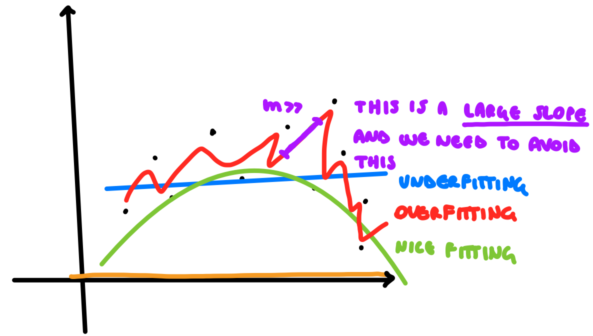
REGULARIZATION

7/17/2023

IN PARAMETRIC MODELS, θ IS EXTREMELY DATA DEPENDENT AND THIS CAN BE A PROBLEM
THANKS TO REGULARIZATION, WE START FROM AN UNDERFITTING PREDICTOR h_θ AND OBTAIN SOMETHING OPTIMAL



RECALL THE LOSS FUNCTION

$$J(\theta) = \underbrace{\sum_{i=1}^{n}\left(y^{(i)} - \theta^T x^{(i)}\right)^2}_{1} + \underbrace{\lambda \; REG}_{2}$$

1. ACTUAL SQUARED LOSS, $\nabla_\theta(1)$ AND $\arg\min_{\hat{\theta}}$ MAKES THE MODEL OVERFITTING

IN A CERTAIN SENSE, WE NEED TO MAKE $J(\theta)$ WORSE BY A TERM REGULARIZATION

$$\boxed{REG = \lambda \underbrace{\|\theta\|_2^2}_{\substack{L_2 \text{ NORM} \\ \sum_i \theta_i^2}}, \quad \lambda \underbrace{\|\theta\|}_{\substack{L_1 \text{ NORM} \\ \sum_i |\theta_i|}}}$$

$\|\theta\| >> \longrightarrow$ COST $>>$

BAYESIAN INTERPRETATION OF REGULARIZATION AND M.A.P.

2 APPROACHES

- FREQUENTISTIC : MAX LIKELIHOOD, COMPLETELY DATA DEPENDENT

$$\hat{\ell}(\theta) = \log \prod_{i=1}^{n} P(y^{(i)} | x^{(i)}, \theta) \rightarrow \hat{\theta} = \arg\max_{\theta} \ell(\theta) \rightarrow \text{PREDICTION}$$

(IN LOG. REG AND LIN. REG IF THE MEAN OF THE DISTRIBUTION IS 0, THEN MIN(COST) = MAX(LIKEL.)

- BAYESIAN : PRIOR $P(\theta)$ → DATA OBSERVATION → POSTERIOR $P(\theta | y, x)$ → BAYES

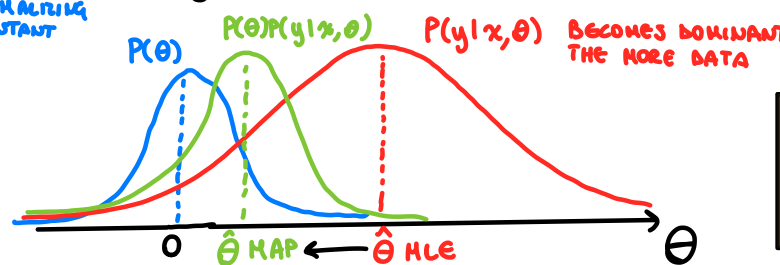
$$P(\theta | y, x) = \frac{P(\theta) P(y | x, \theta)}{\int P(\theta) P(y | x, \theta)} \rightarrow \text{POSTERIOR PREDICTIVE DISTRIBUTION}$$

M.A.P. ESTIMATION, NEW APPROACH! (MAXIMUM A POSTERIOR EST.)
IT'S A COMPROMISE BETWEEN THE 2 PREVIOUSLY MENTIONED APPROACHES
CONSISTS OF CALCULATING θ POINT BY MAXIMIZING THE POSTERIOR DISTRIBUTION

$$\left(\sum \log P(y^i | x^i, \theta)\right)$$

$$P(\theta | x, y) = \frac{P(\theta)\, P(y | x, \theta)}{\underbrace{K}_{\text{NORMALIZING CONSTANT}}} \rightarrow \arg\max_{\hat{\theta}} P(\theta | x, y) = \arg\max \underbrace{P(\theta)}_{\text{PRIOR}} \underbrace{P(y | x, \theta)}_{\text{LIKELYHOOD}} = \arg\max \log P(\theta) + \log P(y | x, \theta)$$

BY NORMALIZING BY THE PRIOR IS AS IF WE WERE PULLING $\hat{\theta}_{MLE}$ TOWARDS Ø



THE CHOICE OF $P(\theta)$ DETERMINES THE REGULARIZATION (ex $L_1, L_2, \ldots$)

NEURAL NETWORKS AND DEEP LEARNING

UNTIL NOW, WE COULD INTRODUCE NON-LINEARITY IN OUR MODELS THANKS TO <u>FEATURE MAPS</u>

$$\bar{y} = h_\theta(x) = \theta^T \phi(x) \qquad \left(\phi : \mathbb{R}^d \longmapsto \mathbb{R}^p \right)$$

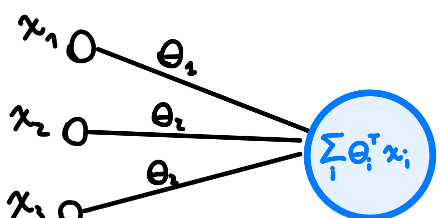
NON-LINEAR!

AND WE CAN SEE THIS AS ADDING "WIDTH" TO THE MODEL. BUT ANOTHER WAY TO INTRODUCE NON-LINEARITY IS <u>DEEP LEARNING</u>, BY INTRODUCING <u>DEPTH</u> TO THE MODEL

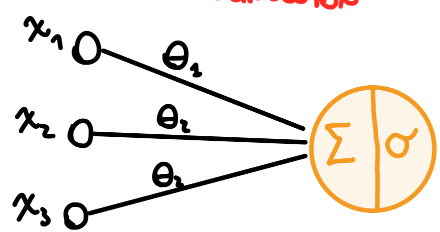
<u>ex</u>

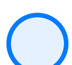
$$x \in \mathbb{R}^3 \longrightarrow \begin{cases} x_1 : \text{SIZE} \\ x_2 : \text{n BEDS} \\ x_3 : \text{ZIPCODE} \end{cases}$$

LINEAR REGRESSION



LOGISTIC REGRESSION



 AND  ARE CALLED NEURONS!

ex NEURAL NETWORK

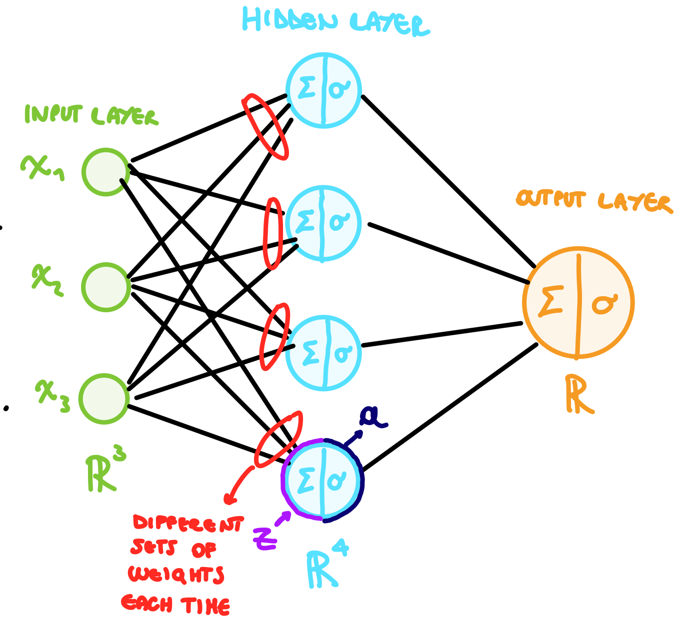
INPUT LAYER TAKES INPUT $x \in \mathbb{R}^d$

HIDDEN LAYER(S) PERFORMS LINEAR COMBINATION, APPLIES NON-LINEARITY

$z \rightarrow \bigcirc \rightarrow a$

USUALLY PERFORM THE SAME OPER. BUT WITH DIFFERENT PARAMETERS

OUTPUT LAYER OUTPUTS $\hat{y}$



WE CAN WORK WITH THE ARCHITECTURE BY PLAYING AROUND :

- N OF NEURONS
- OPERATIONS
- NON-LINEARITY
- OTHER CHOICES

TERMINOLOGY

- INPUTS : $x \in \mathbb{R}^d$
- $a_i^{[l]}$, l : LAYER , $i = 1, \dots, n$ → THING COMING OUT A HD NEURON, TRANSFORMED SCALAR (VECTOR
  FOR INSTANCE, WE COULD DEFINE INPUTS x SUCH AS $x = a^{[0]} \; (x_1 = a_1^{[0]}, x_2 = a_2^{[0]}, \dots)$
- $z^{[l]}$, l : LAYER → THING GOING INSIDE A LAYER , IS A LINEAR COMBINATION (SCALAR) DEFINED AS FOLLOWS

$$z^{[l]} = \sum_i a_i^{[l-1]} \theta_i + b_i$$

$$z^{[l]} = \underbrace{W^{[l]}}_{\text{WEIGHT MATRIX}} a^{[l-1]} + \underbrace{\vec{b}}_{\text{BIAS}}$$

N OF NEURONS ON PREV. LAYER

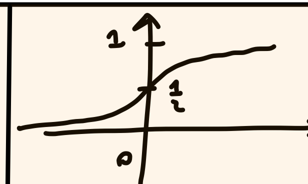
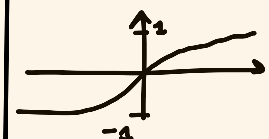
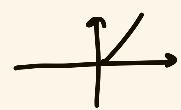
$$z = \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix} \updownarrow \text{N OF NEURONS CURR. LAYER} \times \begin{bmatrix} \\ \\ \\ \end{bmatrix} \updownarrow \text{NN PREV. LAYER} \Rightarrow \begin{bmatrix} \\ \\ \\ \end{bmatrix} \updownarrow \text{CURR} + \overset{b}{\begin{bmatrix} \\ \\ \\ \end{bmatrix}} \updownarrow \text{CURR}$$

$W^{1 \; ? \times d}$, $b^{1 \; ?}$

THEN WHAT HAPPENS IS

$$\boxed{z^{[l]} \longrightarrow \sigma \text{ (NON-LINEARITY)} \longrightarrow a_i^{[l]}}$$

THE MOST COMMON NON-LINEARITY FUNCTION ARE :

SIGMOID	$\sigma(z) = \dfrac{1}{1+e^{-z}}$	
TANH	$\tanh(z) = \dfrac{e^{z}-e^{-z}}{e^{z}+e^{-z}}$	
RELU	$\mathrm{ReLU}(z) = \max\{z, 0\}$	

SO, TO RECAP

$$\boxed{a^{[l]} = \sigma(z^{[l]})}$$

AND, FINALLY

$$\boxed{\hat{y} = a^{[L]}}$$