Q7 In conventional ~~m~~ ML models involving linear regression, the "hypothesis" is a linear/affine function, i.e of the form.

$$z = W^T X + b.$$

An activation layer introduces non-linearity in the model, e.g.

$$a = \sigma(z) = \sigma(W^T X + b)$$

where $\sigma(x) = \dfrac{1}{1+e^{-x}}$.

Mathematical Requirements

(1) Non-linearity $\quad a(\lambda z_1 + \mu z_2) \neq \lambda a(z_1) + \mu a(z_2)$

(2) Differentiability

Although ReLU is not diff'ble, it is extensively used. At 0, the issue is handled by employing subgradient.

(3) Boundedness $\exists M > 0$ s.t. $|a(x)| \leq M \ \forall x \in \mathbb{R}$.

e.g σ, tanh etc.

However, there are activation fns that are unbdd.

e.g ReLU.

Date
Page

Some of the Types

(1) Sigmoid. $\sigma(x) = \frac{1}{1+e^{-x}}$

→ Used for binary classification problems.

→ $\sigma \in (0, 1)$ ~~can be~~ ~~returns~~

→ $\sigma(x)$ can be interpreted as probability.

(2) ReLU

$$\text{ReLU}(x) = \max(0, x) = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

→ Used to avoid vanishing gradients
→ Used in CNNs almost always [other alternatives → GELU, Leaky ReLU]

(3) Softmax

$$[\text{softmax}(z)]_i = \frac{e^{z_i}}{\sum_{j=1}^{c} e^{z_j}}$$

c → no. of classes

$$z = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_c \end{bmatrix}$$

→ Used for multi-class classification problems

(4) Leaky ReLU

$f(x) = \max\{\alpha x, x\}$, where $\alpha > 0$ is fixed.

→ Used in CNNs

→ Avoids dying ReLU issue, i.e. $f'(x) = \alpha \neq 0$

[as opposed to $\frac{d}{dx}\mathrm{ReLU}(x) = 0 \;\forall x < 0$]