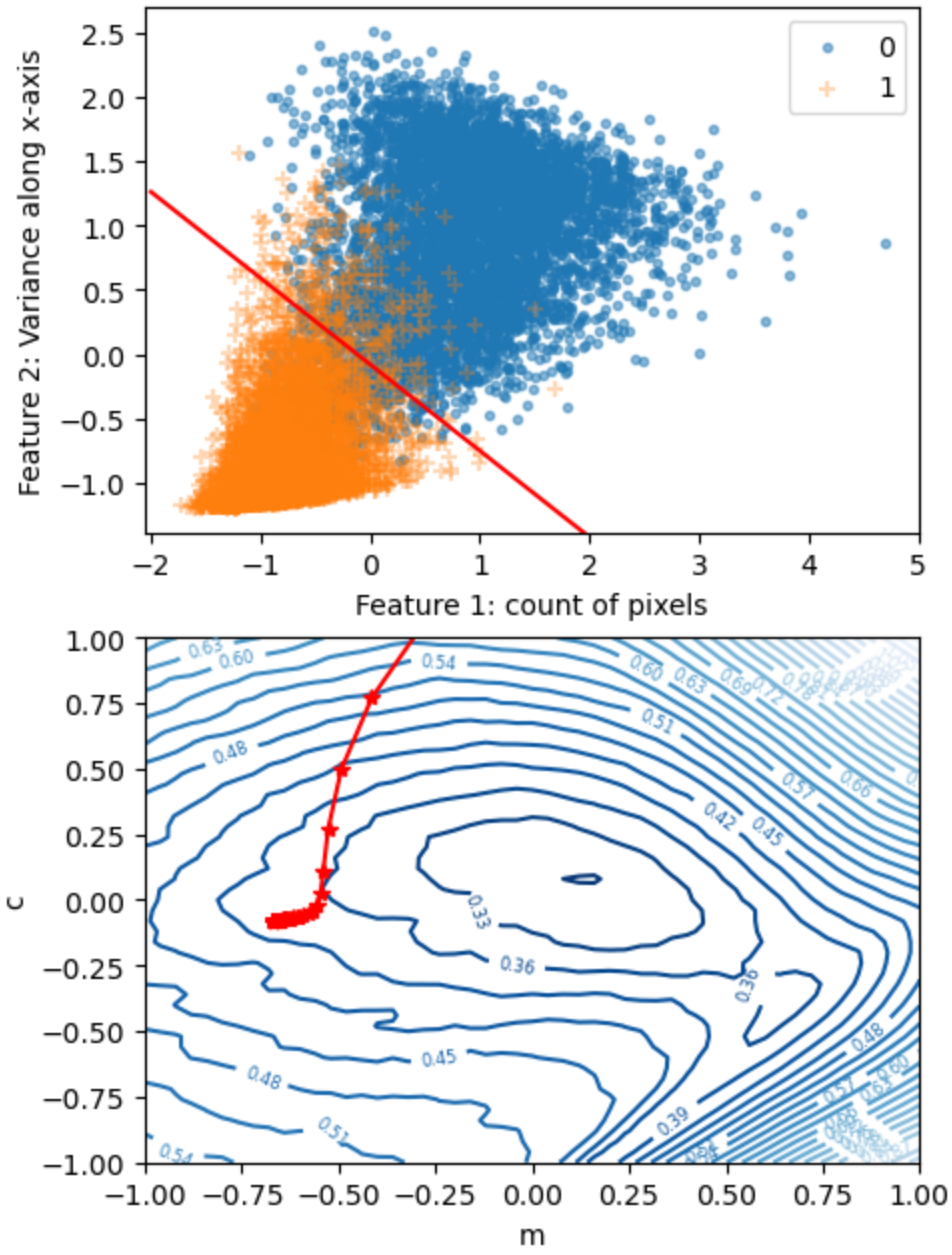
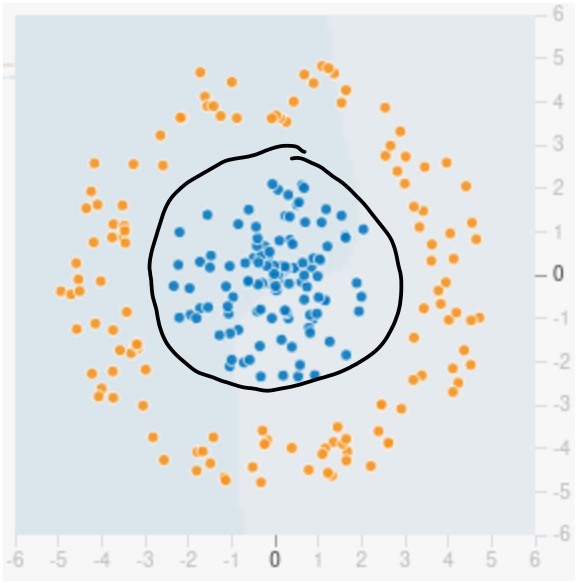
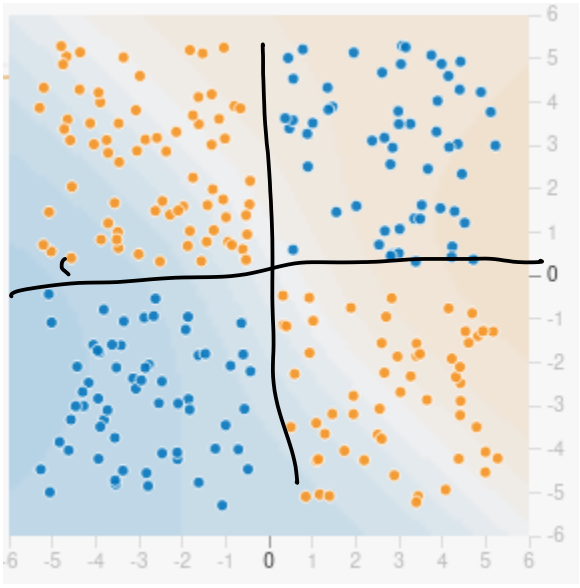
Multi layer Perceptron

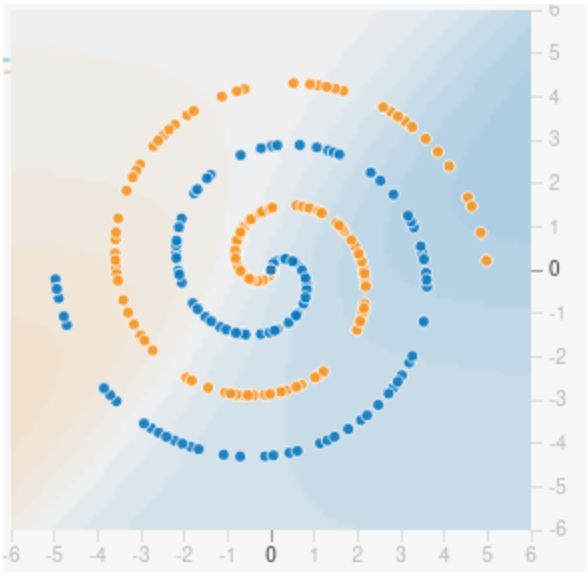
All figures are from Chapter 3 of UDLBook. https://github.com/udlbook/udlbook

Recall the single layer perceptron









Not linearly separable

Linear $(x) = l = f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + w_0$ (parameters, input)

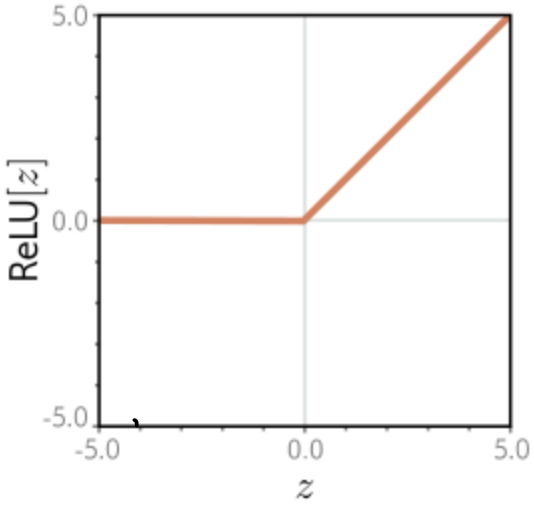
$y = m x + c$

$y = \underline{w}^T \underline{x} + w_0$ (weights, Bias)

Multi Layer Perceptrons

ReLU activation function $\mathrm{ReLU}(z) = \max\{0, z\}$

Rectified Linear unit



$\text{Linear}_1(\text{Linear}_2(\underline{x}))$

Non linear? in terms of x

$f(\underline{x}) = W_2\left(\underset{1\times n}{\underline{w}_1^T}\ \underset{n\times 1}{\underline{x}} + \underset{1\times 1}{w_0}\right) + w_{02}$

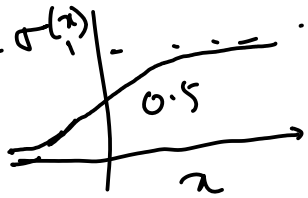
$f(\underline{x}) = \underset{1\times m}{\underline{w}_2^T}\left(\underset{m\times n}{W_1}\ \underset{n\times 1}{\underline{x}} + \underset{m\times 1}{\underline{w}_0}\right) + w_{02}$

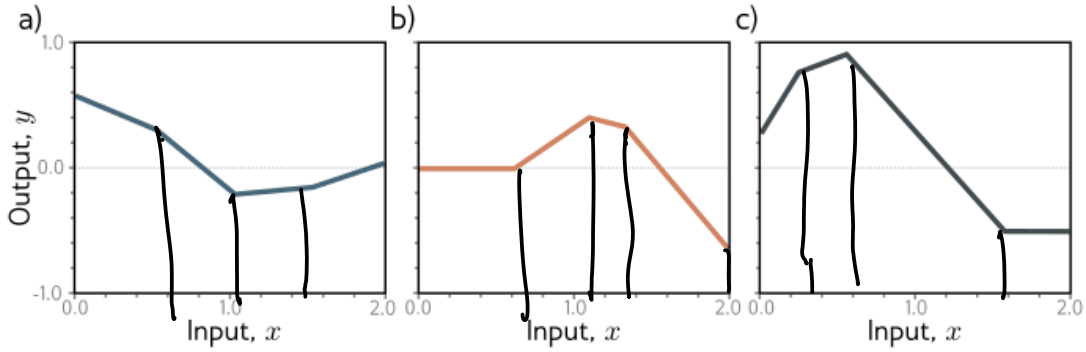
$= \underbrace{\underline{w}_2^T W_1}_{\underline{w}_3^T}\ \underline{x} + \underbrace{\underline{w}_2^T \underline{w}_0 + w_{02}}_{w_{03}}$

Two layer Perceptron

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

$$y = f(\mathbf{x}) = \text{Linear}(\ \underbrace{\text{ActivationFunction}}_{\text{Non-linear}}(\ \text{Linear}(x)))$$





with ReLU the model is Piecewise Linear function

Example

$a(x) = \mathrm{ReLU}(x)$

Activation/Hidden unit

$$\begin{aligned} h_1 &= a(\theta_{10} + \theta_{11}x) \\ h_2 &= a(\theta_{20} + \theta_{21}x) \end{aligned}$$ Linear

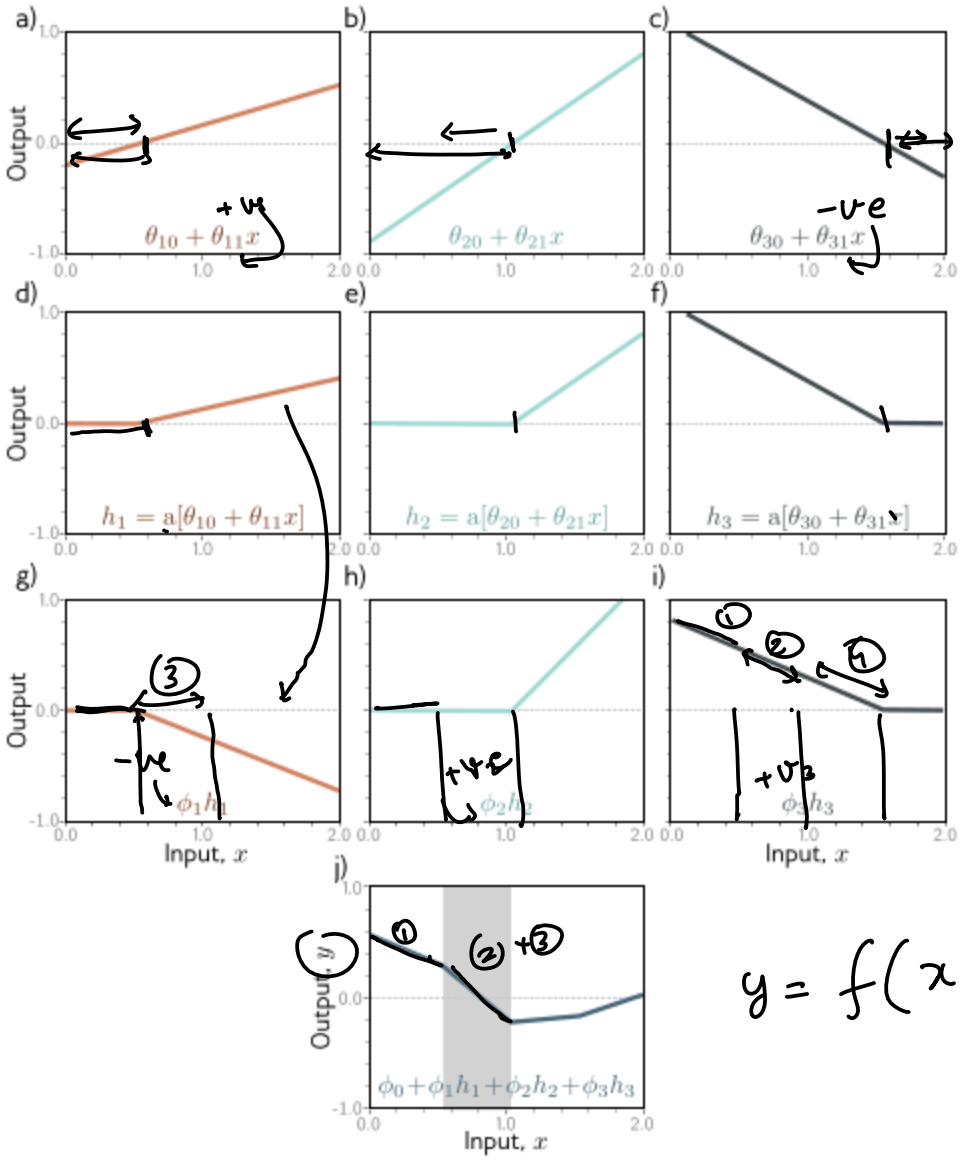
$x \in \mathbb{R}$

$$\begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix} = a\left(\begin{bmatrix} \theta_{11} \\ \theta_{21} \\ \theta_{31} \end{bmatrix} x + \begin{bmatrix} \theta_{10} \\ \theta_{20} \\ \theta_{30} \end{bmatrix}\right)$$

$$h_3 = a(\theta_{30} + \theta_{31}x)$$

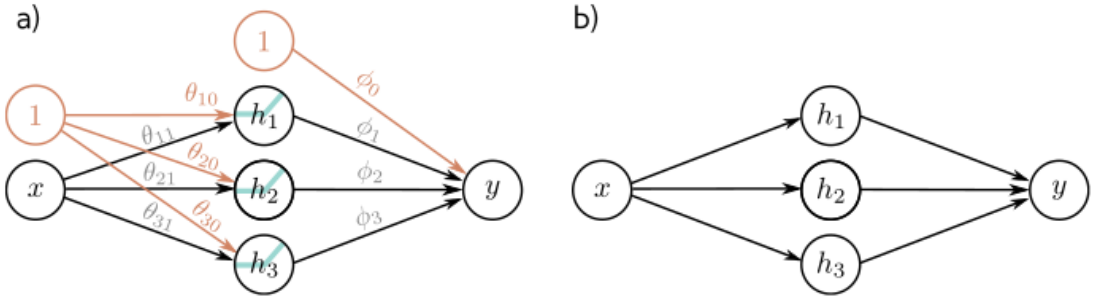
$$y = \phi_0 + \phi_1 h_1 + \phi_2 h_2 + \phi_3 h_3$$

Linear



$$y = f(x; \theta_{10}, \ldots, \theta_{13})$$

Depicting Neural Networks



$L_2(L_1(x))$

① $y = L_2(L_1(a(x)))$

② $y = L_2(a(L_1(x)))$ — multi dimensional

③ $y = a(L_2(a(L_1(x))))$ — softmax

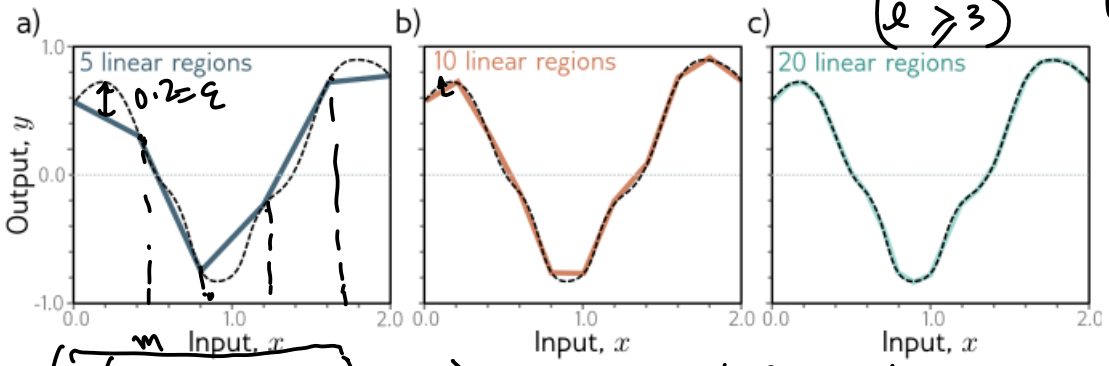
$y \in [0,1]$ $\quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ $\quad y_i \in [0,1]$, $\sum_i y_i = 1$

Two layer Perceptron = Multi-layer Perceptron (MLP) = Shallow Neural Network ($l=2$)

Universal Approximation Theorem

$m \to \infty$ $\epsilon \to 0$

Deep Neural Network ($l \geq 3$)



approx

$$f(x;\cdot) = \underline{w}_2^T\left(a\left(\underbrace{W_1}_{m\times n}\underbrace{\underline{x}}_{n\times 1} + \underline{b}_1\right) + b_2\right)$$

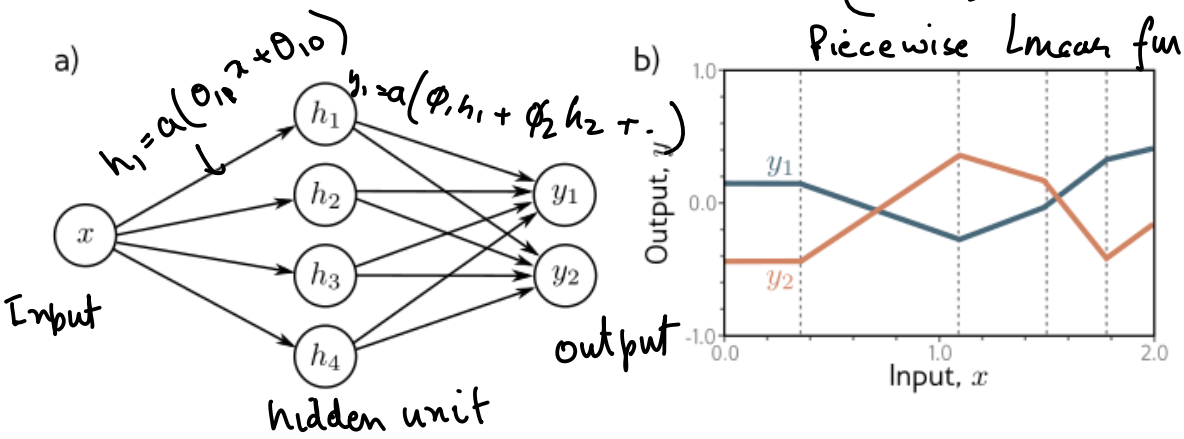
$$\begin{bmatrix} h_1 \\ \vdots \\ h_m \end{bmatrix} = a\left(W_1\underline{x} + \underline{b}_1\right)$$

hidden unit

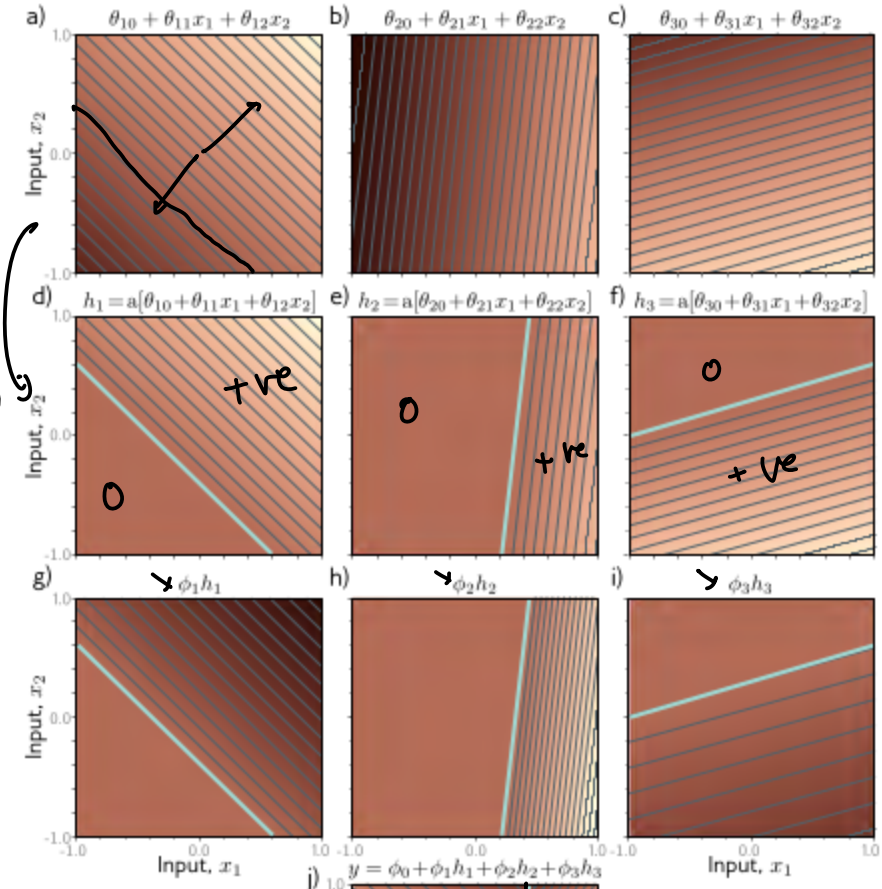
(ReLU)

Multivariate outputs

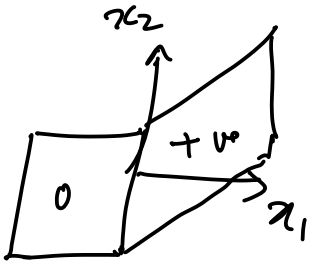
Piecewise Linear function

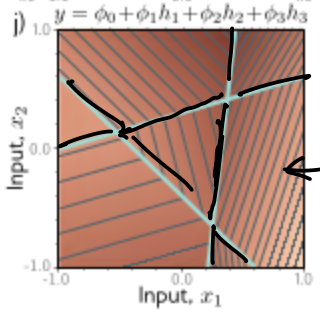


Multiple inputs

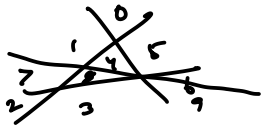


ReLU(a)

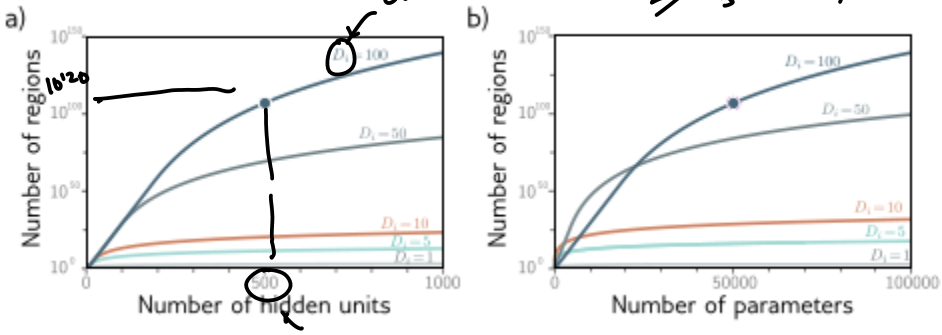




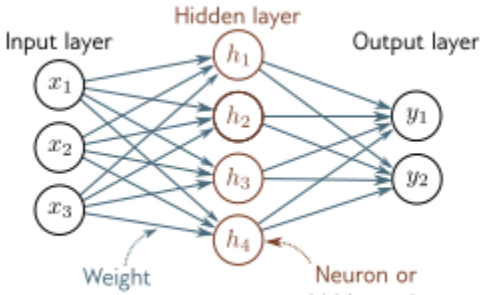
Polygons

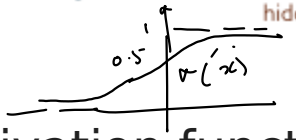


dimension of input



10^{120}





$\sigma(x) = \frac{1}{1+\exp(-x)}$ logistic regression

Sigmoid $\in (0,1)$

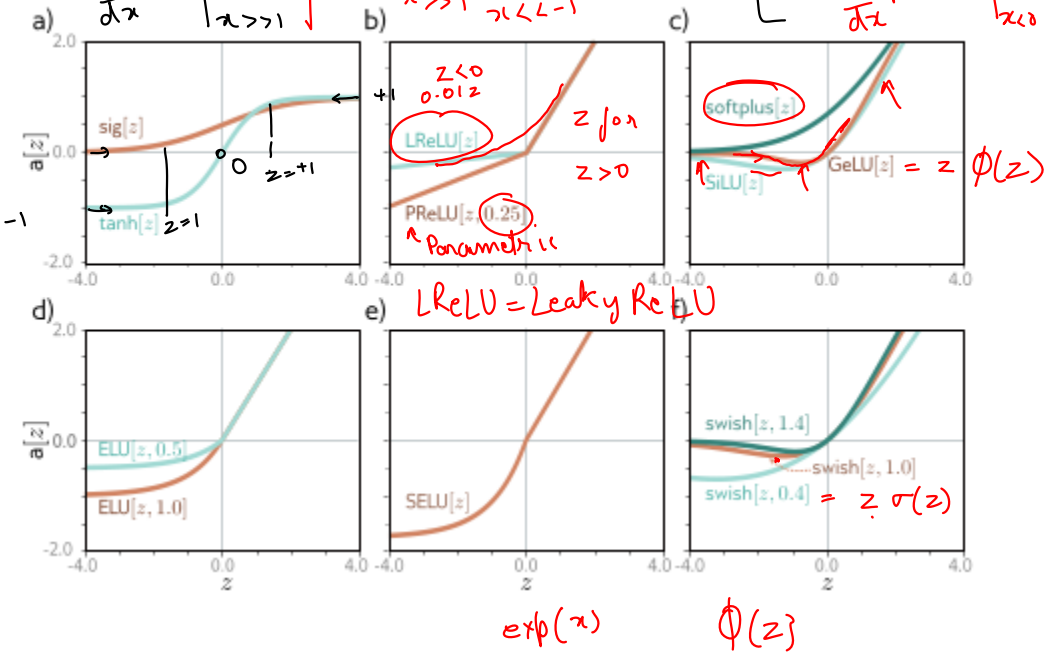
$\tanh(x) \in (-1,1) = \frac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)}$

ReLU

Activation functions

$\frac{d}{dx}\sigma(x)\Big|_{x>>1} \approx 0$ Saturation problem $x >> 1$ $x << -1$

$\frac{d}{dx}\mathrm{ReLU}(x)\Big|_{x<0} = 0$



LReLU = Leaky ReLU

$\exp(x)$ $\Phi(z)$

Hard Swish

Hard Sigmoid

Sigmoid

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

Hyperbolic tangent

$$\tanh(z) = \frac{exp(z) - exp(-z)}{exp(z) + exp(-z)}$$

Parametric Rectified Linear Unit

$$\mathrm{PReLU}(z, \alpha) = \begin{cases} z & \text{if } z > 0 \\ \alpha z & \text{if } z \leq 0 \end{cases}$$

Leaky Rectified Linear Unit

$$\text{LReLU}(z) = \text{PReLU}(z, \alpha = 0.01)$$

Softplus

$$\text{softplus}(z) = \frac{1}{\beta}\log(1 + \exp(\beta z))$$

Gaussian error Linear Units

$$\text{GELU}(z) = z\Phi(z)$$

where $\Phi(z)$ is the error function or the cumulative distribution function of a Gaussian distribution.

Sigmoid Linear Unit

$$\text{SiLU}(z) = z\sigma(z)$$

Exponential Linear Unit

$$\text{ELU}(z, \alpha) == \begin{cases} z & \text{if } z > 0 \\ \alpha(\exp(z) - 1) & \text{if } z \leq 0 \end{cases}$$

Scaled exponential linear unit

$$\text{SELU}(z) = 1.0507 * \text{ELU}(z, 1.673)$$

Swish

$$\text{Swish}(z, \beta) = z\sigma(\beta z)$$

HardSwish

$$\text{HardSwish}(z) = \begin{cases} 0 & z < -3 \\ z(z+3)/6 & -3 \leq z \leq 3 \\ z & z > 3 \end{cases}$$