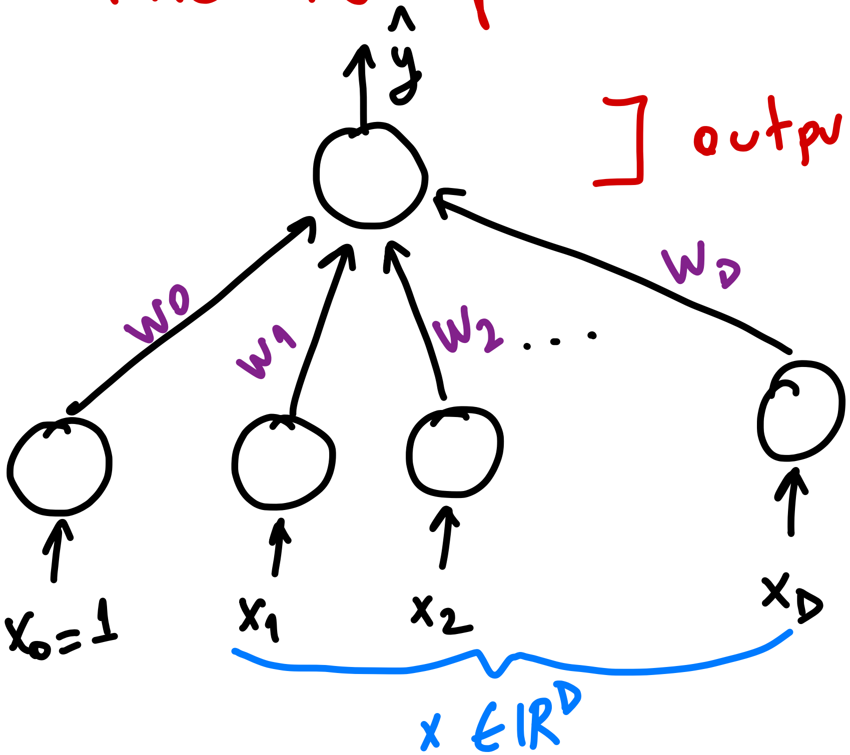
The Perceptron



] output layer

] input layer

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

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

$$= \underbrace{w^T . x + w_0}_{\text{total weighted signal received by output}}$$

pain threshold

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

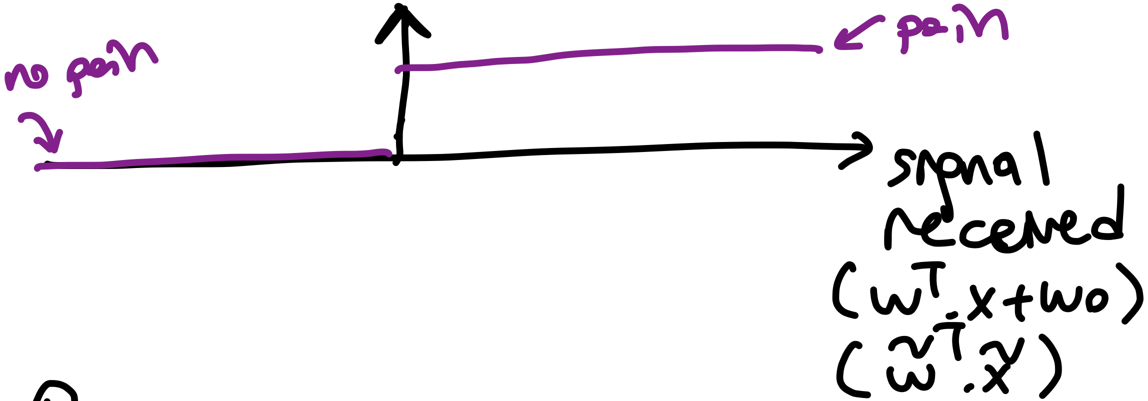
$w^T . x < -w_0 \Leftarrow w^T . x + w_0 < 0$ do not feel pain

$$\hat{y} = [w_1 \ w_2 \ \dots \ w_D] \begin{bmatrix} \overset{1}{x_1} \\ x_2 \\ \vdots \\ x_D \end{bmatrix} + w_0$$

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

Activation Function

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



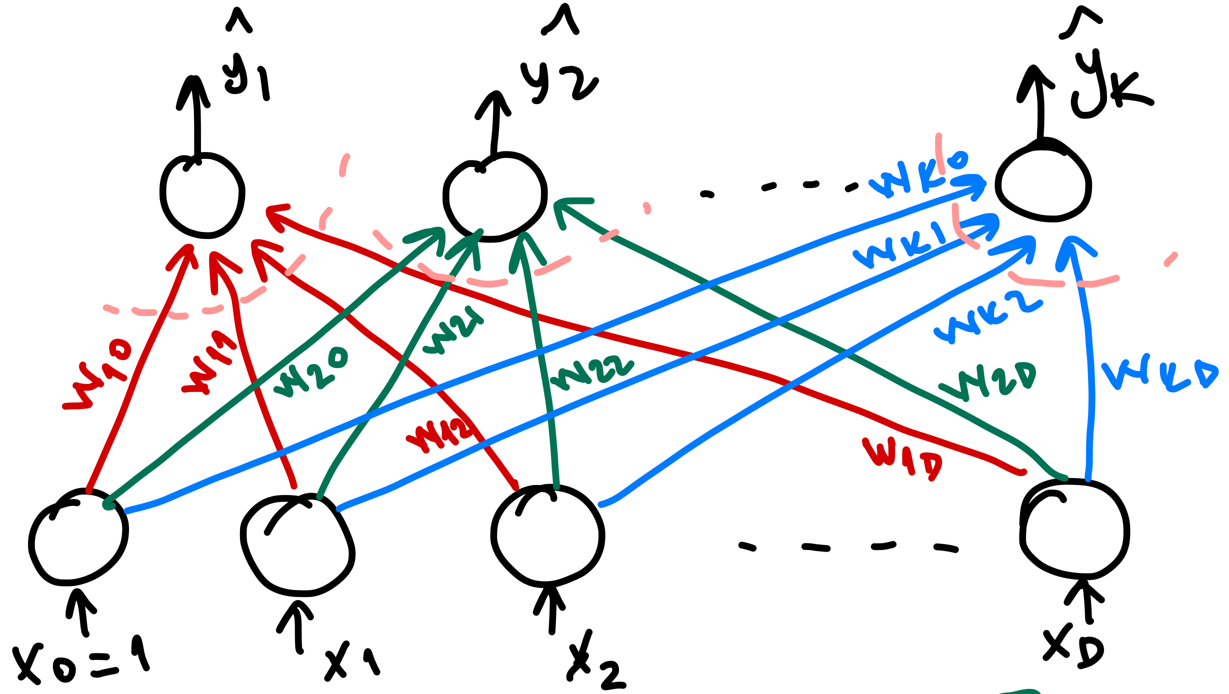
$$s(\tilde{w}^T . \tilde{x}) = \begin{cases} 1 & \text{if } \tilde{w}^T . \tilde{x} > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$s(\tilde{w}^T . \tilde{x}) = \frac{1}{1 + \exp\left[-\tilde{w}^T . \tilde{x}\right]} \quad \Big\} \text{ binary classification}$$

→ sigmoid activation

$$s(\tilde{w}^T . \tilde{x}) = \tilde{w}^T . \tilde{x} \quad \Big\} \text{ regression}$$

→ unit activation
linear activation



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

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

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

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

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

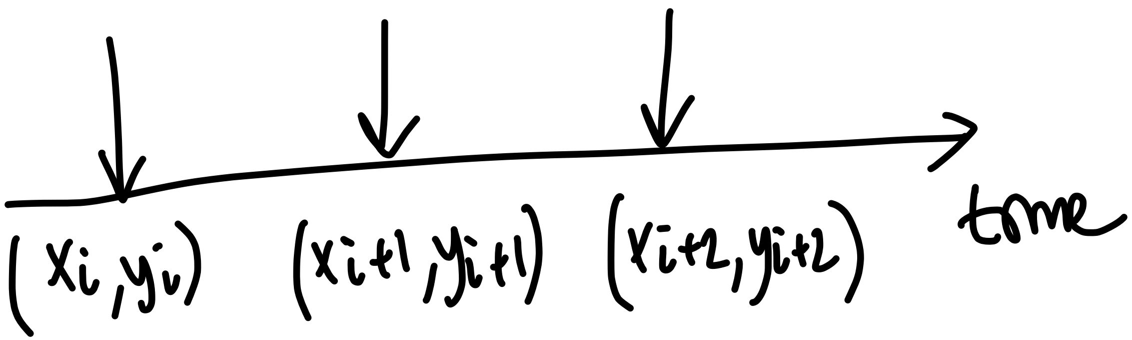
$$\hat{y} = \tilde{W} \cdot \tilde{x}$$

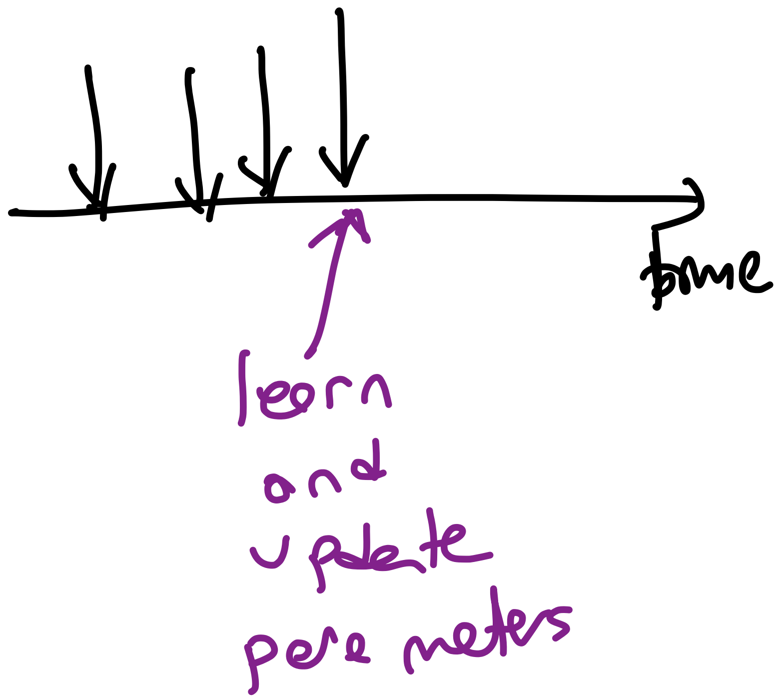
$$\left.\hat{y}_c = \frac{\exp(\tilde{w}_c^T \tilde{x})}{\sum_{d=1}^{K} \exp(\tilde{w}_d^T \cdot \tilde{x})}\right\} \text{softmax activation}$$

a new data point $x_{N+1} \Rightarrow$ choose $\hat{y} = \arg\max_c \hat{y}_c$

LEARNING

Online Learning vs Batch Learning.





→ samples are coming one by one.

Regression:

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

$x_i \in \mathbb{R}^D$

$y_i \in \mathbb{R}$

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

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

$$= \frac{1}{2}\left(y_i - s(\tilde{w}^T \cdot \tilde{x}_i)\right)^2$$

unit act.

$$= \frac{1}{2}\left(y_i - \tilde{w}^T \cdot \tilde{x}_i\right)^2$$

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

$$= -(y_i - \hat{y}_i) \cdot \tilde{x}_i$$

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

Binary Classification

$$-\sum_{i=1}^{N}\left(y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right)$$

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

$x_i \in \mathbb{R}^D$

$y_i \in \{0,1\}$

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

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

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

$$\frac{\partial \text{Error}_i}{\partial \tilde{w}} = \underline{-(y_i - \hat{y}_i) \cdot \tilde{x}_i}$$

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

Multiclass Classification

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

$x_i \in \mathbb{R}^D$

$y_i \in \{1, 2, \dots, K\}$

$$-\sum_{i=1}^{N}\sum_{c=1}^{K} y_{ic} \log(\hat{y}_{ic})$$

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

$$\hat{y}_{ic} = \frac{\exp[\tilde{w}_c^T \cdot \tilde{x}_i]}{\sum_{d=1}^{K} \exp[\tilde{w}_d^T \cdot \tilde{x}_i]}$$

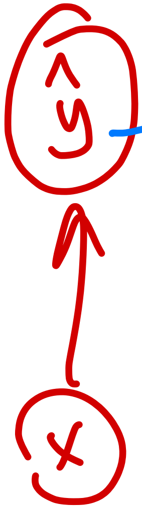
$$= -\sum_{c=1}^{K} y_{ic} \log\left[\frac{\exp[\tilde{w}_c^T \tilde{x}_i]}{\sum_{d=1}^{K} \exp[\tilde{w}_d^T \cdot \tilde{x}_i]}\right]$$

$$\frac{\partial \text{Error}_i}{\partial \tilde{w}_c} = -\underline{(y_{ic} - \hat{y}_{ic}) \cdot \tilde{x}_i}$$

$$\Delta \tilde{w}_c = -\eta \frac{\partial \text{Error}_i}{\partial \tilde{w}_c} = \eta \cdot (y_{ic} - \hat{y}_{ic}) \cdot \tilde{x}_i$$

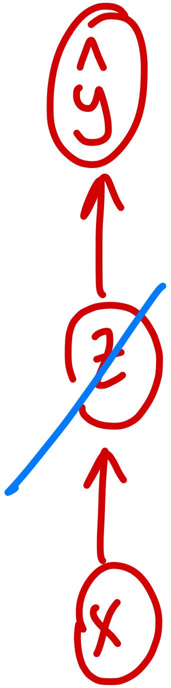
All three problems ↓

Update = (Learning Factor) x [True output - predicted output] x (Input)



$\hat{y} = f(x)$

$= 8x+7$



$\hat{y} = g(z) = g(f(x))$

linear

$\rightarrow z = \boxed{f(x)}$

linear

$g(z) = 4z - 5$

$f(x) = 2x+3$

$8x+7$