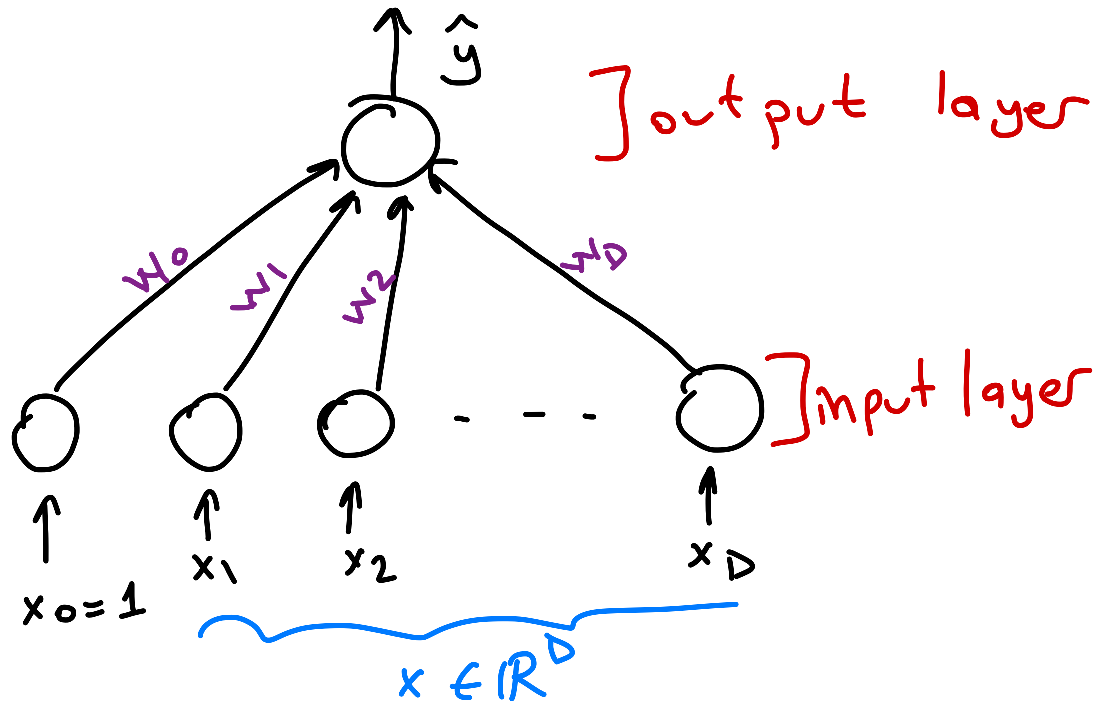
The Perceptron



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

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

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

total weighted signal received

$w^T . x + w_0 > 0 \Rightarrow$ feel pain

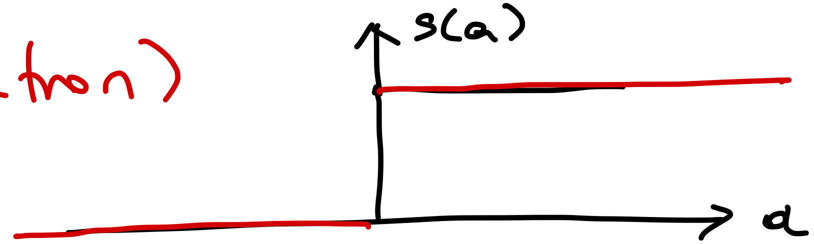
$w^T . x + w_0 < 0 \Rightarrow$ does not ~~feel~~ feel

$w^T . x > \boxed{-w_0} \Leftarrow$ threshold

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

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

threshold function (activation function)



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

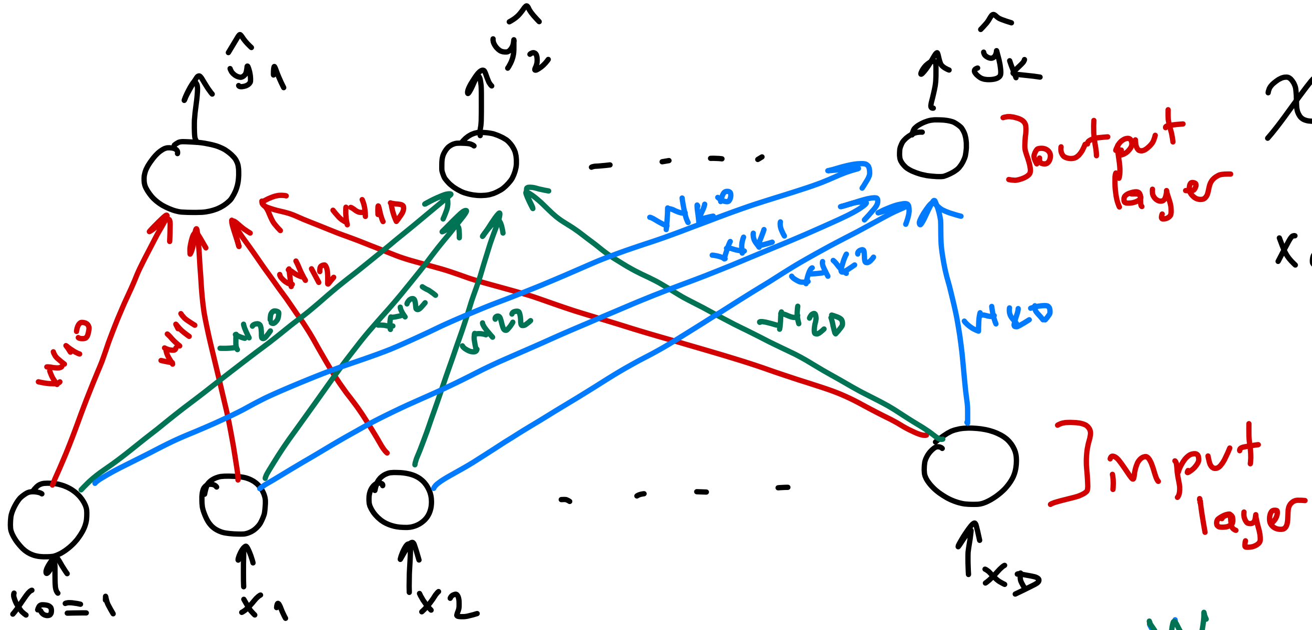
$$s(w^T.x) = \begin{cases} 1 & \text{if } w^T.x > 0 \\ 0 & \text{otherwise} \end{cases}$$

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

↳ sigmoid activation

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

↳ linear activation



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

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

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

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

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

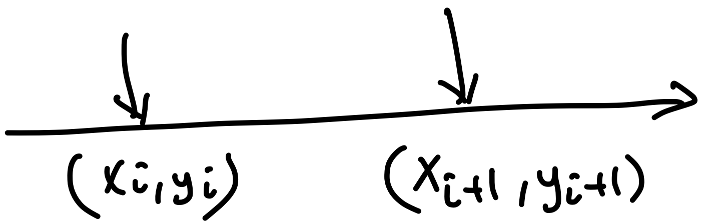
$$\underset{K \times 1}{\hat{y}} = \underset{K \times (D+1)}{W} \cdot \underset{(D+1) \times 1}{x}$$

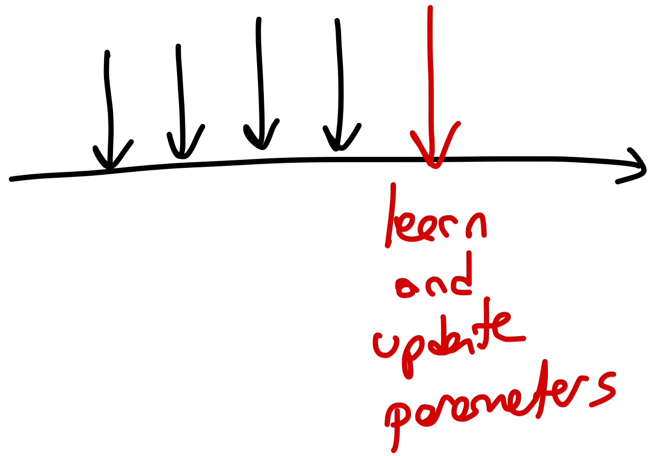
$$\hat{y}_c = \frac{\exp(w_c^T . x)}{\sum_{d=1}^{K} \exp(w_d^T . x)}$$ softmax activation

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

LEARNING

Online Learning versus Batch Learning





→ samples are coming one by one

Regression

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

$y_i \in \mathbb{R}$

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

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

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

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

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

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

Binary Classification

$\{(x_i, y_i)\}_{i=1}^{N} \Rightarrow y_i \in \{0,1\}$

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

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

Hint: $\dfrac{\partial \log(\hat{y}_i)}{\partial w}$

$$\frac{\partial \log(f(w))}{\partial w} = \frac{1}{f(w)}\frac{\partial f(w)}{\partial w}$$

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

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

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

Multiclass Classification

$$\text{Error}_i\left(\{w_c\}_{c=1}^{K} \mid x_i, y_i\right) = -\sum_{c=1}^{K} y_{ic} \log(\underbrace{\hat{y}_{ic}}_{})$$

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

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

$$\frac{\partial \text{Error}_i}{\partial w_c} = -(y_{ic} - \hat{y}_{ic}) . x_i$$

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

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

$f(x) = 2x$ $g(x) = 3x$

↳ linear func ↳ linear func.

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

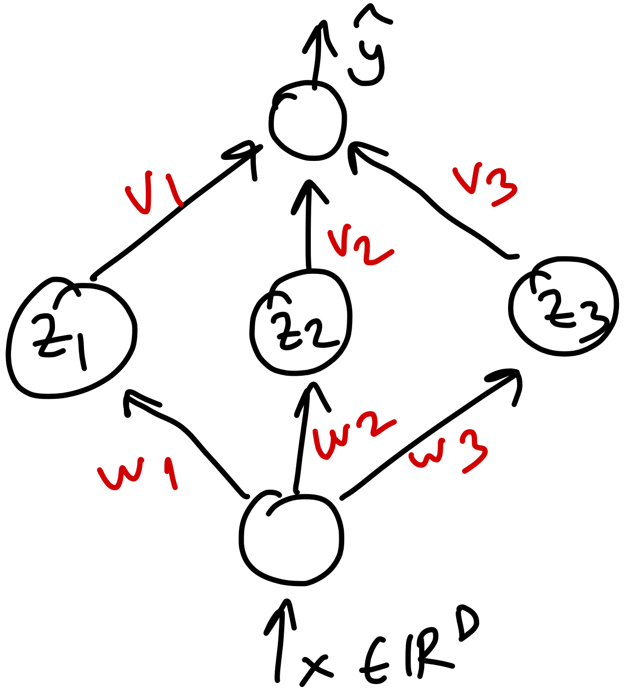
↳ linear func.

$f(x) = 2x$ $g(x) = \log(x)$

↳ linear ↳ nonlinear

$f \circ g(x) = 2\log(x)$

↳ nonlinear



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

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

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

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

$$\hat{y} = \underbrace{v_1 \cdot w_1^T}_{\tilde{w}_1^T} \cdot x + \underbrace{v_2\, w_2^T}_{\tilde{w}_2^T} \cdot x + \underbrace{v_3\, w_3^T}_{\tilde{w}_3^T} \cdot x$$

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

$$= (\tilde{w}_1 + \tilde{w}_2 + \tilde{w}_3)^T \cdot x$$