \frac{(1-\hat{\boldsymbol{y}})}{(1-\boldsymbol{y})} \times \frac{\partial}{\partial \boldsymbol{y}}(1-\boldsymbol{y})\right\} = -\left\{\frac{\hat{\boldsymbol{y}}}{\boldsymbol{y}} - \frac{(1-\hat{\boldsymbol{y}})}{(1-\boldsymbol{y})}\right\}$$

$$= -\left\{\frac{\hat{\boldsymbol{y}}(1-\boldsymbol{y}) - (1-\hat{\boldsymbol{y}})\boldsymbol{y}}{\boldsymbol{y}(1-\boldsymbol{y})}\right\} = -\left\{\frac{\hat{\boldsymbol{y}} - \hat{\boldsymbol{y}}\boldsymbol{y} - \boldsymbol{y} + \boldsymbol{y}\hat{\boldsymbol{y}}}{\boldsymbol{y}(1-\boldsymbol{y})}\right\} = -\left\{\frac{\hat{\boldsymbol{y}} - \boldsymbol{y}}{\boldsymbol{y}(1-\boldsymbol{y})}\right\} = \frac{\boldsymbol{y} - \hat{\boldsymbol{y}}}{\boldsymbol{y}(1-\boldsymbol{y})}$$

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{z}} = \frac{\partial}{\partial \boldsymbol{z}}\boldsymbol{\sigma}(\boldsymbol{z}) = \frac{\partial}{\partial \boldsymbol{z}}\left(\frac{1}{1+e^{-z}}\right) = -\frac{1}{(1+e^{-z})^2} \times \frac{\partial}{\partial \boldsymbol{z}}(1+e^{-z})$$

$$= -\frac{1}{(1+e^{-z})^2} \times (-e^{-z}) = \frac{e^{-z}}{(1+e^{-z})^2} = \frac{1}{1+e^{-z}} \times \frac{e^{-z}}{1+e^{-z}} = \boldsymbol{\sigma}(\boldsymbol{z})\{1-\boldsymbol{\sigma}(\boldsymbol{z})\}$$

$$\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{f_2}} = \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{z}} = \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{y}} \times \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{z}} = \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{f_1}}\,\boldsymbol{\sigma}(\boldsymbol{z})\{1-\boldsymbol{\sigma}(\boldsymbol{z})\} = \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{f_1}} \times \boldsymbol{y}(1-\boldsymbol{y}) \quad \left(= \frac{\boldsymbol{y}-\hat{\boldsymbol{y}}}{\boldsymbol{y}(1-\boldsymbol{y})} \times \boldsymbol{y}(1-\boldsymbol{y}) = \boldsymbol{y} - \hat{\boldsymbol{y}}\right)$$

$$\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{w}} = \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{z}} \times \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{w}} = \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{z}} \times \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{w}}(\boldsymbol{w}\boldsymbol{x} + \boldsymbol{b}) = \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{f_2}} \times \boldsymbol{x} \quad (= (\boldsymbol{y} - \hat{\boldsymbol{y}})\boldsymbol{x})$$

$$\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{b}} = \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{z}} \times \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{b}} = \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{z}} \times \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{b}}(\boldsymbol{w}\boldsymbol{x} + \boldsymbol{b}) = \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{f_2}} \times 1 = \frac{\partial \boldsymbol{L}}{\partial \boldsymbol{f_2}} \quad (= \boldsymbol{y} - \hat{\boldsymbol{y}})$$