MLP (Multi-Layer Perceptron)

- Neural Networks
  - Computing systems vaguely inspired by the biological neural networks that constitute animal brains
  - **Function approximators that stack affine transformations followed by nonlinear transformations.**
- Linear Neural Networks
  - Data: $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$
  - Model: $\hat{u} = wx + b$
  - Loss: $\text{loss} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2$

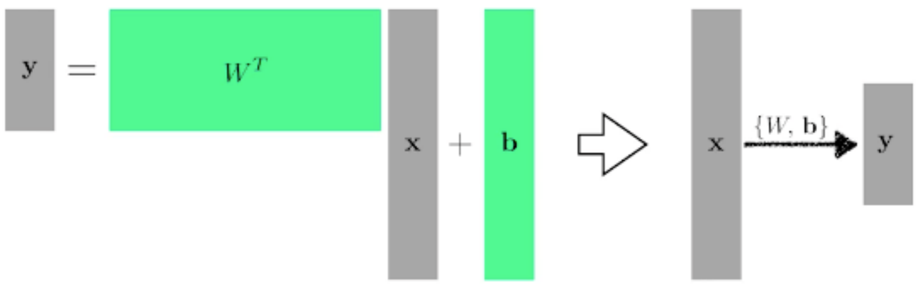
$$\begin{aligned}\frac{\partial\,\text{loss}}{\partial w} &= \frac{\partial}{\partial w}\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2 \\ &= \frac{\partial}{\partial w}\frac{1}{N}\sum_{i=1}^{N}\left(y_i - wx_i - b\right)^2 \\ &= \frac{1}{N}\sum_{i=1}^{N} -2\left(y_i - wx_i - b\right)x_i\end{aligned}$$

$$\begin{aligned}\frac{\partial\,\text{loss}}{\partial b} &= \frac{\partial}{\partial b}\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2 \\ &= \frac{\partial}{\partial b}\frac{1}{N}\sum_{i=1}^{N}\left(y_i - wx_i - b\right)^2 \\ &= \frac{1}{N}\sum_{i=1}^{N} -2\left(y_i - wx_i - b\right)\end{aligned}$$

$w \leftarrow w - \eta\frac{\partial\,\text{loss}}{\partial w}$

$b \leftarrow b - \eta\frac{\partial loss}{\partial b}$

  - η (Stepsize) 적절하게 잡는 것 중요



$y = W_3^T h_2 = W_3^T \rho\left(W_2^T X h_1\right) = W_3^T \rho\left(W_2^T \rho\left(W_1^T X\right)\right)$

  - Non-linear transform (ρ), activation functions(활성함수)이 들어야 네트워크를 깊게 쌓았을 때 의미 있음
- Activation functions
  - ReLU (Rectified Linear Unit)
  - sigmoid
  - tanh (Hyperbolic Tangent)

- Universal approximation theory
    - 히든 레이어가 1개 있는 신경망의 표현력은 일반적인 continuous function들을 포함함
    - 존재성에 대한 증명 (어떻게 찾는지까지에 대한 이론은 아님)
- Loss functions
    - Regression Task

      $\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\sum_{d=1}^{D}\left(y_i^{(d)} - \hat{y}_i^{(d)}\right)^2$
    - Classification Task

      $\mathrm{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{d=1}^{D} y_i^{(d)} \log \hat{y}_i^{(d)}$
    - Probabilistic Task

      $\mathrm{MLE} = \frac{1}{N}\sum_{i=1}^{N}\sum_{d=1}^{D} \log \mathcal{N}\left(y_i^{(d)}; \hat{y}_i^{(d)}, 1\right) \quad (= \mathrm{MSE})$