COL865
Deep Learning
July 31, 2017

Last Class:-

(I) Motivating Applications

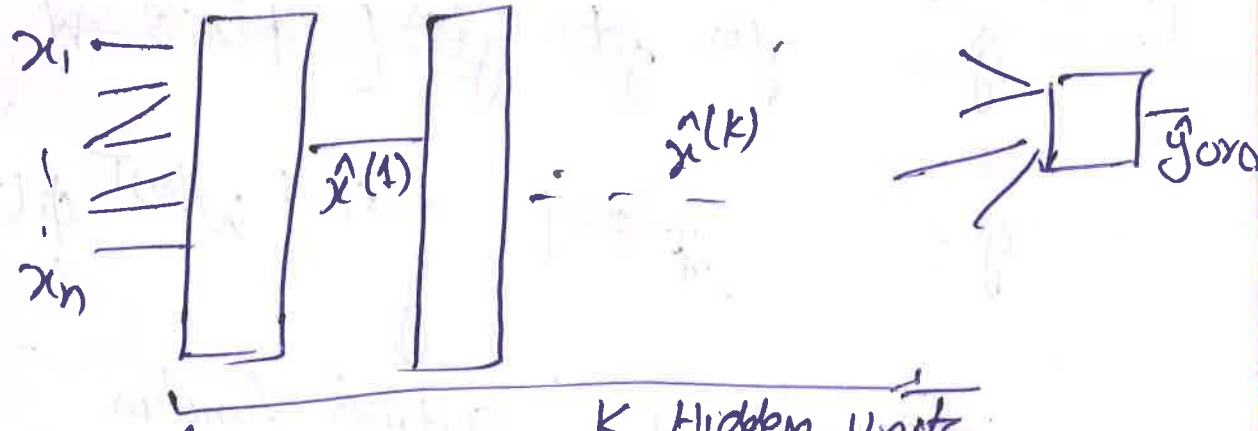
(II) Standard Architectures
↳ 5 of them
CNNs, RNNs, Autoencoders, GANs, DeepRL

(III) Implementation:- Pytorch

(IV) Feedforward Networks

L:- # of Layers

Notation:



m:- # of examples

n:- Dimensionality of x

$x^{(0)} \equiv x$

$n^{(k)}$:- Dimensionality of output coming out of layer k.

$W^{(k)} \in \mathbb{R}^{n^{(k-1)} \times n^{(k)}}$ $(k \geq 1)$

$W^{(k)}$ ⇓ Weights in layer K.

$\theta = (\theta^{(1)} \cdots \theta^{(k)})$

Feature representation $\phi(x; \theta)$

$\theta^{(0)}$

$\theta = (\theta, \theta^{(0)})$

$\phi = \phi^{(k)}(\phi^{(k-1)} \cdots \phi^{(1)}(x, \theta^{(1)}) \cdots)$

$f^*(x;\theta) = \hat{y}$

~~$\hat{y} \in \mathbb{R}^{(0)}$~~ ⇒ ~~output layer~~

$\hat{y} \in \mathbb{R}\{0, \ldots L\}$

⇓

# of classes

$y = f^{(0)}(\phi(x;\theta); \theta^{(0)})$

$\hat{x}^{(k+1)} = g^{(k)}$ ⇒ Elementwise Application

$\hat{x}^{(k)} = g^{(k)}[\ W^{(k)T}\hat{x}^{(k-1)} + b^{(k)}]$

⇓

ReLU

$h^{(k)}(x) = g^{(k)}(W^{(k)T}x + b^{(k)})$ Rectified Linear Unit

~~$\hat{y} = g^{(0)}\ g^{(1)}\ g^{(0)}(\phi(x;\theta\ W$~~

$\hat{y} = g^{(0)}(W^{(0)T}\phi(x;\theta) + b^{(0)})$

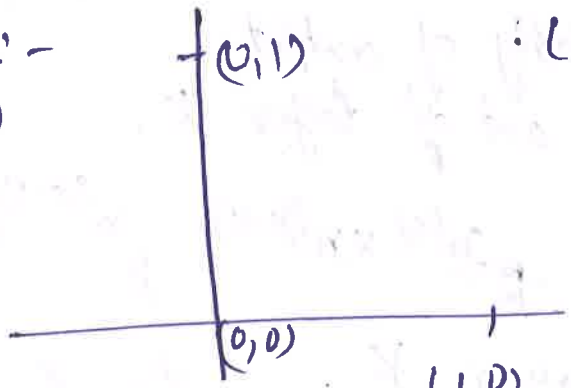
⇓

Logistic / Softmax function

Now:- Example (Non-Linearity).

XOR:-

$g(x) = \max(0, x)$



Not linearly separable

Transform:- ReLU

$g(W^Tx + b) = W = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$

$b = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$,

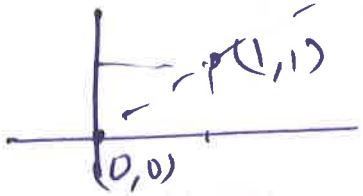
(2)

$$X = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \qquad W = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

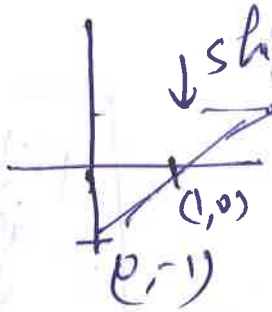
$$W^T X + b = \begin{bmatrix} 0 & 1 & 1 & 2 \\ 0 & 1 & 1 & 2 \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

(-ve, +ve, +ve, -ve)

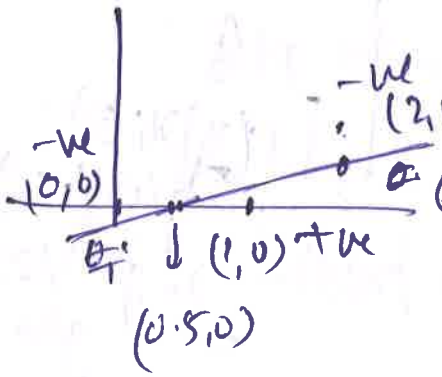
(2,2)



$$\Rightarrow \begin{bmatrix} 0 & 1 & 1 & 2 \\ -1 & 0 & 0 & 1 \end{bmatrix}$$



$$g(W^T X + b) = \begin{bmatrix} 0 & 1 & 1 & 2 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$



$\Rightarrow$ linearly separable

After the transformation

Assume:- $y \in \{0,1\}$

Output Unit :-

$y = f^{(0)}(x;\theta) = g^{(0)}\left(w^{(0)T}\phi(x;\theta) + b^{(0)}\right)$

what is its form?

① $g^{(0)} =$ ~~Identity~~ identity

$f^{(0)}(x,\theta) = w^{(0)T}\phi(x;\theta) + b^{(0)}$

Not a good idea for 0/1 loss

~~Understand the loss function:~~

~~$L(y, \hat{y}; \theta)$~~

part of input or can be learned. (6.2.4)

$\phi(x; \theta) \equiv \hat{x}^{(k)}$

Assume $y|x \sim \mathcal{N}(\mu, \sigma^2)$, $\mu = \hat{y}$

① $y \in \mathbb{R}$.

$g^{(o)} \equiv$ Identity

$\hat{y}^{(o)} = \phi(x; \theta), \theta) = w^{(o)T} \phi(x; \theta) + b^{(o)}$

Linear Output Unit

② $y \in \{0, 1\}$ ~~Bernoul~~ $y \sim$ Bernoulli(ϕ)

$g^{(o)} \equiv$ ~~identity~~

$\neq g^{(o)} =$ sigmoid

$g^{(o)}(z) = \frac{1}{1+e^{-z}}$

$P(y|x;\theta) = g^{(o)}((2y-1)z)$

~~$g^{(o)}(w^{(o)T}$~~

$z = w^{(o)T} \hat{x}^{(k)} + b^{(o)}$

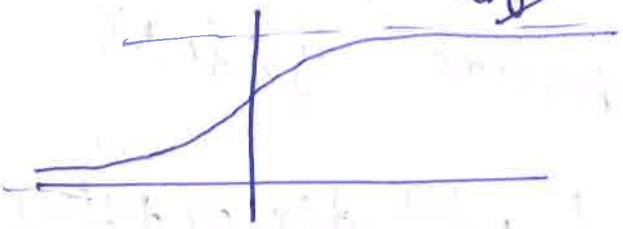
PT

$g^{(o)}(z) = \frac{1}{1+e^{-z}}$ ⟹ P ~~ŷ~~

$P(y|x;\theta) = g^{(o)}(z)$

$P(y=0|x;\theta) = 1 - g^{(o)}(z) = g^{(o)}(-z)$

logit



③ ~~$y \in \{0, \dots, L\}$~~ $y \in \{1 \dots L\}$

$g^{(o)} \equiv$ Softmax function $z = w^{(o)T} \hat{x}^{(k)} + b^{(o)}$

↓ Vector

③

$$g^{(0)}(z)_k = \frac{e^{z_k}}{\sum_{c=1}^{K} e^{z_c}}$$

softmax function (generalization of logistic).

Defines a multinoulli distribution.

over $y \in \{0, 1, \ldots, K\}$.

$$\boxed{P(y = k \mid x; \theta) = g^{(0)}(z_k)}$$

Now, what is the right loss function?

$$L(y, \hat{y}, \theta) = ?$$

$$= \|y - f(x, \theta)\|^2$$

squared difference

$$J(\theta) = \sum_{x \in X} \|y - f(x, \theta)\|^2$$

X ↓ All instances

$$\hat{J}(\theta) = \sum_{\{x^{(i)}, y^{(i)}\}_{i=1}^{m}} \|y^{(i)} - f(x^{(i)}, \theta)\|^2$$

⇓ Not a good cost function

$g(\hat{x}^{(k)})$

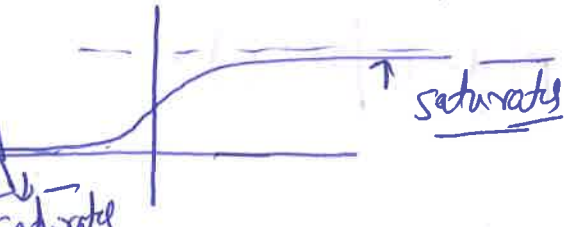
$g(w^{(0)T} \hat{x}^{(k)} + b^{(0)})$

$\nabla_{w^{(0)}} g(w^{(0)T} \hat{x}^{(k)} + b^{(0)})$

$= g(w^{(0)T} \hat{x}^{(k)} + b^{(0)})$

$(1 - g(w^{(0)T} \hat{x}^{(k)} + b^{(0)}))$

$* \hat{x}^{(k)}$



$g'(z) = g(z)(1 - g(z)) \to 0$

as $z \to \infty$

$z \to -\infty$

Another idea

~~$J(\theta) =$~~ —

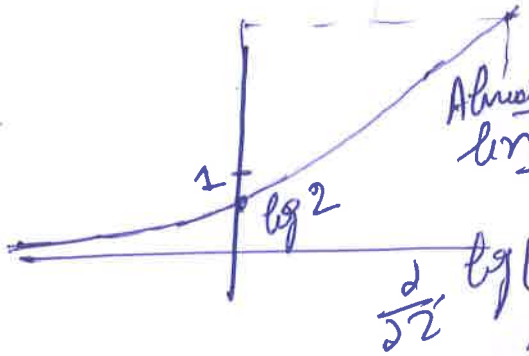
~~$L(y, \hat{y})$~~ $L(y, \vec{y}; \theta) = -\log P(y|x;\theta)$ $\quad z' = (1-2y)z$

Bernoulli:-

$= -\log [g(-z')]$

$= \log \left[\frac{1}{g(-z')}\right]$

$= \boxed{\log(1+e^{+z'})}$ ↓ softplus function



$z' >> 1$ $\quad \log(1+e^{z'}) \approx z'$

~~Does~~ Does not saturate when $z' \geq 0$

$J(\theta) = E_{(x,y)\sim Dist.} -\log[P(y|x;\theta)]$

$\hat{J}(\theta) = \frac{1}{m}\sum_{i=1}^{m} -\log P(y^{(i)}|x^{(i)};\theta)$ ~~y & z do not match~~

↳ $y>0$ & $z>0$] High gradient

$y=0$ & $z<0$

Good.

Multinoulli:-

$L(y, \vec{y}; \theta) = -\log P(y|x;\theta)$

$P(y=e|x;\theta) = \frac{e^{z_e}}{\sum_{e=1}^{L} e^{z_e}}$

$\log P(y=e|x;\theta) = z_e - \log\left[\sum_{e=1}^{L} e^{z_e}\right]$

~~if z_k is the label with highest z_e, then~~

(4) Note:-

if $e^* \in$ ~~Ze~~ s.t $2e^* >> 2e' \quad \forall e' \neq e$

then $\log \sum_{\ell=1}^{L} e^{Z_\ell} \approx Z_{e^*} + \log \left[\frac{e^{2e}}{e^{2e^*}} \right]$

$\overline{Ze^*} = 2e^* + \log \sum_{\ell=1}^{L} e^{(Ze - 2e^*)}$

$\Rightarrow$ if $\ell = e^* \quad = 2e^* + \log \sum_{\ell=1}^{L} (1 + e^{(Ze - 2e^*)})$

then:- $\log p(y=\ell \mid x, \theta)$ saturates. $\ell \neq e^*$ ↓

soft plus fn.

Saturates near 0.

Ex,

$-\log p(y=\ell \mid x, \theta) = -[Z_\ell - Z_{e^*}]$

$J(\theta) = \mathbb{E}_{(x,y) \sim \text{dist}} \; -\log p(y \mid x, \theta)$

$\hat{J}(\theta) = \frac{1}{m} \sum_{i=1}^{m} -\log \left[p(y^{(i)} \mid x^{(i)}, \theta) \right]$

Note:-

$\text{Softmax}(z) = \text{Softmax}(z - \max_i z_i)$

↓ Stable version

Finally, Gaussian

$y \sim \mathcal{N}(\hat{y}, \sigma^2)$ $\quad p(y \mid \hat{y}, \theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{\left[\frac{(y-\hat{y})^2}{2\sigma^2} \right]}$

$z = w^{(0)T} h^{(0)} + b^{(0)} = \hat{y}$

$-\log p(y \mid x, \theta) = C + (1/2) * \frac{\|\hat{y} - y\|^2}{2\sigma^2}$

squared error.

Summary:- Output unit

Output:- ① Logistic / Softmax / Identity

$$J(\theta) = \text{cost} - \sum_{i=1}^{n} \log p(y^{(i)} | x^{(i)}; \theta)$$

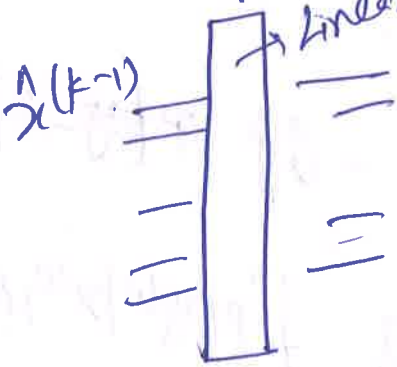
~~Hidden Units~~ - Book:- Mixture of components

Mixture density networks

More advanced! Later

Hidden Units:-

$W^{(k)}, b^{(k)}$ → Linear units followed by activation



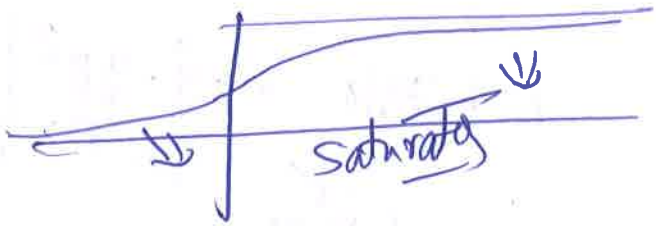
$$\theta^{(k)} = (W^{(k)}, b^{(k)})$$

$$h^{(k)} = g^{(k)}\left[W^{(k)T} h^{(k-1)} + b^{(k)} \right] = f(h^{(k-1)}; \theta^{(k)})$$

⇓

What is the right form for $g^{(k)}$?

① $g^{(k)}(z) \equiv \sigma(z) = \frac{1}{1+e^{-z}}$



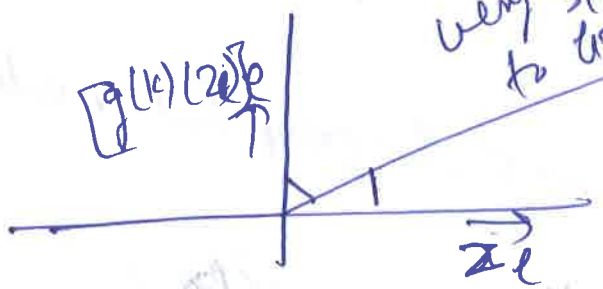
(5)

(1)

ReLu:- Rectified Linear Unit:-

$$z = W^{(k)T} \vec{x}^{(k-1)} + b^{(k)}$$

$$g^{(k)}(z)_i = \max(0, z_i)$$



$g^{(k)\prime}(z) = 1$ when the unit is active

Works very well in practice

$\frac{\partial}{\partial z_i}\left[g^{\prime(k)}(z)_i\right]_{z_i=0}$ is not defined

not an issue:-

$g''(z) = 0$ curvature.

(I) Typically may not reach exact minima (early stopping)

(II) Use sub-gradient.

(III) Initialize:- $b^{(k)}$'s to be slightly +ve numbers. ⇒ unit is active & ($w^{(k)}$'s to close to zero).

Variations:- $g^{(k)}(z, \alpha)_i = \max(0, z_i) + \alpha_i \min(0, z_i)$

when $\alpha = -1$, $g(z)_i = |z_i|$

Absolute value rectification

Features invariant to sign reversal

Leaky ReLu:-

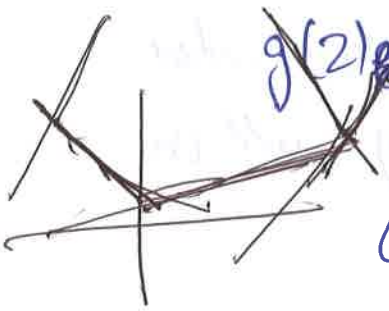
$\alpha_i = 0.01$

Parametric & Leaky ReLu:-

α_i - learnable

Date dependent non-linearities

Maxout Units:-



$g(z)_i = \max_{j \in G^{(i)}} z_j$ ⇒ Each group has r values

$G^{(i)} = \{ (i-1)*k + 1 \cdots \}$

Can learn a piecewise linear (convex) function with upto k pieces.

saturation issue

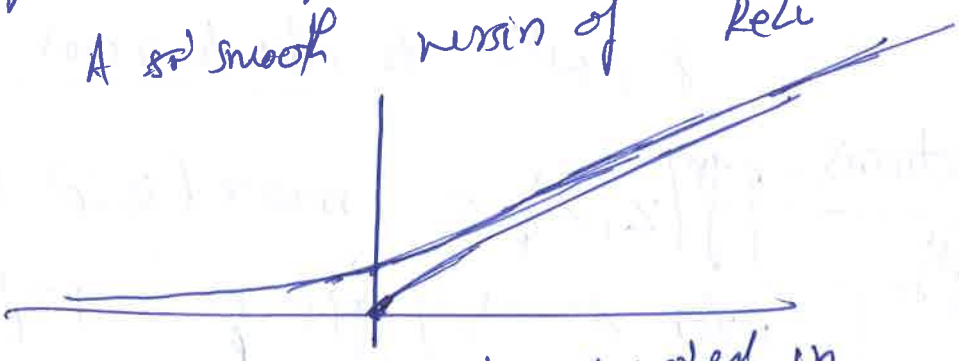
eg $\tanh(z) =$

$g(z)_i = \tanh(z_i) = \frac{e^{z_i} - e^{-z_i}}{e^{z_i} + e^{z_i}} = 2\sigma(2z_i) - 1$

Softplus:-

$g(z)_i = \log(1 + e^{z_i})$

A smooth version of ReLu



Some other units described in the book