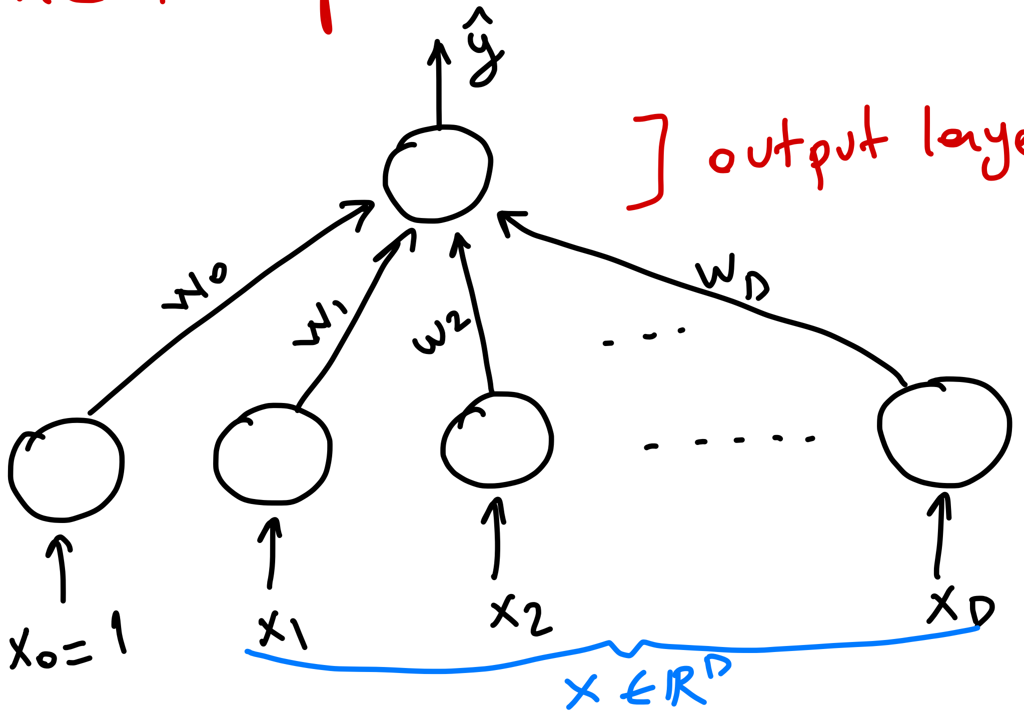
Multilayer Perceptrons

The Perceptron



] output layer

] input layer

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

$$= w_0.x_0 + \underbrace{\sum_{d=1}^{D} w_d . x_d}_{w^T.x}$$

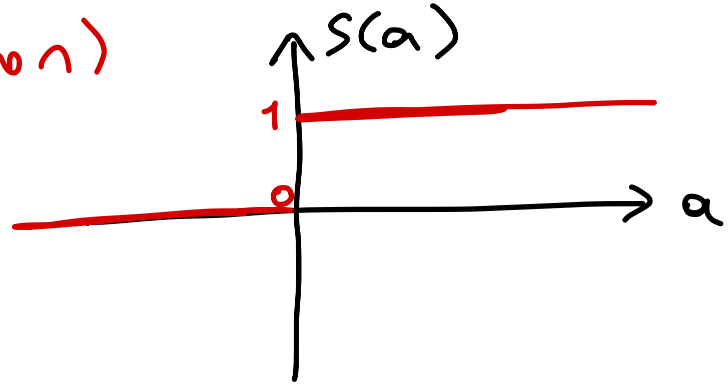
$$= w^T.x + w_0$$

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

$$= \begin{bmatrix} w_0 & w_1 & w_2 & \dots & w_D \end{bmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} = \underset{1 \times (D+1)}{w^T} . \underset{(D+1) \times 1}{x}$$

threshold function (activation function)

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



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

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

→ sigmoid activation

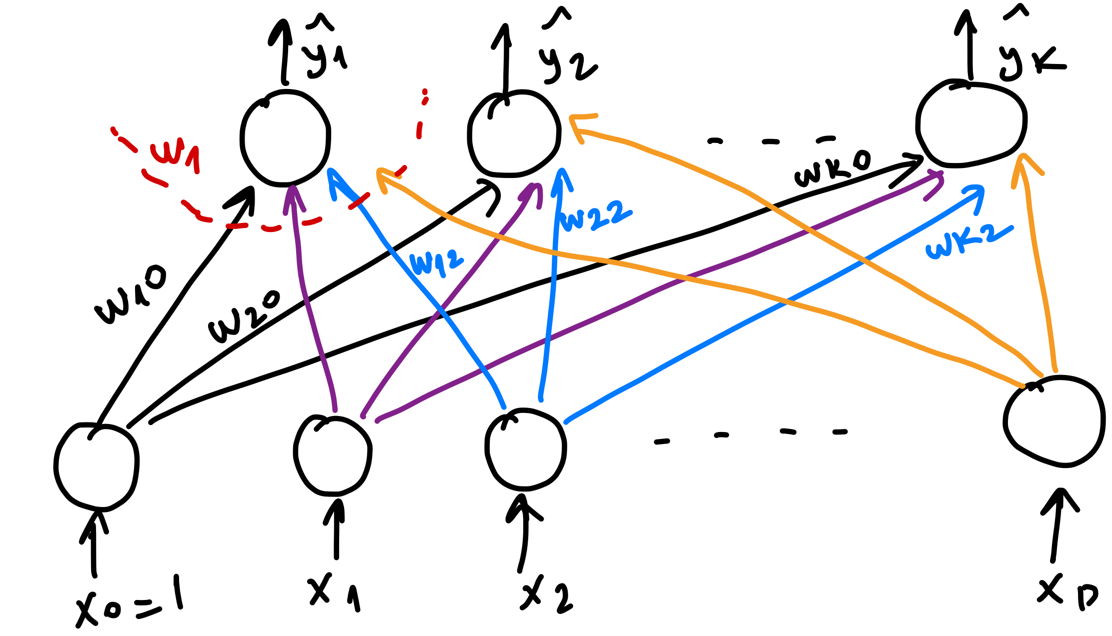
} binary classification

$$s(w^T.x) = w^T.x$$

→ linear activation

} regression

{ 10^{10} neurons
each neuron is connected to 10–1000 neurons.



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

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

$$w_1 = \begin{bmatrix} w_{10} \\ w_{11} \\ w_{12} \\ \vdots \\ w_{1D} \end{bmatrix} \quad w_2 = \begin{bmatrix} w_{20} \\ w_{21} \\ w_{22} \\ \vdots \\ w_{2D} \end{bmatrix} \quad \cdots \quad w_K = \begin{bmatrix} w_{K0} \\ w_{K1} \\ w_{K2} \\ \vdots \\ w_{KD} \end{bmatrix}$$

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

$$\underbrace{\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_K \end{bmatrix}}_{K\times 1} = \underbrace{\begin{array}{c} \\ \end{array}\begin{array}{r} W_1^T \\ W_2^T \\ \\ W_K^T \end{array}\begin{bmatrix} w_{10} & w_{11} & w_{12} & \cdots & w_{1D} \\ w_{20} & w_{21} & w_{22} & \cdots & w_{2D} \\ & & \vdots & & \\ w_{K0} & w_{K1} & w_{K2} & \cdots & w_{KD} \end{bmatrix}}_{K\times(D+1)} \underbrace{\overset{X}{\begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix}}}_{(D+1)\times 1}$$

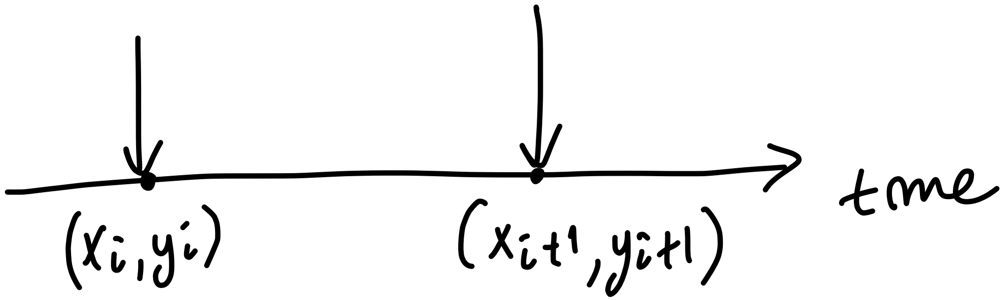
$$\Rightarrow \underbrace{\hat{y}}_{K\times 1} = \underbrace{W}_{K\times(D+1)} . \underbrace{X}_{(D+1)\times 1}$$

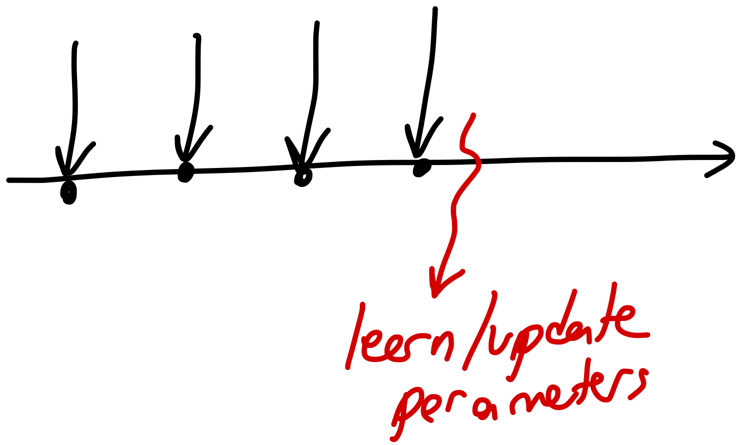
$$\hat{y}_c = \frac{\exp(w_c^T . x)}{\sum_{d=1}^{K} \exp(w_d^T . x)}$$ } softmax activation

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

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

$$= \frac{1}{2}(y_i - s(w^T.x_i))^2$$

$$= \frac{1}{2}(y_i - w^T.x_i)^2$$

$$f(x) = w_0.\cancel{x_0}x_0 + w_1.x_1 + w_2.x_2$$

$$\frac{\partial f(x)}{\partial w_0} = x_0 \qquad \frac{\partial f(x)}{\partial w_1} = x_1 \qquad \frac{\partial f(x)}{\partial w_2} = x_2$$

$$\frac{\partial \text{Error}_i}{\partial w} = \frac{1}{\cancel{2}}(y_i - w^T.x_i)^{\cancel{2-1}\,1} . \cancel{2} . \frac{\partial(y_i - w^T.x_i)}{\partial w} \rightarrow -x_i$$

$$= (\underbrace{y_i - w^T.x_i}_{\hat{y}_i}).(-x_i) = -(y_i - \hat{y}_i).x_i$$

$$\frac{\partial f(x)}{\partial W} = \begin{bmatrix} \frac{\partial f(x)}{\partial w_0} \\ \vdots \\ \frac{\partial f(x)}{\partial w_2} \end{bmatrix} = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} = X$$

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

$$= \sum_{i=1}^{N}\text{Error}_i$$

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

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

Binary Classification

$$Error_i(w \mid x_i, y_i) = -\left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right]$$

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

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

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

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

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

Multiclass Classification

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

$$= -\sum_{c=1}^{K} y_{ic} \log\left[\frac{\exp[w_c^T \cdot x_i]}{\sum_{d=1}^{K} \exp[w_d^T \cdot x_i]}\right]$$

$$\hat{y}_{ic} = \frac{\exp[w_c^T \cdot x_i]}{\sum_{d=1}^{K} \exp[w_d^T \cdot x_i]}$$

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

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

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

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

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

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

$f(x) = 2x$

$2x$

↳ linear function

$g(x) = 3x$

$\log(x)$

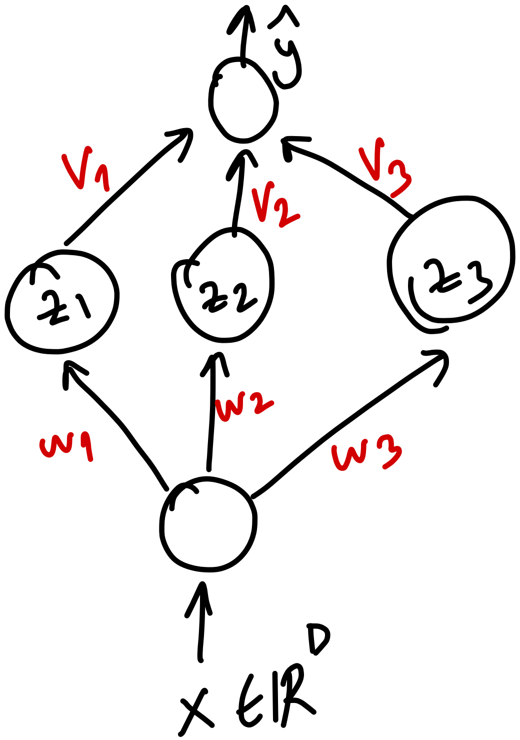
↳ linear function

$f \circ g(x) = 6x$

↳ linear function

$2\log(x)$

at least $f(x)$ or $g(x)$ should be nonlinear.



$$z_1 = w_1^T \cdot x$$

$$z_2 = w_2^T \cdot x$$

$$z_3 = w_3^T \cdot x$$

$$\hat{y} = v_1 \cdot z_1 + v_2 \cdot z_2 + v_3 \cdot z_3$$

$$\hat{y} = \underbrace{v_1 \cdot w_1^T}\cdot x + \underbrace{v_2 \cdot w_2^T}\cdot x + \underbrace{v_3 \cdot w_3^T}\cdot x$$

$$\hat{y} = \tilde{w}_1^T \cdot x + \tilde{w}_2^T \cdot x + \tilde{w}_3^T \cdot x$$

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