FoDA L28 : Neural Networks

+ other classifiers

Classification

Input $X \subset \mathbb{R}^d$ **labels** $y \in \{-1, +1\}^n$

$x_1, \dots, x_n$

goal **function** $g : \mathbb{R}^d \to \mathbb{R}$

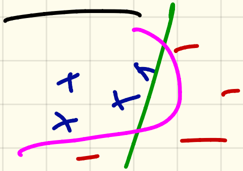
want $\text{sign}(g(x_i)) = y_i$

linear $g_\alpha(x) = \langle \overset{w}{\alpha}, x \rangle + \overset{b}{\alpha_0}$

(non-linear) **kernel** $g_\alpha(p) = \sum_{j=1}^{n} \alpha_j K(x_j, p) + \alpha_0$

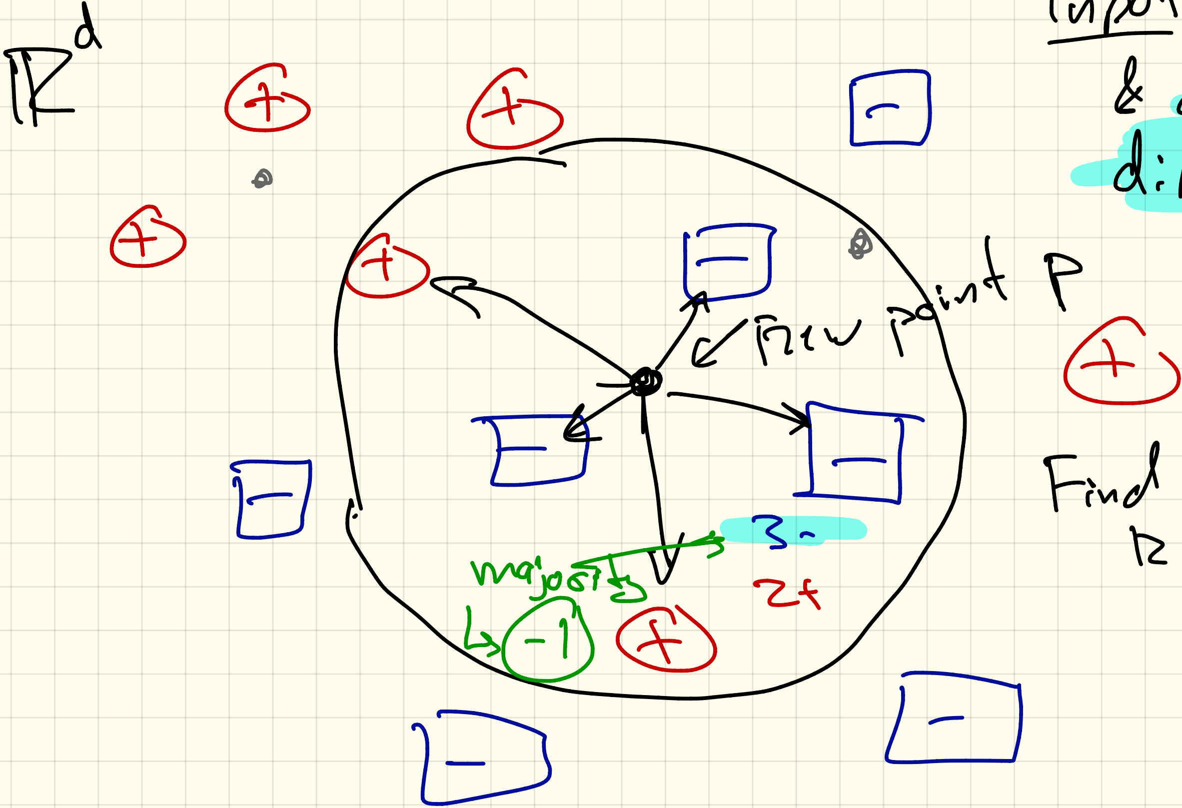
↳ still be optimized w/ GD

convex



KNN Classifiers

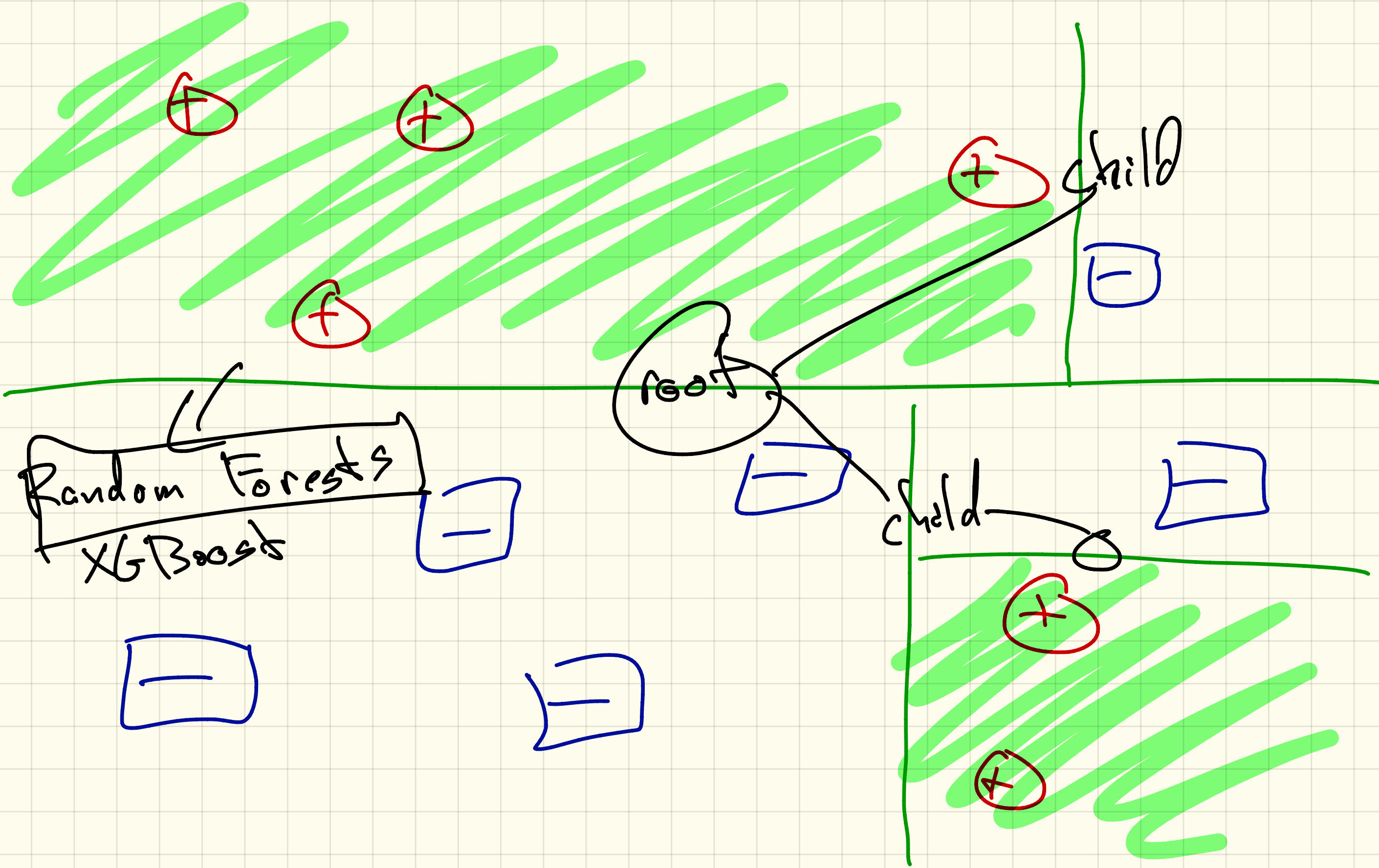
Input (X, y)
& distance $d: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$



$(k=5)$

Find closest k neighbors

Decision Trees

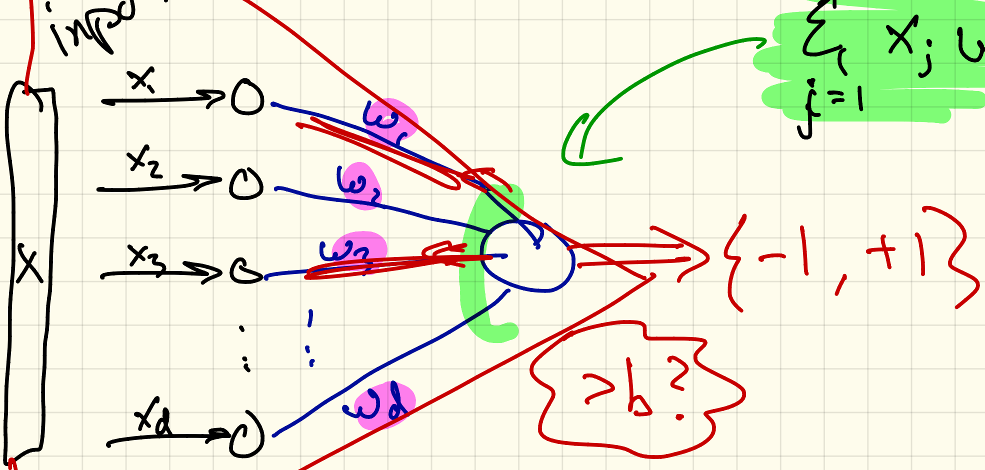


Random Forests
XGBoost

Neuron

e.g. uniform $w_j = \frac{1}{d}$

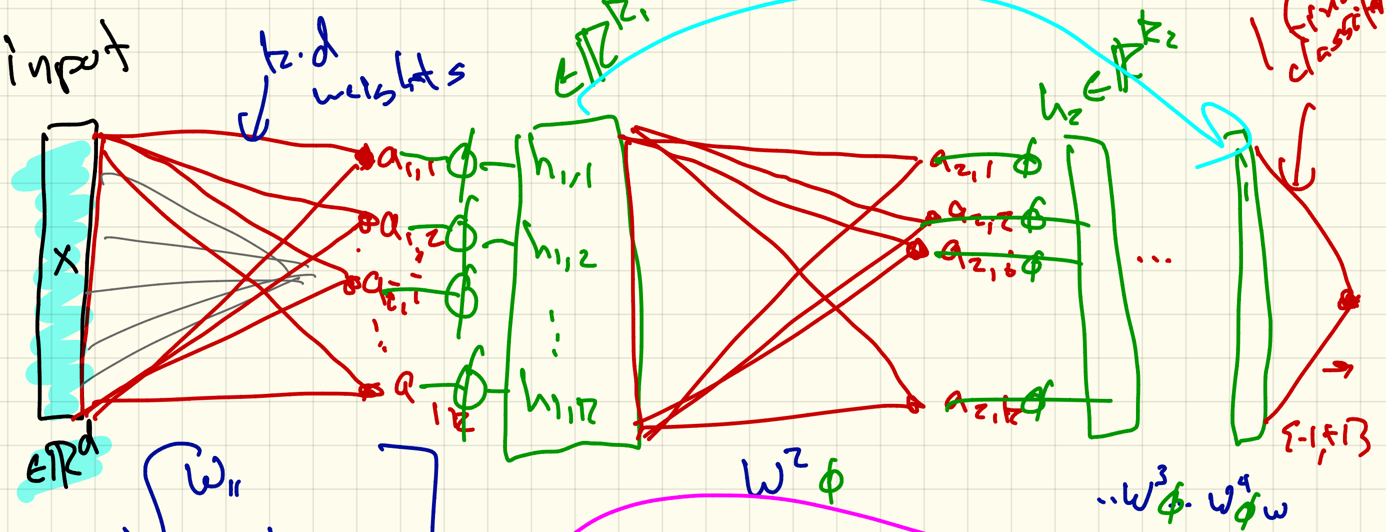
$$\sum_{j=1}^{d} x_j w_j \overset{?}{>} b$$



linear classifier

$$\sum_{j=1}^{d} x_j w_j > b \iff \langle x, w \rangle - b > 0$$

Neural Network



$$W^1 \begin{bmatrix} w_{11} & & \\ & \ddots & \\ & & w_{d,k} \end{bmatrix}$$

$$a_1 = Wx$$

ϕ activation

$$\phi : \mathbb{R} \to [-1, +1]$$

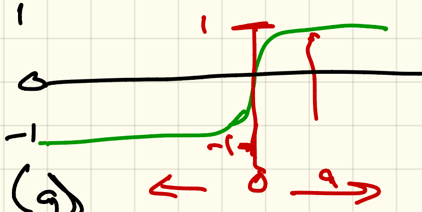
$$[0, 1]$$

$$[0, \infty)$$

Activation Functions ϕ

goal: ① proxy $f: \mathbb{R} \to \begin{cases} +1 & \text{if } \geq 0 \\ -1 & \text{if } < 0 \end{cases}$

② differentiable

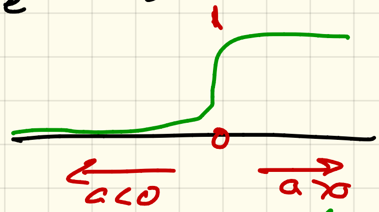


hyperbolic tangent

$$\phi(a) = \tanh(a) = \frac{e^a - e^{-a}}{e^a + e^{-a}} \in [-1, +1]$$

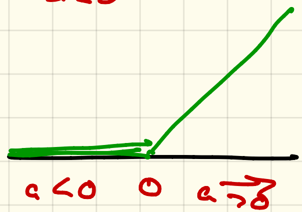
sigmoid

$$\phi(a) = \frac{e^a}{e^a + 1} \in [0, 1]$$



ReLU

$$\phi(a) = \max\{0, a\} \in [0, \infty)$$



Neural Network Function

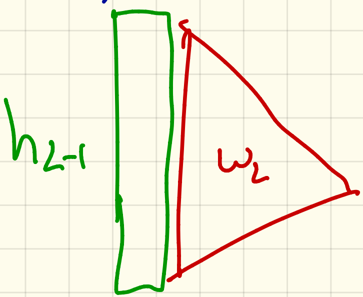
$$g_\alpha(x) = g(x) = b + w^L \underbrace{\phi\left(w^{L-1} \cdots \phi\left(w^2 \cdot \phi\left(w^1 x\right)\right) \cdots\right)}_{h_{L-1}}$$

$$\#\text{parameters} \left(1 + k_{L-1} + k_{L-1}\cdot k_{L-2} + \dots + k_1 \cdot k_2 + d \cdot k_1\right) \to \alpha \in \mathbb{R}^{\#P}$$

Back Propagation

$\nabla g_\alpha(x)$

$\nabla g_{w^L}(x)$



h_{L-1}

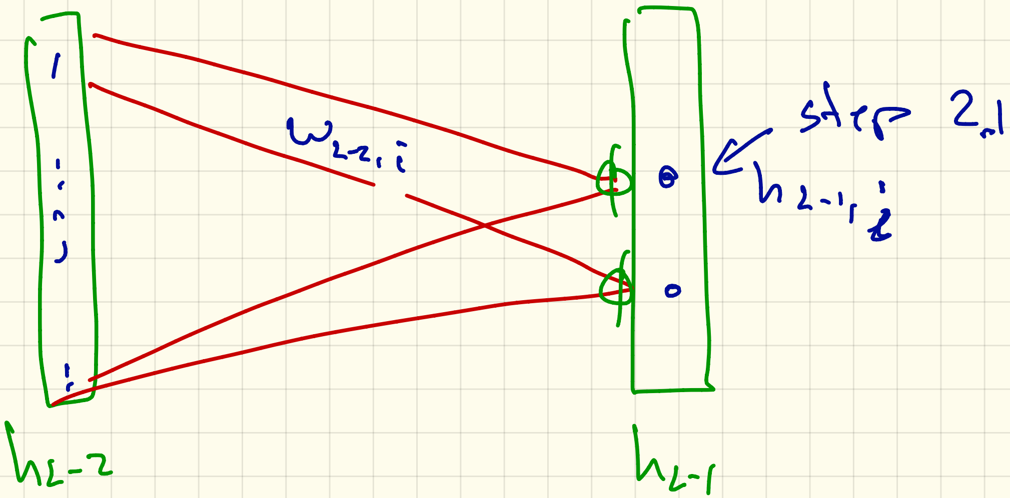
$\ell(z_i)$

$z_i = y_i g(x_i)$

$g(x_i)$

y_i

1. compute $g(x)$
   - 1a $h_1(x)$
   - 1b $h_2(x) = h_2(h_1(x))$
   - 1_{L-1} $h_{L-1}(x) = h_{L-1}(\dots h_1(x))$
2. $\nabla g(x)$
   - 2a $\nabla_{w_L}(x, y, h_{L-1})$
   - ⋮



$$\nabla_{w_{L-2,ij}}(x, y) =$$

instead of $\ell(z_i)$

$\phi(\nabla h_{L-1,i})$

chain rule

↳ dynamic programming