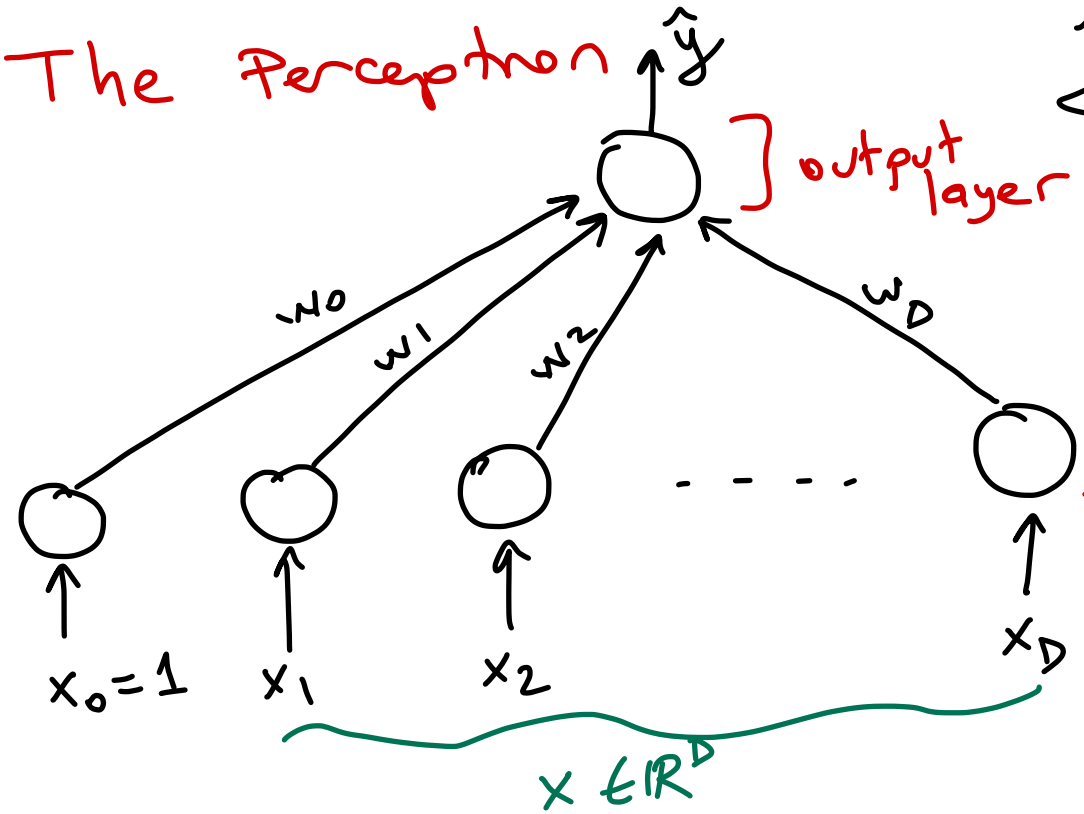
Multilayer Perceptrons

The Perceptron



$$\hat{y} = w_0 \cdot x_0 + w_1 x_1 + w_2 x_2 + \dots + w_D x_D$$

$$= w_0 \cdot \underbrace{x_0}_{1} + \sum_{d=1}^{D} w_d \cdot x_d$$

$$= w^T \cdot x + w_0$$

$w^T x$

$$\hat{y} = [w_1 \; w_2 \; \dots \; w_D] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} + w_0$$

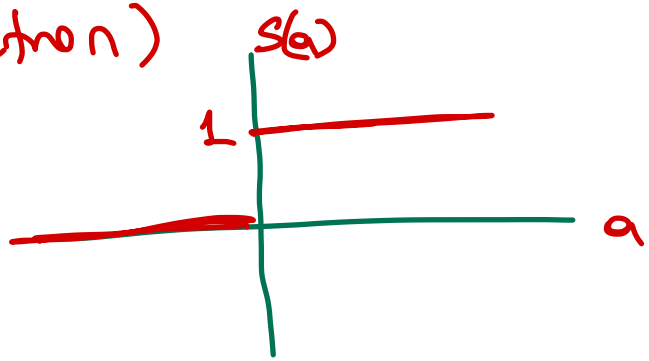
$$= [w_0 \; w_1 \; w_2 \; \dots \; w_D] \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} = w^T \cdot x$$

$1 \times (D+1)$ $\qquad$ $(D+1) \times 1$

threshold function (activation function)

$$s(a) = \begin{cases} 1 & \text{if } a > 0 \\ 0 & \text{otherwise} \end{cases}$$

↑ received message



$$s(w^T \cdot x) = \begin{cases} 1 & \text{if } w^T \cdot x > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$s(w^T \cdot x) = \frac{1}{1 + \exp[-w^T \cdot x]}$$

↳ sigmoid activation

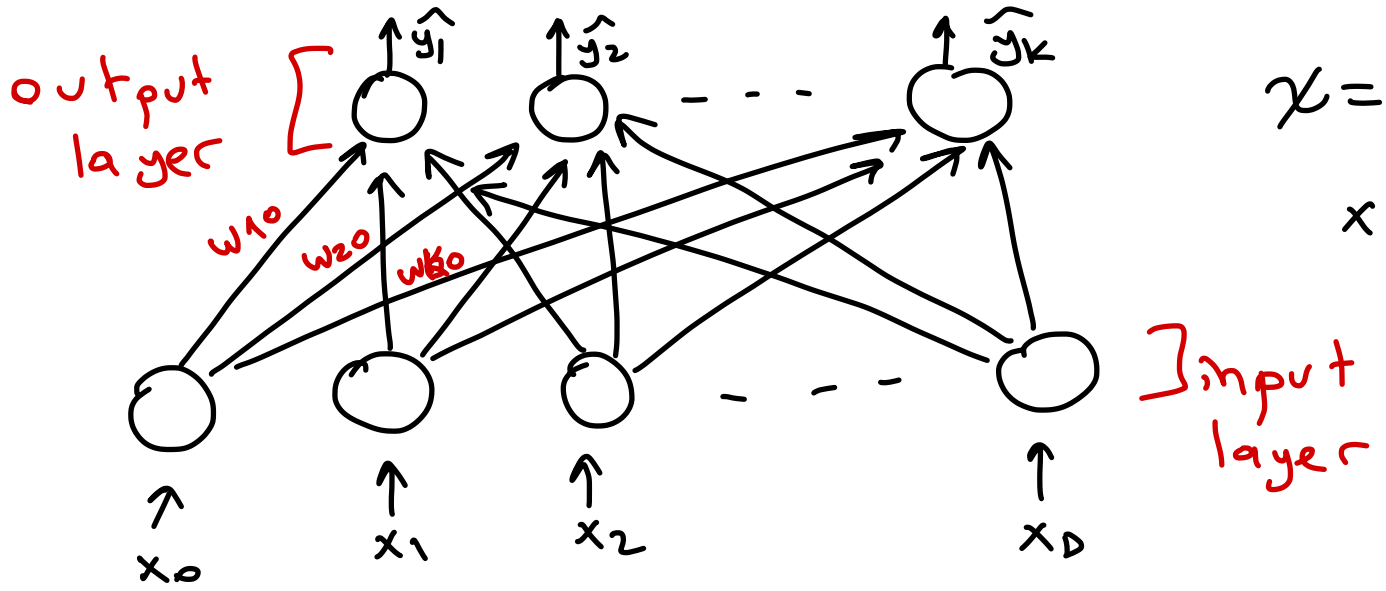
} binary classification

$$s(w^T \cdot x) = w^T \cdot x$$

↳ linear activation

} regression

$$s(a) = a$$



$$\mathcal{X} = \{(x_i, y_i)\}_{i=1}^{N}$$

$$x_i \in \mathbb{R}^D \qquad y_i \in \{1, 2, \dots, K\}$$

$$W_1 = \begin{bmatrix} w_{10} \\ w_{11} \\ w_{12} \\ \vdots \\ w_{1D} \end{bmatrix} \qquad W_2 = \begin{bmatrix} w_{20} \\ w_{21} \\ w_{22} \\ \vdots \\ w_{2D} \end{bmatrix} \qquad \cdots \qquad W_K = \begin{bmatrix} w_{K0} \\ w_{K1} \\ w_{K2} \\ \vdots \\ w_{KD} \end{bmatrix}$$

$$\hat{y}_c = \sum_{d=1}^{D} w_{cd} \cdot x_d + w_{c0} = W_c^T \cdot x$$

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_K \end{bmatrix}_{K \times 1} = \begin{bmatrix} w_{10} & w_{11} & w_{12} & \cdots & w_{1D} \\ w_{20} & w_{21} & w_{22} & \cdots & w_{2D} \\ \vdots & \vdots & \vdots & & \vdots \\ w_{K0} & w_{K1} & w_{K2} & \cdots & w_{KD} \end{bmatrix}_{K \times (D+1)} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix}_{(D+1) \times 1}$$

$$\Rightarrow \underbrace{\hat{y}}_{K \times 1} = \underbrace{W}_{K \times (D+1)} \cdot \underbrace{x}_{(D+1) \times 1}$$

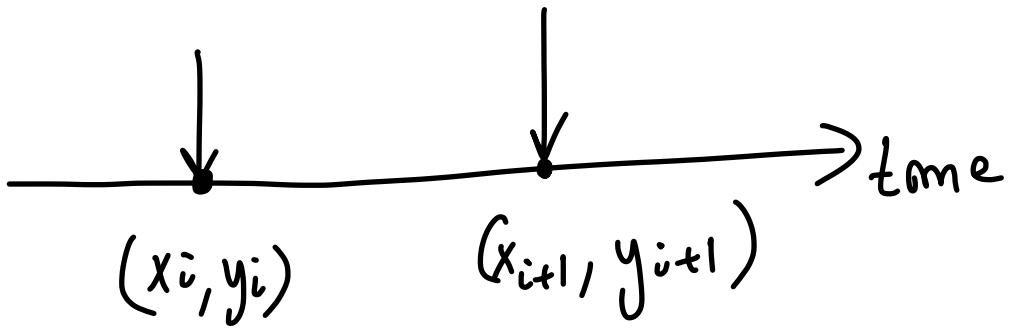
$$\hat{y}_c = \frac{\exp\left[w_c^T \cdot x\right]}{\sum_{k=1}^{K} \exp\left[w_k^T \cdot x\right]}$$

} softmax activation

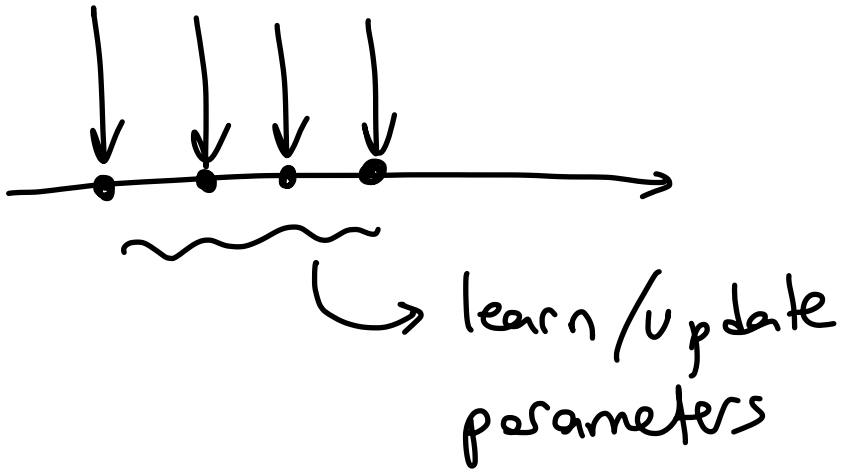
a new data point x^* $\Rightarrow$ choose $\hat{y}^* = \arg\max_c \hat{y}_c$

LEARNING

Online Learning vs Batch Learning



- samples are coming one by one



Regression

$$\text{Error}_i(w \mid x_i, y_i) = \frac{1}{2}(y_i - \hat{y}_i)^2 \quad \leftarrow \text{squared error}$$

$$= \frac{1}{2}\left[y_i - s(w^T \cdot x_i)\right]^2$$

$$= \frac{1}{2}\left[y_i - w^T \cdot x_i\right]^2$$

$$\frac{\partial \text{Error}_i}{\partial w} = \frac{1}{2} \cdot 2 \cdot \left[y_i - w^T \cdot x_i\right] \frac{\partial \left[y_i - w^T \cdot x_i\right]}{\partial w}$$

$$= \left[y_i - w^T \cdot x_i\right](-x_i) = -(y_i - \hat{y}_i) \cdot x_i$$

1×1 $\quad$ $(D+1) \times 1$

$y_i - \hat{y}_i$

$$\frac{1}{2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \frac{1}{2}(y_1 - \hat{y}_1)^2 + \frac{1}{2}(y_2 - \hat{y}_2)^2 + \cdots + \frac{1}{2}(y_N - \hat{y}_N)^2$$

$$\frac{\partial w^T \cdot x_i}{\partial w} = \frac{\partial x_i^T \cdot w}{\partial w} = x_i$$

$$\frac{\partial \text{Error}_i}{\partial w} = -(y_i - \hat{y}_i).x_i$$

$$\Delta w = -\eta \cdot \frac{\partial \text{Error}_i}{\partial w} = \boxed{\eta (y_i - \hat{y}_i).x_i}$$

Binary Classification

$$\text{Error}_i(w \mid x_i, y_i) = -\underbrace{\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right]}_{\text{log-likelihood}}$$

$$\hat{y}_i = s(w^T.x_i) = \frac{1}{1+\exp[-w^T.x_i]}$$

$$= -\left[y_i \log\left[\underbrace{\frac{1}{1+\exp[-w^T.x_i]}}_{f(w)}\right] + (1-y_i)\log\left[1 - \frac{1}{1+\exp[w^T x_i]}\right]\right]$$

Hint: $\frac{\partial \log(\hat{y}_i)}{\partial w} \Rightarrow \frac{\partial \log[f(w)]}{\partial w} = \frac{1}{f(w)} \cdot \frac{\partial f(w)}{\partial w}$

$$\frac{\partial Error_i(w \mid x_i, y_i)}{\partial w} = -(y_i - \hat{y}_i).x_i$$

$$\Delta w = -\eta \cdot \frac{\partial Error_i}{\partial w} = \boxed{\eta \cdot (y_i - \hat{y}_i).x_i}$$

Multiclass Classification

$$Error_i(\{w_c\}_{c=1}^{K} \mid x_i, y_i) = -\sum_{c=1}^{K} y_{ic} \log(\hat{y}_{ic})$$

log-likelihood

$$\hat{y}_{ic} = \frac{\exp[w_c^T . x_i]}{\sum_{k=1}^{K} \exp[w_k^T . x_i]}$$

$f(w)$

$$= -\sum_{c=1}^{K} y_{ic} \log\left[\frac{\exp[w_c^T . x_i]}{\sum_{k=1}^{K} \exp[w_k^T . x_i]}\right]$$

$$\frac{\partial Error_i(\{w_k\}_{k=1}^{K} \mid x_i, y_i)}{\partial w_c} = -(y_{ic} - \hat{y}_{ic}).x_i$$

$$\Delta w_c = -\eta \frac{\partial Error_i}{\partial w_c} = \boxed{\eta (y_{ic} - \hat{y}_{ic}).x_i}$$

Update = (Learning Factor) × (True Output − Predicted output) × (Input)

$[\eta] \times [(y_i - \hat{y}_i)] \times [x_i]$ ⇒ Regression

$[\eta] \times [(y_i - \hat{y}_i)] \times [x_i]$ ⇒ Binary Classification

$[\eta] \times [(y_{ic} - \hat{y}_{ic})] \times [x_i]$ ⇒ Multiclass Classification

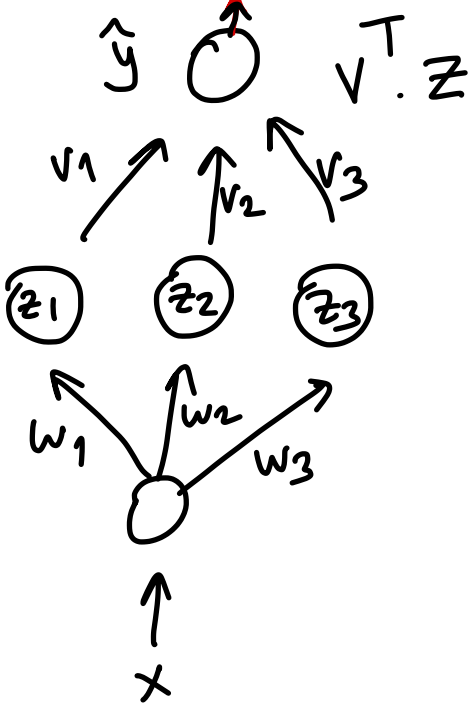
---

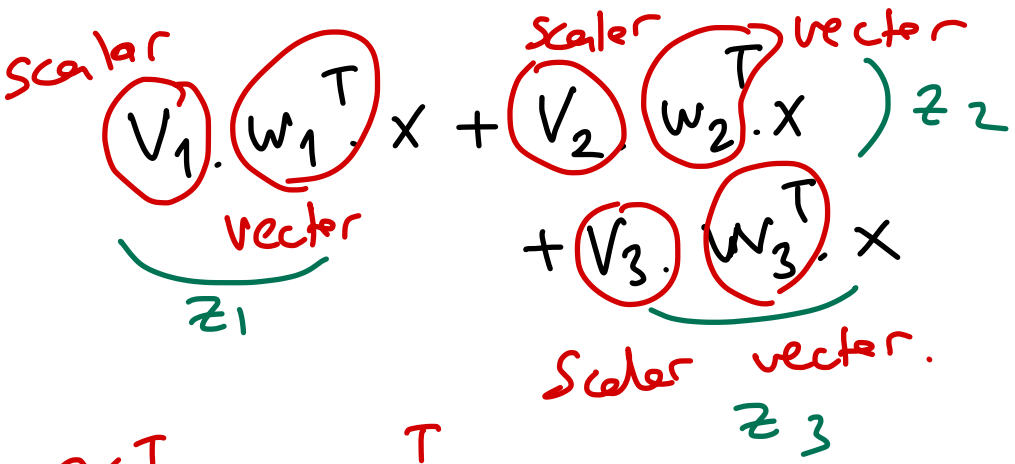
$f(x) = 2x$ ↳ linear

$g(x) = 3x$ ↳ linear

$f \circ g(x) = 6x$ ↳ linear

atleast $f(x)$ or $g(x)$ should be nonlinear

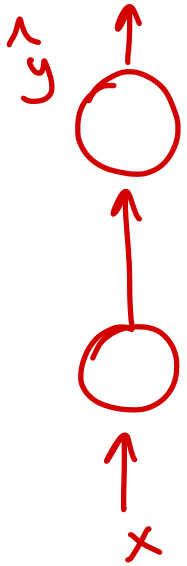




$$\tilde{w}_1^T = v_1 . w_1^T$$

$$\tilde{w}_2^T = v_2 . w_2^T$$

$$\tilde{w}_3^T = v_3 . w_3^T$$



$$\hat{y} = \left[\tilde{w}_1^T + \tilde{w}_2^T + \tilde{w}_3^T \right] . x$$