DASC 550 : Applied Deep Learning

→ Grading : Discussions ($\boxed{10\%}$)
Homeworks (90%) ⇒ 3 homeworks / projects

→ Lectures : Fridays ⇒ mathematical details
Saturdays / Sundays ⇒ implementation

Jupyter Notebooks → Python 3.X
→ numpy
scipy
pandas
sklearn
matplotlib

seaborn
tensorflow
⋮

→ Hybrid (SNA 104 and Zoom)

office : ENG 118

mehmetgonen@ku.edu.tr

Supervised Learning

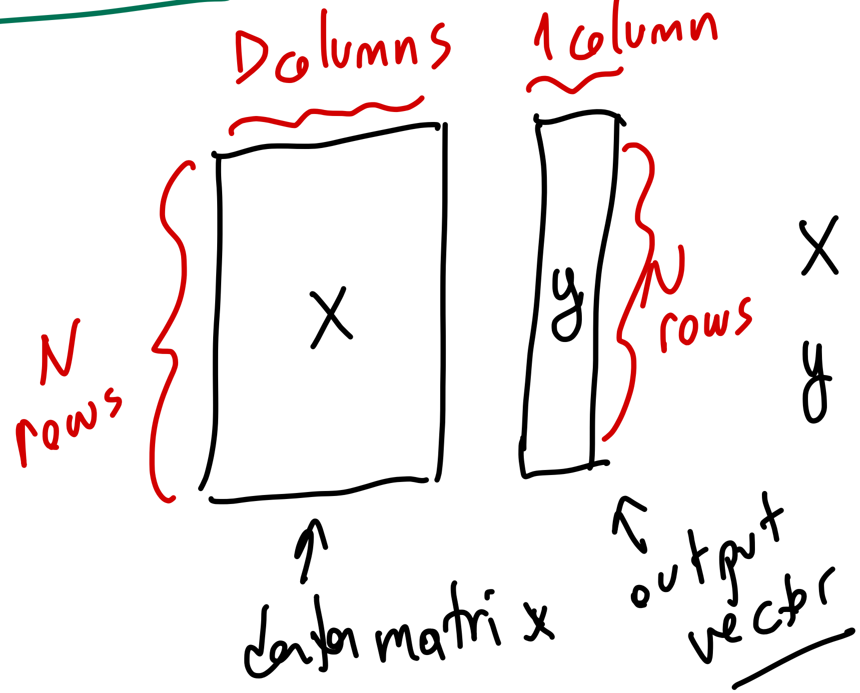
- ↳ regression ✓
- ↳ classification ✓
- ↳ multioutput regression ✓
- ↳ multilabel classification ✓

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

$x_i \in \mathbb{R}^D$ ↓ data point

$y_i \in \mathbb{R}$ ↓ output (target)

Regression



$X \in \mathbb{R}^{N\times D}$

$y \in \mathbb{R}^{N\times 1}$

Classification

$$\mathcal{X} = \{(x_i, y_i)\}_{i=1}^{N} \qquad x_i \in \mathbb{R}^D \qquad y_i \in \{0,1\}$$

or

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

$$y = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \\ 3 \\ 3 \end{bmatrix} \begin{array}{l} \leftarrow \text{dog} \\ \leftarrow \text{cat} \\ \\ \leftarrow \text{lion} \\ \\ \\ \end{array} \Rightarrow Y = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

one-hot-encoding

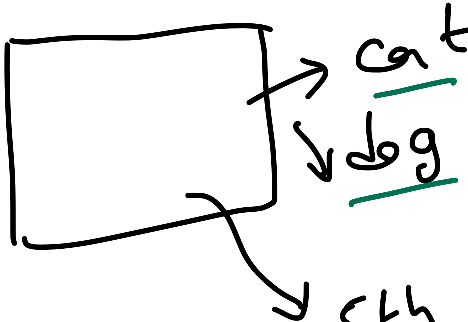
label vector

label matrix

mutually exclusive

softmax activation

Multilabel Classification



5th picture (x_5)

$$\begin{array}{ccc} \text{cat} & \text{dog} & \text{lion} \end{array}$$

$$Y = \begin{bmatrix} \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \end{bmatrix}$$

5th row

not mutually exclusive.

separate sigmoid activations.

$2^3 = 8$ possible labels

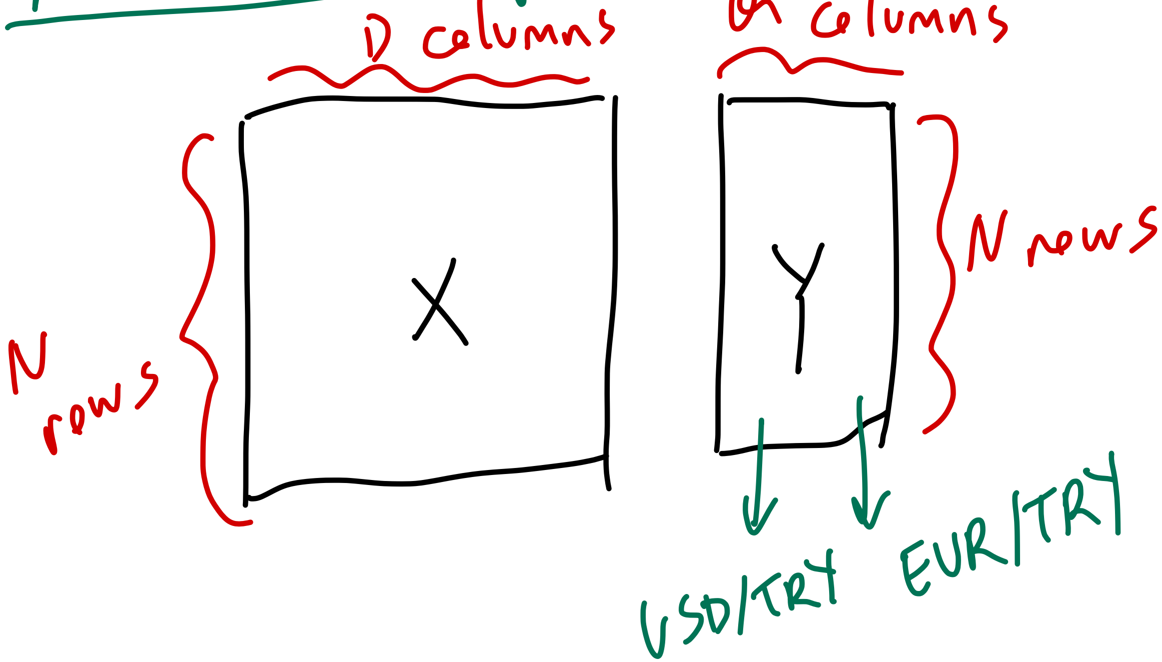
$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \end{bmatrix}$$

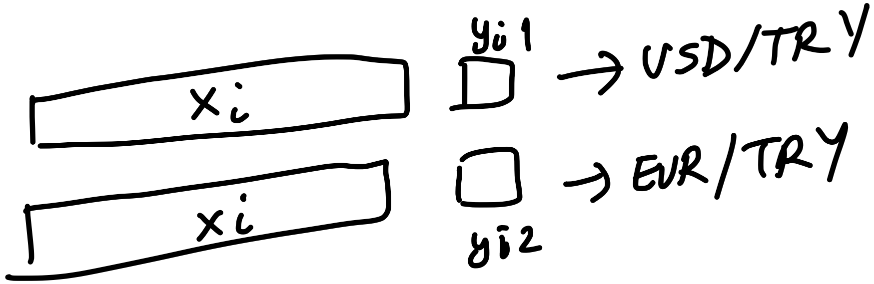
no animal | cat | cat & dog | cat & dog & lion | dog | dog & lion | cat & lion | lion

\# of animals : 10 $\Rightarrow$ \# of classes $= 2^{10} = 1024$

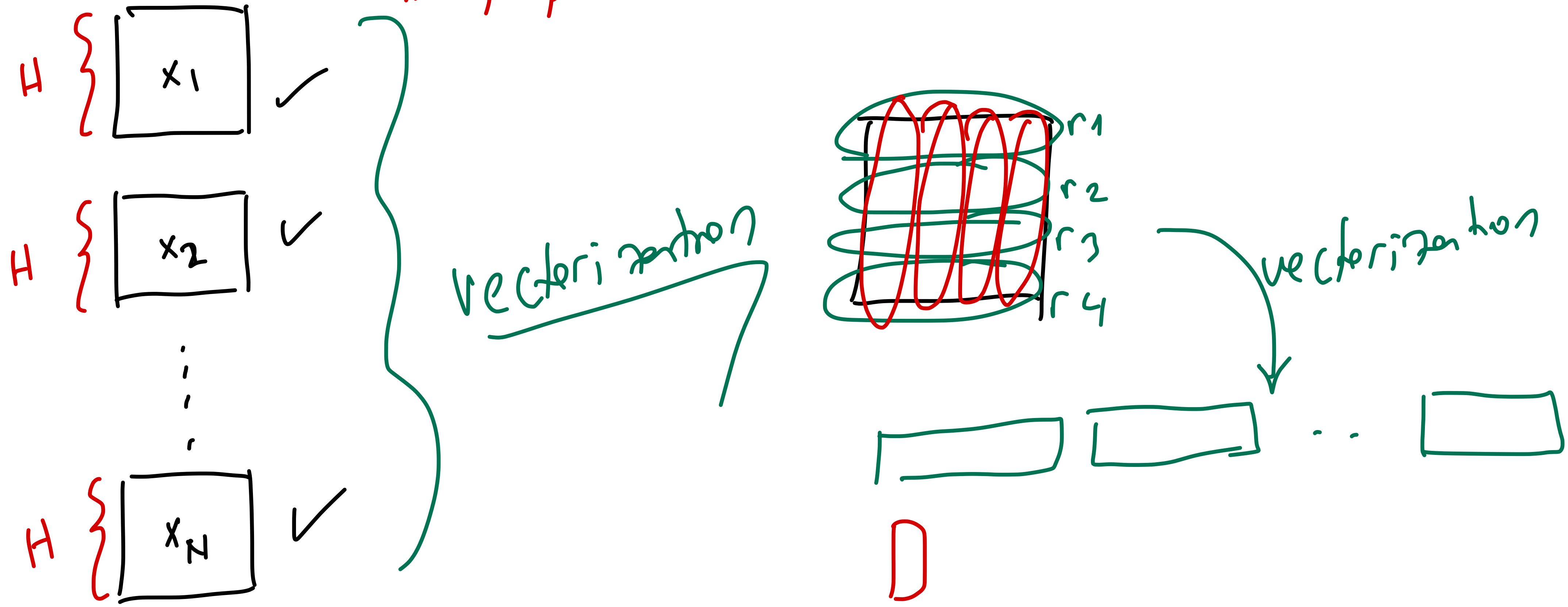
Multioutput Regression

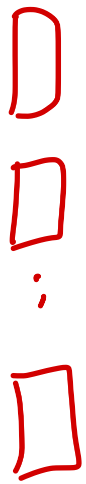
$$\mathcal{X} = \{(x_i, y_i)\}_{i=1}^{N} \qquad x_i \in \mathbb{R}^D \qquad y_i \in \mathbb{R}^Q, \quad Q \geq 2$$





# of pixels $= H \times W$

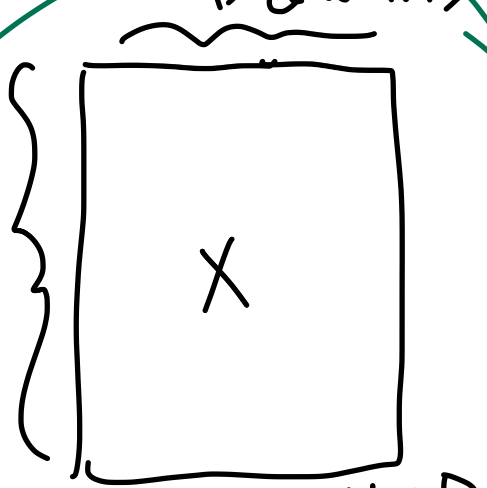




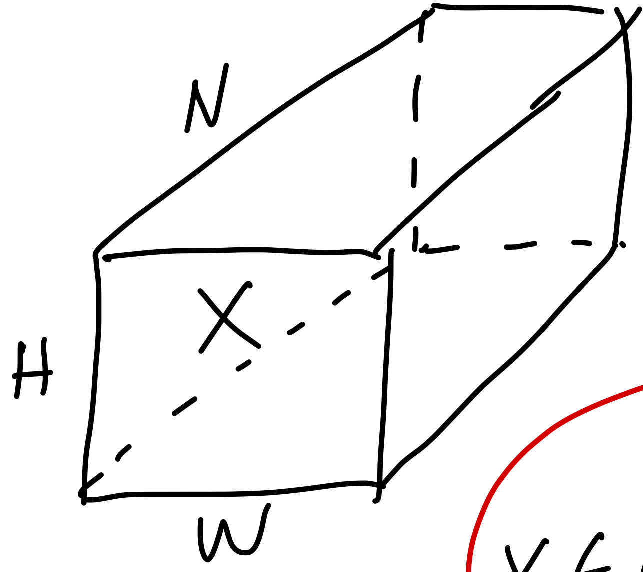
$$x_i \in \mathbb{R}^{H \times W} \longrightarrow x_i \in \mathbb{R}^{D}$$

matrix

where $D = H \times W$



$X \in \mathbb{R}^{N \times D}$

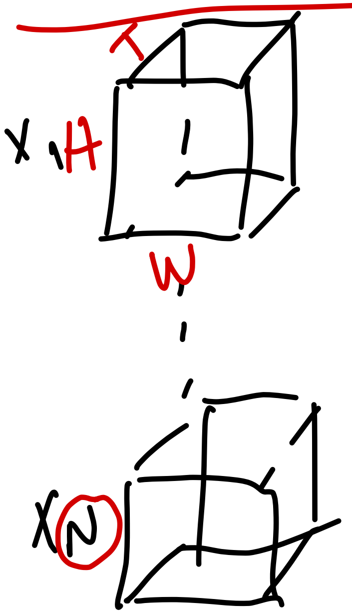


\# of data points

$X \in \mathbb{R}^{N \times H \times W}$

or

$X \in \mathbb{R}^{H \times W \times N}$



→ video recording.
(2s)
$T = 120$

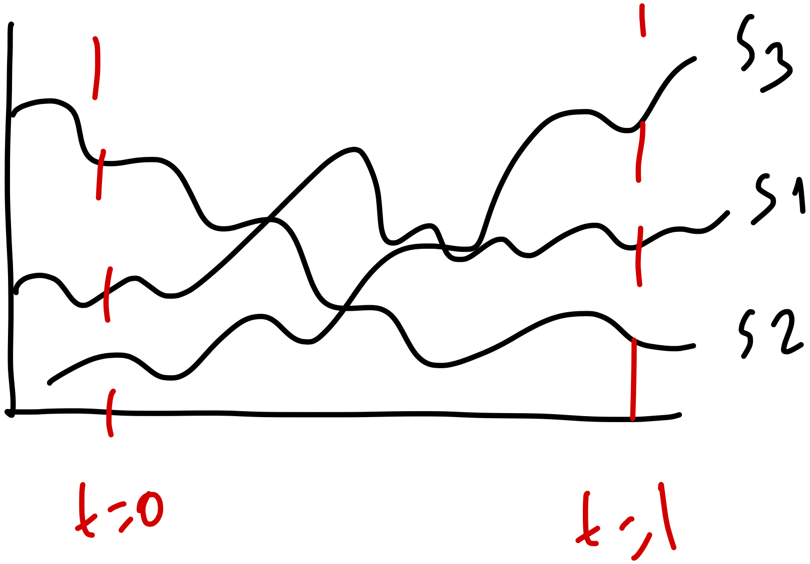
(2s)

30 fps
60 fps

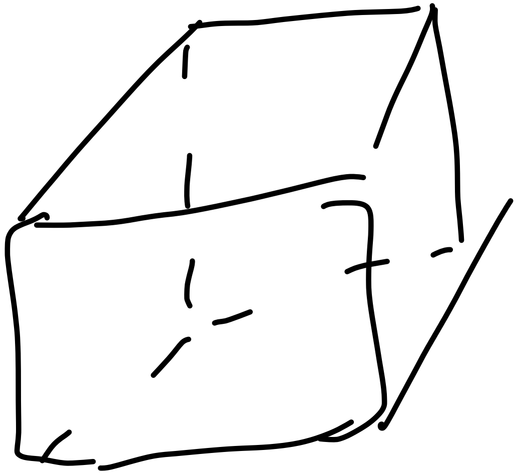
\# of data points, height, width, frame count

$X \in \mathbb{R}^{N \times \boxed{H \times W \times T}}$

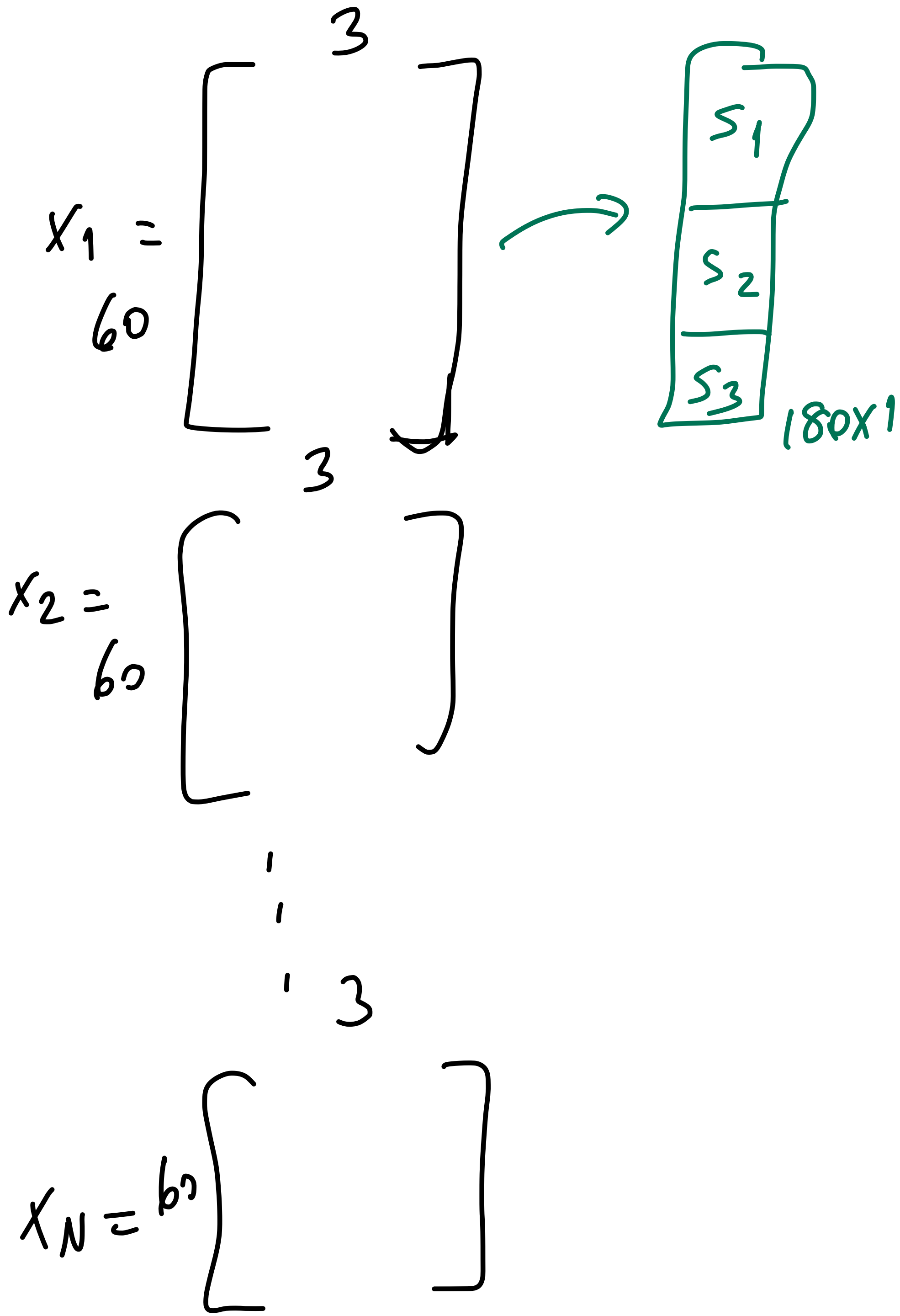
$D = H \times W \times T$



$X =$



$X \in \mathbb{R}^{N \times 60 \times 3}$



Vector to Real Valued $\quad \mathbb{R}^D \to \mathbb{R}$

Matrix to $\to$ $\quad \mathbb{R}^D \to \{0, 1\}$

Tenser to $\to$ $\quad \mathbb{R}^D \to \{1, 2, \dots, K\}$

Sequences

7 words S1 I did not like this movie at all $\Rightarrow$ $-$

8 words S2 We spent two hours at the movie theatre. $\Rightarrow$ $\varnothing$

10 words We liked the movie and watched for the second time. $\Rightarrow$ $+$

Sequence-to-Real Valued Problems $\to$

BoW approach

(Bag-of-Words)

TfIdf approach

sequence ⇒ vector

	like	not	spent	hours	two
S1	1	1	0	0	0
S2	0	0	1	1	1
S3	1	0	0	0	0

Sequence-to-Sequence problems

Machine Translation (MT)

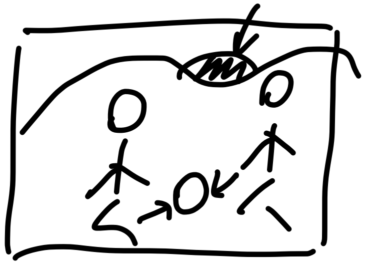
This course is a great course. (EN)

⇓

Bu ders, süper bir ders. (TR)

Vector-to-Sequence problems

Image-to-Text (Image Captioning)



Two kids are playing soccer on the field.

while sun is setting, two happy kids are playing together.

⋮

P_1 → $P_1 - S_1$, $P_1 - S_2$

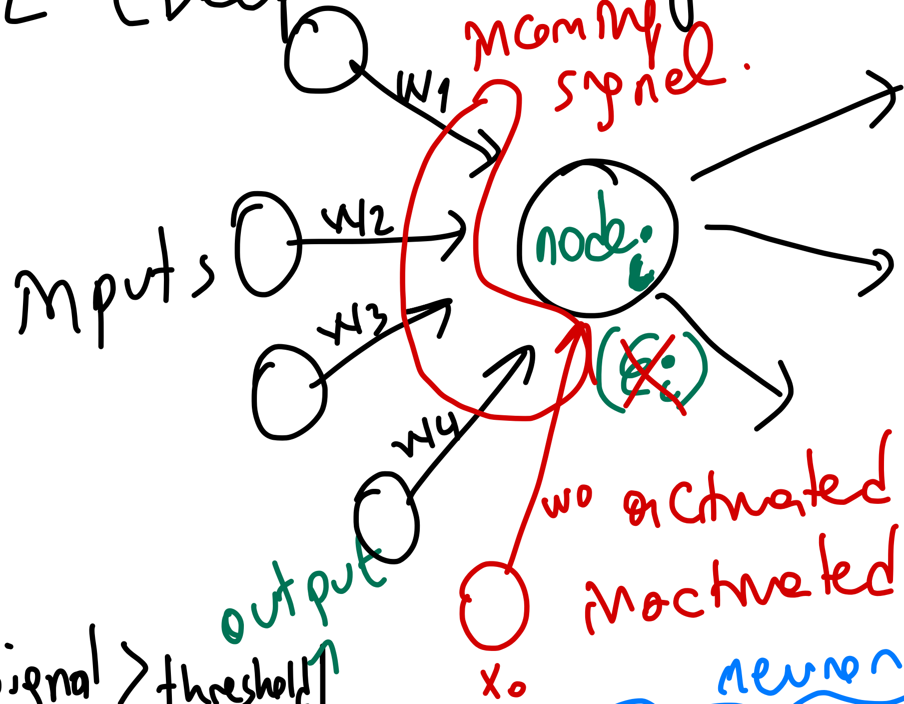
P_2 → $P_2 - S_1$, $P_2 - S_2$, $P_2 - S_3$

P_{N+1} → ⋮

Mathematical Foundations

ANNs (Artificial Neural Networks)

DL (Deep Learning)

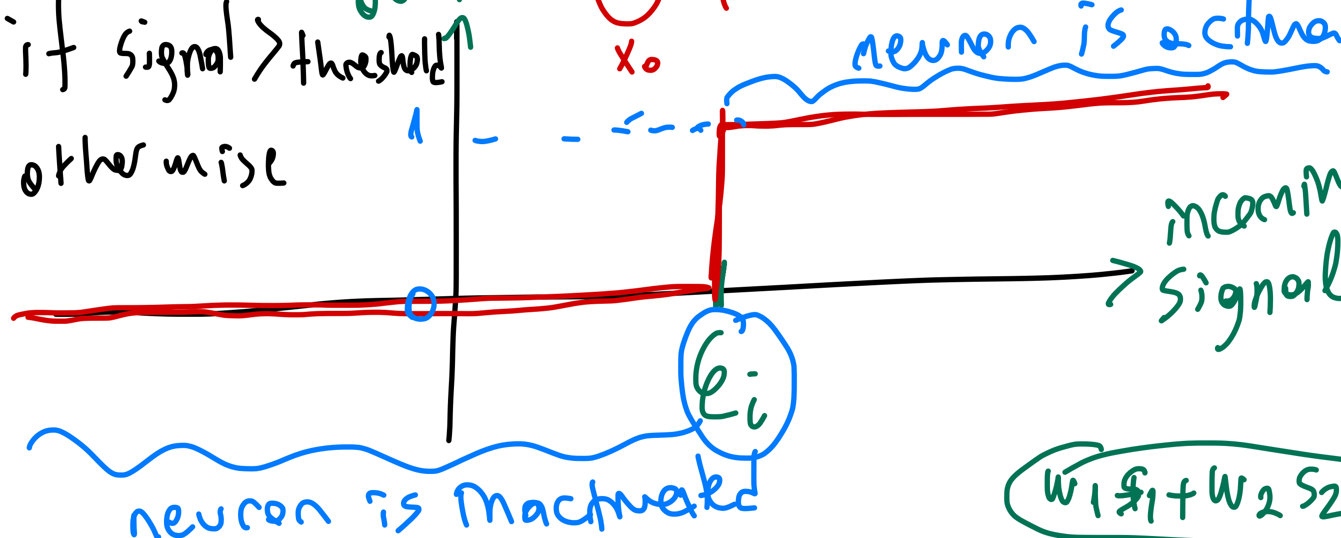


θ_i = threshold of node i

$$\boxed{\theta_i = -w_0 s_0}$$

1

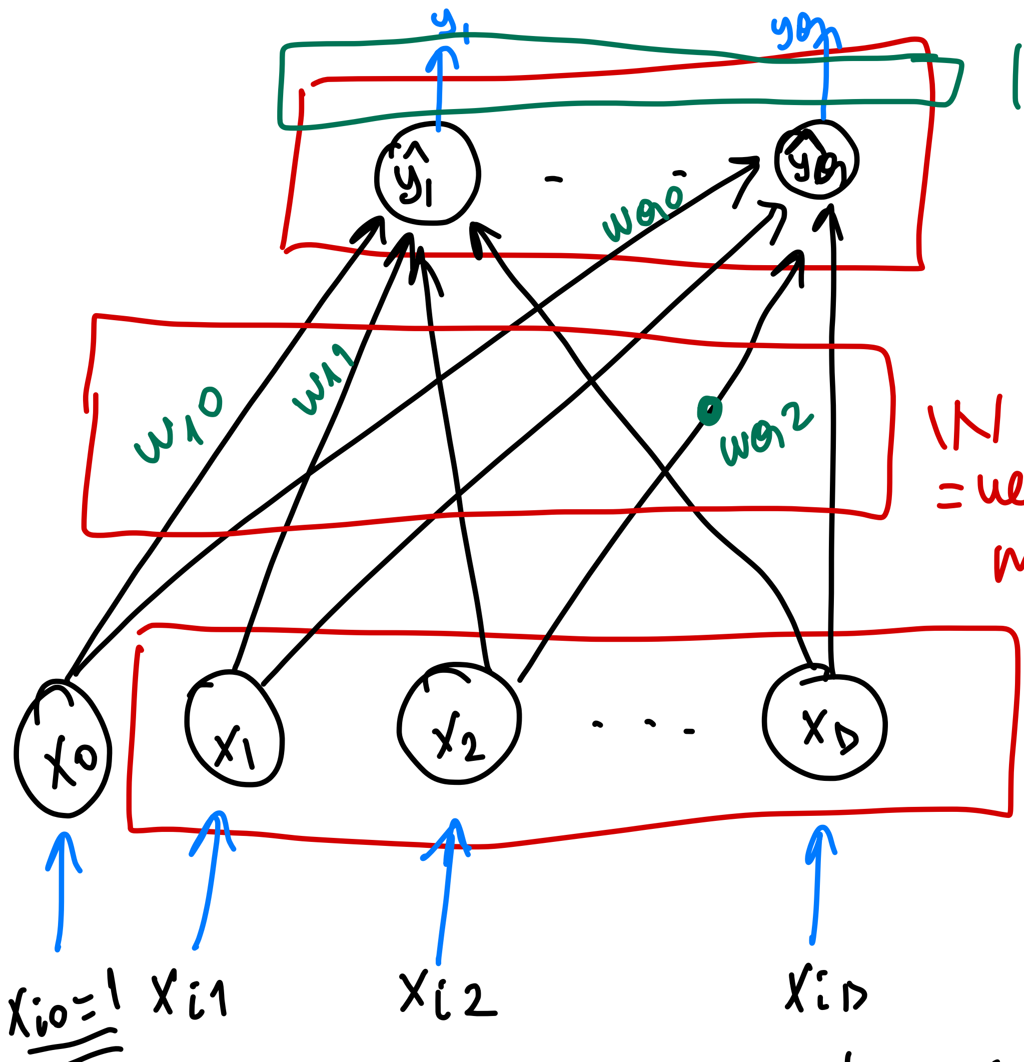
$$\begin{cases} 1 & \text{if signal} > \text{threshold} \\ 0 & \text{otherwise} \end{cases}$$



$$w_1 . s_1 + w_2 s_2 + w_3 . s_3 + w_4 . s_4 > \theta_i$$

$$w_1 s_1 + w_2 s_2 + w_3 s_3 + w_4 s_4 + w_0 s_0 > 0$$

loss function $(y_i, \hat{y}_i)$



← output layer

W = weigh matrix.

w_{qd} = the edge that connect $\hat{y}_q$ with x_d.

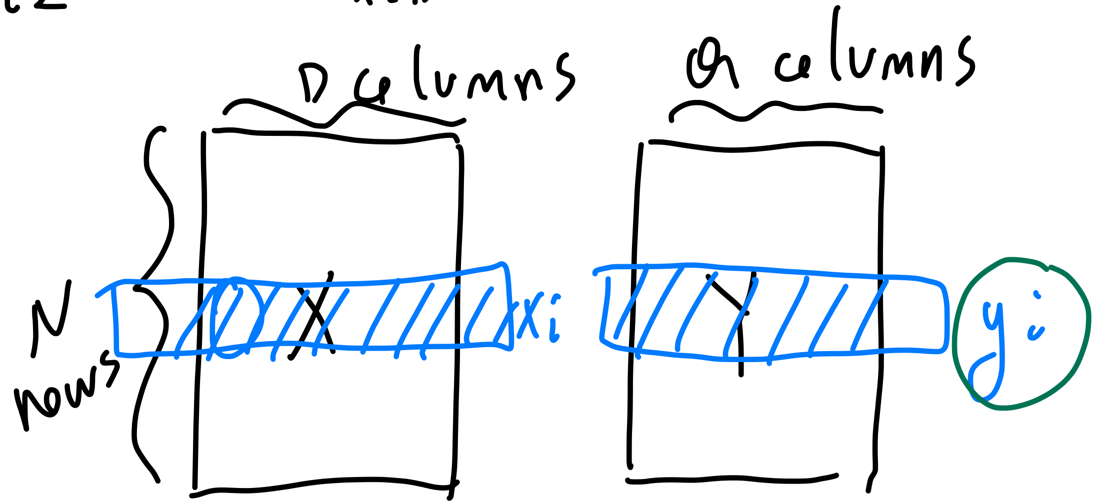
← input layer.
(sensors)
(receptors)

$q \times (D+1)$

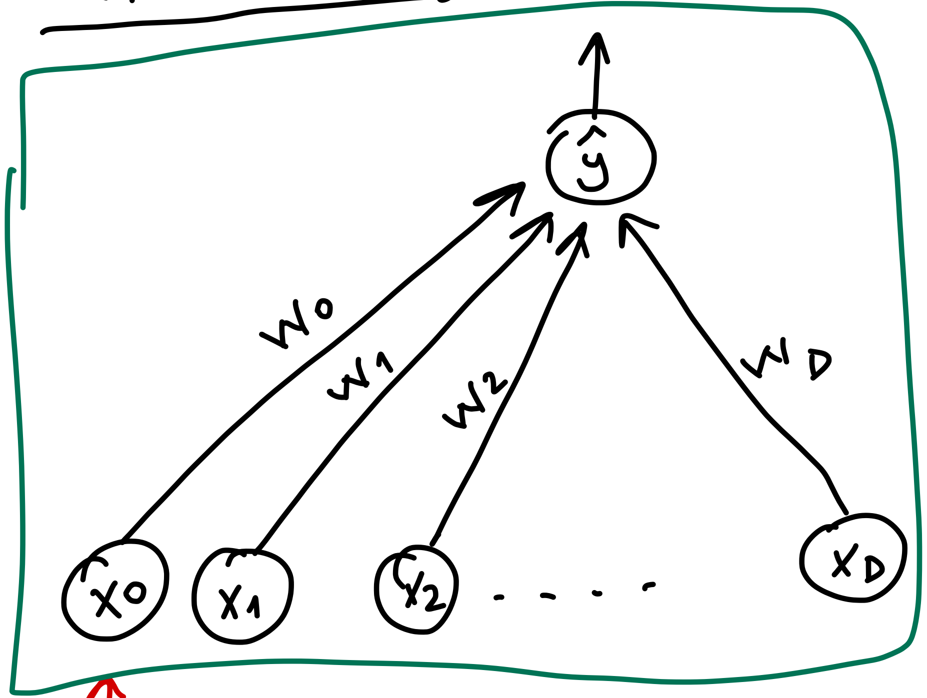
$W \in \mathbb{R}$

$\Downarrow$

$W \in \mathbb{R}^{(D+1)\times q}$



Linear Regression Model.



Use_bias

$$\hat{y}_i = w_1 \cdot x_{i1} + w_2 \cdot x_{i2} + \dots + w_D x_{iD} + w_0$$

$$\boxed{\begin{aligned}
\hat{y}_1 &= w_0 + w_1 x_{11} + \dots + w_D x_{1D} \\
\hat{y}_2 &= w_0 + w_1 x_{21} + \dots + w_D x_{2D} \\
&\vdots \\
\hat{y}_N &= w_0 + w_1 x_{N1} + \dots + w_D x_{ND}
\end{aligned}}$$

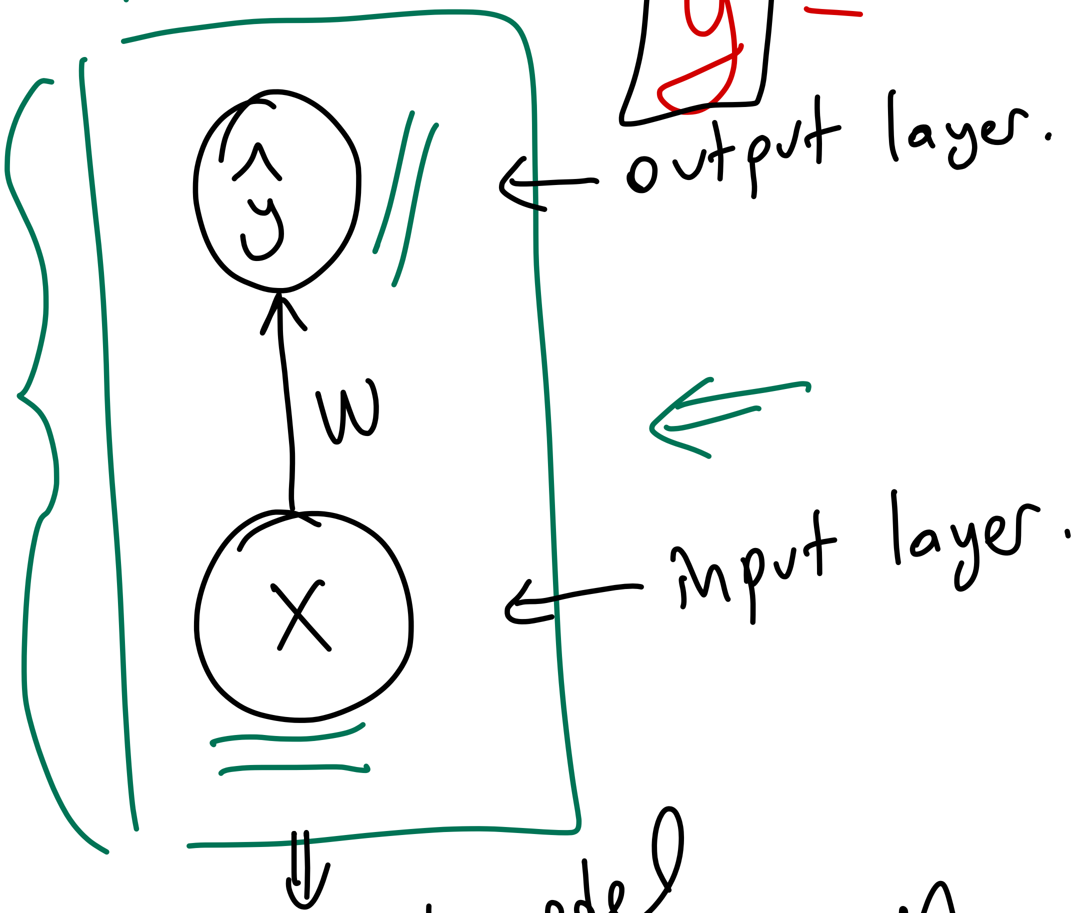
X_0 $\quad X$

$$\underbrace{\begin{bmatrix}\hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N\end{bmatrix}}_{N \times 1} = \underbrace{\begin{bmatrix}1 & x_{11} & \cdots & x_{1D} \\ 1 & x_{21} & \cdots & x_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{ND}\end{bmatrix}}_{N \times (D+1)} \underbrace{\begin{bmatrix}w_0 \\ w_1 \\ \vdots \\ w_D\end{bmatrix}}_{(D+1) \times 1}$$

W

$N \times D$

Model.



Abstract model that describes relationship between X & y.

$$\hat{y} = \tilde{X} . \tilde{W} \equiv \sigma(X . W)$$

activation function

$$\sigma(x) = x$$

$$\text{loss}(y, \hat{y})$$

$y \rightarrow N \times 1$

$\hat{y} \rightarrow N \times 1$

$$\Downarrow$$

$$(y-\hat{y})^T (y-\hat{y})$$

Squared error $\Rightarrow \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \boxed{(y-\hat{y})^T(y-\hat{y})}$

~~for loop~~

$$[y_1-\hat{y}_1 \; \cdots \; y_N-\hat{y}_N]_{1\times N} \begin{bmatrix} y_1-\hat{y}_1 \\ \vdots \\ y_N-\hat{y}_N \end{bmatrix}_{N\times 1}$$

$$(y_1-\hat{y}_1)^2 + \cdots + (y_N-\hat{y}_N)^2$$

minimize $(y-\hat{y})^T\cdot(y-\hat{y})$ given X, y

minimize $(\overset{\checkmark}{y} - \overset{\checkmark}{X}\cdot\overset{?}{w})^T\cdot(\overset{\checkmark}{y} - \overset{\checkmark}{X}\overset{?}{w}) \Leftarrow \underline{f(w)}$

minimize $\cancel{\sum_{i=1}^{N} \left(y_i - w^T x_i - w_0 \right)^2}$

$\boxed{\text{minimize } f(w)}$

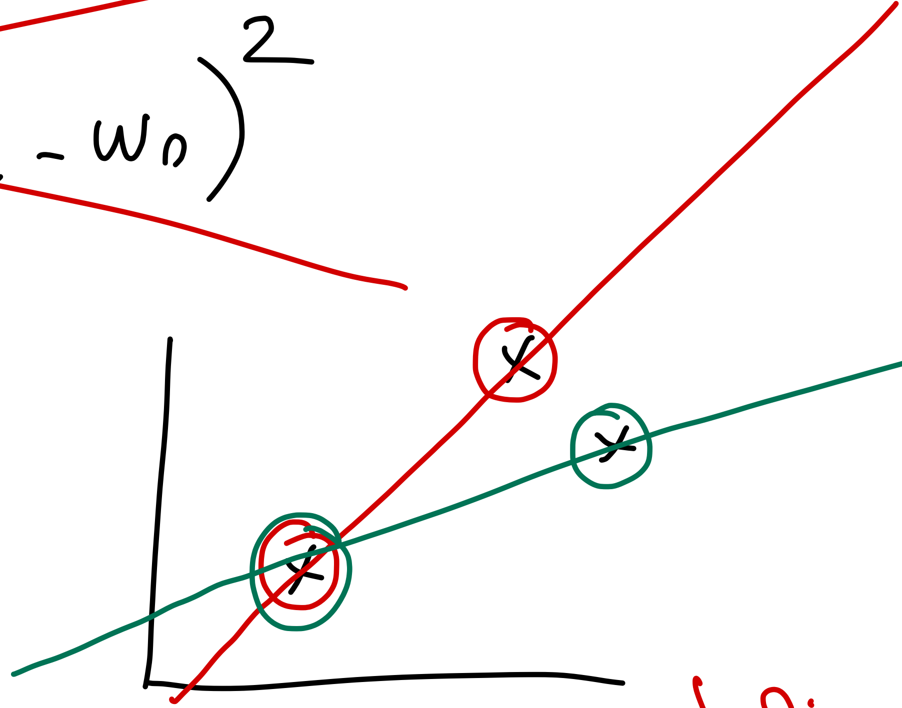
training data. ?

loss function ✓

network ✓

randomness

if we use random initialization.



$$\text{minimize} \quad \underbrace{(y - X.w)^T}_{g(w)} . \underbrace{(y - X.w)}_{h(w)}$$

$$\text{minimize } f(w)$$

$$\Downarrow$$

$$\frac{\partial f(w)}{\partial w} = 0$$

$$\frac{\partial^2 f(w)}{\partial w^2}$$

Convex

concave

$$\frac{\partial\, g(w).h(w)}{\partial w} = \overbrace{\frac{\partial g(w)}{\partial w}}^{D \times 1} . h(w) + g(w) . \frac{\partial h(w)}{\partial w}$$

$$= \underbrace{-\overset{D \times N}{X^T} . \overset{N \times 1}{(y - X.w)}}_{D \times 1} + \underbrace{-\overset{D \times N}{X^T} \overset{N \times 1}{(y - Xw)}}_{D \times 1}$$

$$(a^T . b)^T = b^T . a$$

$$(A . B)^T = B^T . A^T$$

p.s.d.

$$\partial X^T . X$$

$$= -2 . X^T . (y - X.w)$$

$$X^T \cdot (y - Xw) = 0$$

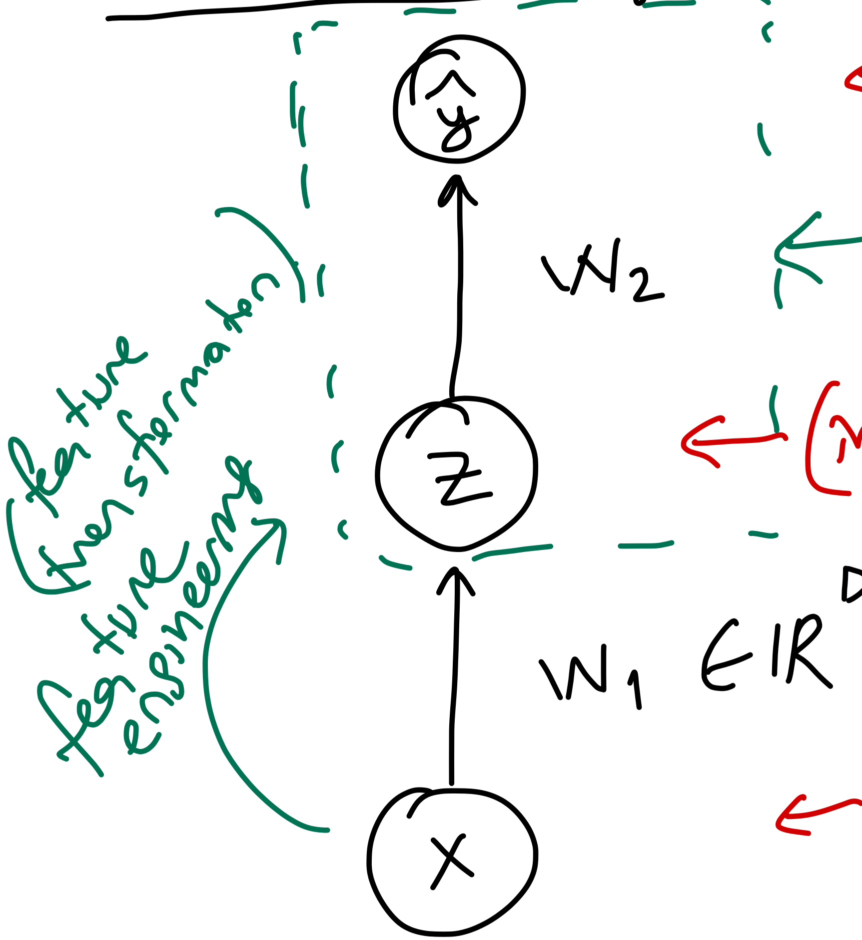
$$(X^T \cdot X)^{-1} X^T \cdot y = (X^T \cdot X)^{-1} (X^T \cdot X) w$$

$$\left((X^T \cdot X)^{-1} \cdot X^T \cdot y \right) = I \cdot w \Rightarrow \boxed{\underline{\underline{w}} = (X^T \cdot X)^{-1} \cdot X^T \cdot y}$$

closed-form solution.

global optimum.

NonLinear Regression

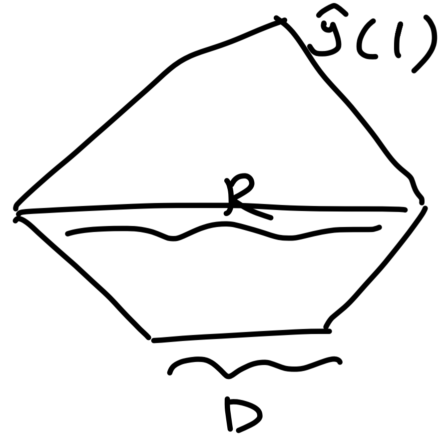


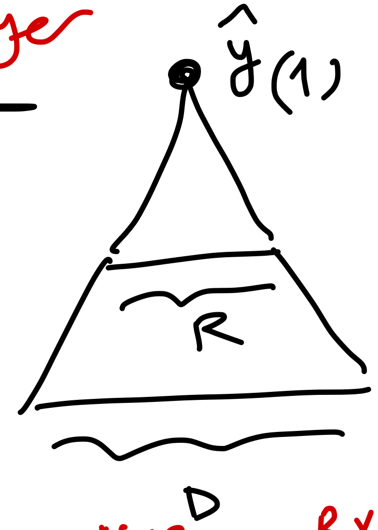
← output layer.

← linear regression.

← (intermediate) hidden layer

← input layer





$$\underbrace{z}_{N \times R} = \sigma(\underbrace{X}_{N \times D} \cdot \underbrace{W_1}_{D \times R})$$

$$\underbrace{\hat{y}}_{N \times 1} = \underbrace{Z}_{N \times R} \cdot \underbrace{w_2}_{R \times 1}$$

if $R < D$, we can consider this as dimensionality reduction.

if $\sigma(\cdot)$ is unit activation $\boxed{\sigma(x) = x}$

$$Z = X \cdot W_1 \qquad \hat{y} = Z \cdot W_2$$

$$\hat{y} = \underset{N \times D}{X} \cdot (\underset{D \times R}{W_1} \cdot \underset{R \times 1}{W_2}) = \boxed{\underset{N \times D}{X} \cdot \underset{D \times 1}{W}}$$

$$\hat{y} = \left[\sigma(X \cdot W_1)\right] \cdot W_2$$

$\sigma(\ldots) \Rightarrow$ $[0,1]$ sigmoid $\longrightarrow \sigma(a) = \dfrac{1}{1+\exp(-a)}$

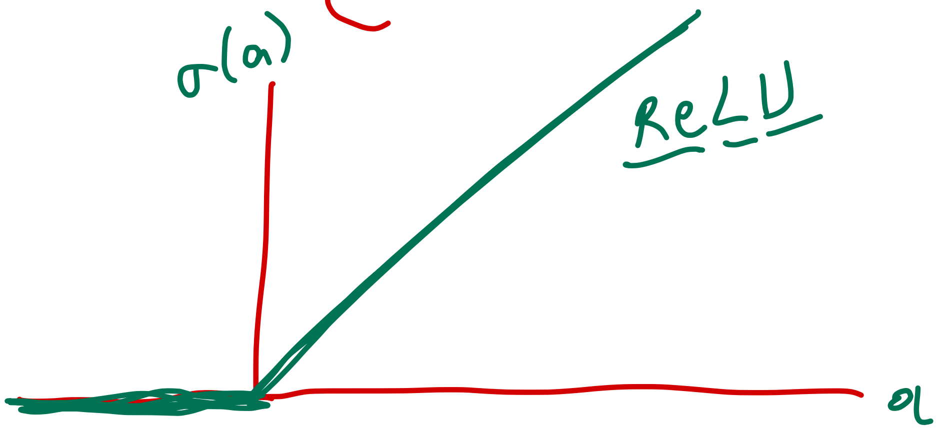
$[-1,+1]$ tanh $\longrightarrow \sigma(a) = \dfrac{\exp(a)-\exp(-a)}{\exp(a)+\exp(-a)}$

ReLU

Leaky ReLU

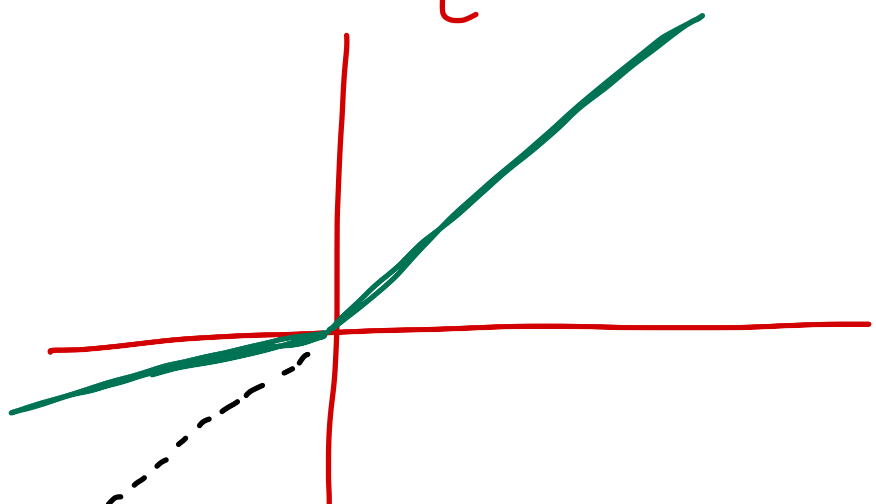
$\vdots$

$$\sigma(a) = \begin{cases} a & a \geq 0 \\ 0 & a < 0 \end{cases}$$



$$\sigma(a) = \begin{cases} a & a \geq 0 \\ \alpha a & a < 0 \end{cases}$$

$\alpha < 1$



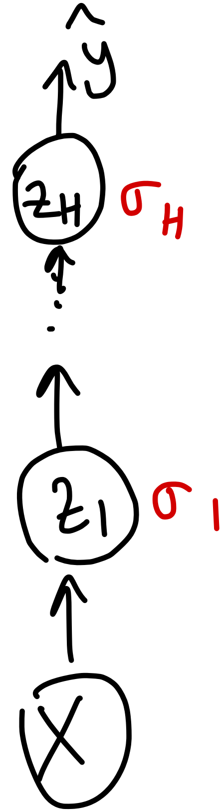
minimize $(y-\hat{y})^T \cdot (y-\hat{y})$

$\hat{y} = X \cdot w$

$\hat{y} = \sigma(X \cdot W_1) \cdot w_2$

minimize $\left(y - \sigma(X \cdot W_1) \cdot w_2\right)^T \cdot \left(y - \sigma(X \cdot W_1) \cdot w_2\right)$

There is no closed-form solution.



$$z_1 = \sigma_1(X \cdot W_1)$$

$$z_2 = \sigma_2(z_1 \cdot W_2)$$

$$\vdots$$

$$z_H = \sigma_H(z_{H-1} \cdot W_H)$$

$$\hat{y} = z_H \cdot w_{H+1}$$

if all of them one unit act.

$$\hat{y} = \sigma_H(\sigma_{H-1}(\sigma_{H-2}(\cdots \sigma_1(X \cdot W_1)W_2)W_3 \ldots)W_H) \cdot w_{H+1}$$

$$X(\underbrace{W_1 \cdot W_2 \cdot W_3 \cdot \ldots W_H \cdot W_{H+1}}_{w})$$

$$f(x) = 2x^2 + 5$$

$$g(x) = 2x^3 - 2x^2 + 7$$

$$h(x) = \log(x)$$

$$f \circ g \circ h(x)$$

$$=$$

Linear Classification

$$X \xrightarrow{w} \hat{y}$$

sigmoid.

$$\sigma(a) = \frac{1}{1+\exp(-a)}$$

$$\hat{y} \neq X.w$$

$$\hat{y} = \sigma(X.w)$$

$\Downarrow$

logistic regression.

NonLinear Classification.

$$X \xrightarrow{W_1} z \xrightarrow{w_2} \hat{y}$$

$$z = \sigma_1(X.W_1)$$

$$\hat{y} = \sigma_2(z.w_2)$$

always sigmoid if binary classification

$$\hat{y} = \sigma_2(\sigma_1(X.w_1).w_2)$$

sigmoid, tanh, ReLU, Leaky ReLU.

Problem	Act. at the output layer
Regression $\rightarrow$	Unit activation.
Binary Class. $\rightarrow$	Sigmoid.
Multiclass Class. $\rightarrow$	Softmax
Multilabel Class. $\rightarrow$	multiple sigmoids
Multioutput Regres. $\rightarrow$	Multiple unit. actv.

loss$(y, \hat{y})$ → for binary classification

└→ most frequently used loss is binary cross entropy.

as Bernolli distributed R.Vs.

maximize log(likelihood)

⇓

minimize −log(likelihood.)

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

1   0

0   1

ratio of + = 98%

ratio of − = 2%

A = 1

B = 49

class_weight

$f(w) \rightarrow$ should be differentiable.