Boosting
Decision Trees
Neural Networks

Notes are available online (see Piazza)

Boosting, Decision Trees, Neural Networks

* Revisit surrogate loss : exp-loss
* Boosting and exp-loss connected
* Decision trees, exp-loss, and boosting connected
* From boosted classifiers to Neural Networks

Notes:

* old-new view that is not covered in one book or a few papers
* Intuitive, simple, "proof-free" (almost)
* Focus on empirical loss $\mathcal{L}_S(f)$ sample

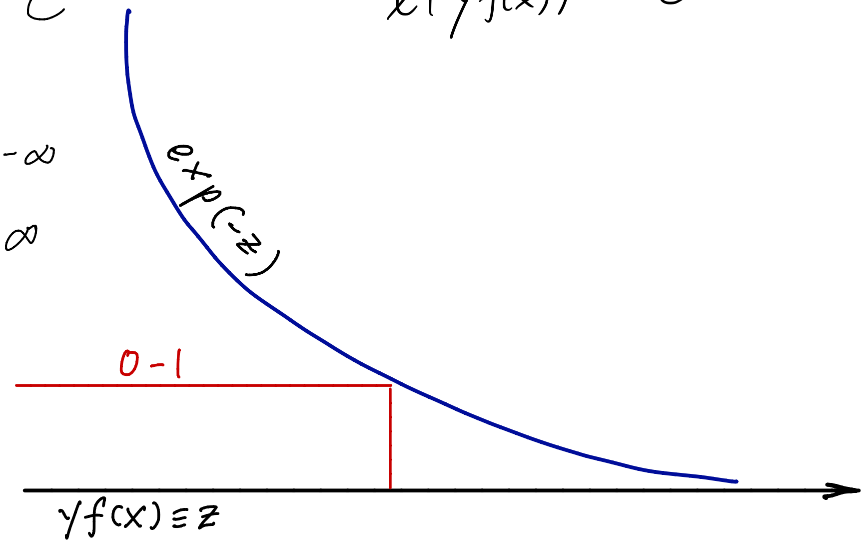
Exponential Loss

$f : \mathcal{X} \to \mathbb{R}$ $\qquad |f(x)|$ – confidence in prediction

$\mathrm{sign}(f(x))$ – predicted outcome

$$\mathcal{L}_S^{\exp}(f) \equiv \frac{1}{|S|} \sum_{i=1}^{|S|} e^{-y_i f(x_i)} \qquad \ell(yf(x)) = e^{-y f(x)}$$

* grows fast ↗ as $z \to -\infty$
* gets small ↘ as $z \to \infty$
* "nice" properties



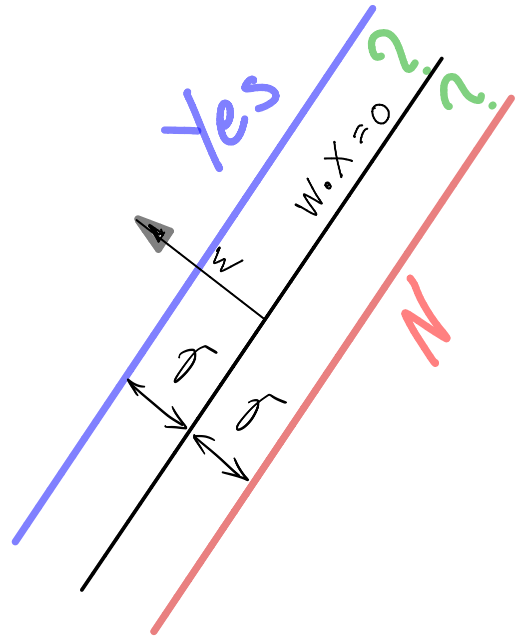
Predictors which can abstain

$f(x) \in \{-1, 0, +1\}$ $\qquad$ $f : \mathcal{X} \longrightarrow \{-1, 0, +1\}$

no
negative

IDK
IDC

yes
positive



Example

$x \in \mathbb{R}^d \qquad w \in \mathbb{R}^d$

$$f(x) = \begin{cases} -1 & w \cdot x < -\gamma \\ 0 & |w \cdot x| \le \gamma \\ +1 & w \cdot x > \gamma \end{cases}$$

Calibrating Predictors

* Assume that $f: \mathcal{X} \to \{-1, 0, +1\}$ is given us
* Need to find α such that $\mathcal{L}_S^{exp}(\alpha f(x))$ is minimized

Can use α_+, α_-

* Process called calibration / rescaling / reweighing

$$\min_{\alpha \in \mathbb{R}} \frac{1}{|S|} \sum_{i=1}^{|S|} e^{-y_i (\alpha f(x_i))}$$

$\Downarrow$

$\tilde{x}_i \in \mathbb{R}$

Search for α

* Exploit the fact that $f : \mathcal{X} \to \{-1, 0, +1\}$

$$S_+ = \{ i \mid y_i f(x_i) = +1 \}$$ Correct

$$S_0 = \{ i \mid y_i f(x_i) = 0 \}$$ Blank

$$S_- = \{ i \mid y_i f(x_i) = -1 \}$$ Mistake

$$\boxed{|S| = m}$$

$$\mathcal{L}_S^{exp}(\alpha f) = \frac{1}{m}\left(\sum_{i \in S_+} e^{-\alpha} + \sum_{i \in S_-} e^{+\alpha} + \sum_{i \in S_0} e^{0} \right)$$

Are we in a better shape?

Closed form solution

Define

$$\mu_+ = \frac{|S_+|}{m} \qquad \mu_- = \frac{|S_-|}{m} \qquad \mu_0 = \frac{|S_0|}{m}$$

μ_+: fraction correct

μ_-: fraction mistake

μ_0: what do thy care

$$\mathcal{L}_S^{exp}(\alpha f) = \mu_+ e^{-\alpha} + \mu_- e^{\alpha} + \mu_0$$

$\mathcal{L}(\alpha f)$ is convex in α Yeh! Yey! Ci!!p!

$$0 = \frac{d\mathcal{L}}{d\alpha} = -\mu_+ e^{-\alpha} + \mu_- e^{\alpha} \Rightarrow \alpha^* = \frac{1}{2}\log\left(\frac{\mu_+ m}{\mu_- m}\right)$$

$$\boxed{\text{Nobody is perfect assumption: } \mu_+ > 0 \quad \mu_- > 0}$$

Further Insights

* If $\mu_+ = \mu_-$ then $\alpha = 0$ $\rightarrow$ $f(x)$ is no better than a random predictor

* If $\mu_- > \mu_+$ then $\alpha < 0$ $\rightarrow$ Negate the prediction of f

---

Total Loss of $\alpha f(x)$:

$$\mathcal{L}_S^{exp}(\alpha f) = \mu_+ e^{-\frac{1}{2}\log\left(\frac{\mu_+}{\mu_-}\right)} + \mu_- e^{\frac{1}{2}\log\left(\frac{\mu_+}{\mu_-}\right)} + \mu_0$$

$$= \mu_+ \sqrt{\frac{\mu_-}{\mu_+}} + \mu_- \sqrt{\frac{\mu_-}{\mu_+}} + \mu_0$$

$$e^{\frac{1}{2}\log a} = e^{\log\sqrt{a}} = \sqrt{a}$$

$$\mu_+ + \mu_- + \mu_0 = 1$$

$$= 2\sqrt{\mu_+\mu_-} + (1 - (\mu_- + \mu_-))$$

Total loss of αf

$$\mathcal{L}_S^{\exp}(\alpha f) = \mu_+ e^{-\frac{1}{2}\log\left(\frac{\mu_+}{\mu_-}\right)} + \mu_- e^{\frac{1}{2}\log\left(\frac{\mu_+}{\mu_-}\right)} + \mu_0$$

$$e^{\frac{1}{2}\log a} = e^{\log\sqrt{a}} = \sqrt{a}$$

$$= \mu_+ \sqrt{\frac{\mu_-}{\mu_+}} + \mu_- \sqrt{\frac{\mu_+}{\mu_-}} + \mu_0$$

$$\mu_+ + \mu_- + \mu_0 = 1$$

$$= 2\sqrt{\mu_+\mu_-} + (1 - (\mu_+ + \mu_-))$$

No normalization $\Longrightarrow m - \mu_+ - \mu_-$

Post Calibration Loss

$$\mathcal{L}(\alpha f) = 2\sqrt{\mu_+ \mu_-} + (1 - (\mu_+ + \mu_-))$$

$$= 1 - (\mu_+ - 2\sqrt{\mu^+}\sqrt{\mu^-} + \mu^-) \qquad (\sqrt{\mu^-})^2$$

$$= 1 - (\sqrt{\mu^+} - \sqrt{\mu^-})^2$$

$\Delta \text{Loss} \equiv$ loss before using f and after

$$\Delta\mathcal{L} \equiv \mathcal{L}(0) - \mathcal{L}(\alpha f) = (\sqrt{\mu^+} - \sqrt{\mu^-})^2$$

Reduction in loss has a closed form

Neat!

Beyond a single predictor

1) Assume we calibrated $\alpha_1 h_1(x)$

2) Provided with a second predictor $h_2(x)$

3) Task: combine $\underbrace{\alpha_1 h_1(x)}_{\text{known}} + \overset{?}{\alpha_2}\, \underset{\text{provided}}{h_2(x)} \equiv f_2(x)$

$\equiv f_1(x)$

$f_1 : \mathcal{X} \to ? \{-\alpha_1, 0, \alpha_1\}$

4) General case: given $f_t(x) = \sum_{i=1}^{t} \alpha_i h_i(x)$ & $h_{t+1}(x)$

find $\alpha_{t+1} \Rightarrow f_{t+1}(x) = \sum_{i=1}^{t+1} \alpha_i h_i(x)$

$= f_t(x) + \alpha_{t+1} h_{t+1}(x)$

Inductive Calibration

As before define:

$$S_+^{t+1} = \{ i \mid h_{t+1}(x_i) = y_i \}$$

$$S_-^{t+1} = \{ i \mid h_{t+1}(x_i) = -y_i \}$$

$$S_0^{t+1} = \{ i \mid h_{t+1}(x_i) = 0 \}$$

Generalize:

$$\mu_+ = \frac{1}{m} \sum_{i \in S_+} e^{-y_i f_t(x_i)}$$

$$\mu_0 = \frac{1}{m} \sum_{i \in S_0} e^{-y_i f_t(x_i)}$$

$$\mu_- = \frac{1}{m} \sum_{i \in S_-} e^{-y_i f_t(x_i)}$$

$$\mathcal{L}(f_{t+1}(x)) = \mu_+ e^{-\alpha_{t+1}} + \mu_- e^{+\alpha_{t+1}} + \mu_0$$

$$f_{t+1}(\cdot) = f_t(\cdot) + \alpha_{t+1} h_{t+1}(\cdot)$$

Solution has the same form!

$$\Delta\mathcal{L}_{t+1} \equiv \mathcal{L}(f_t) - \mathcal{L}(f_{t+1}) = \left(\sqrt{\mu^+} - \sqrt{\mu^-}\right)^2$$

$$(\mu_+ + \mu_- + \mu_0) - \left(2\sqrt{\mu^+\mu^-} + \mu_0\right)$$

Importance Weights

* Given $f_t(x)$ define:

$$q_i^t \sim e^{-y_i f_t(x_i)}$$

importance weight: how "difficult" example (x_i, y_i) is

* Often $\{q_i^t\}$ is normalized $\sum_{i=1}^{m} q_i^t = 1$

* Does not change analysis and algorithm

* Simply implies $\mu_+^t + \mu_-^t + \mu_0^t = 1$

Which can be obtained by simple scalinging

Detour - SGD w/ Weights

$$\mathcal{L}(w) = \sum_{i=1}^{m} q_i f_i(w) \quad \text{s.t.} \quad \sum q_i = 1 \, ; \; q_i > 0$$

Reduction: define $\tilde{f}_i(w) = q_i f_i(w)$

use SGD on: $\frac{1}{m} \sum_{i=1}^{m} \tilde{f}_i(w)$

OR

Importance Sampling: batch size b

Sample b times such that example i is chosen w/ probability q_i

Boosting

Equivalent to previous lecture

Initialize: $\forall i \quad q_i^1 = \frac{1}{m} \quad f_0(x) \equiv 0$

For $t = 1, \dots, T$:

Find $h_t(x)$ s.t. $\sum_i q_i^t y_i h_t(x_i)$ is large

Partition: $S \mapsto S_+^t, \; S_-^t, \; S_0^t$

μ_+^t
μ_-^t
$\cancel{\mu_0^t}$

Calculate: $\alpha_t = \frac{1}{2} \log\left(\frac{\mu_+^t}{\mu_-^t}\right)$

Update:

1. $f_{t+1}(x) = f_t(x) + \alpha_t h_t(x)$
2. $q_i^{t+1} = q_i^t e^{-y_i \alpha_t h_t(x_i)}$

$$q_i^{t+1} = \begin{cases} q_i^t & h_t(x_i) = 0 \\ q_i^t e^{-\alpha} & h_t(x_i) = y_i \\ q_i^t e^{\alpha} & h_t(x_i) \neq y_i \end{cases}$$

Generalization

Instead of $h_t : \mathcal{X} \to \{-1, 0, +1\}$ can use

$$h_t(x) \in [-1, +1]$$ Limit the power of base predictors

where as before

$|h_t(x)|$ confidence

$\text{sign}(h_t(x))$ predicted outcome

$$S_+^t \triangleq \{i \mid y_i h_t(x_i) > 0\}$$
$$S_0^t \triangleq \{i \mid y_i h_t(x_i) = 0\}$$
$$S_-^t \triangleq \{i \mid y_i h_t(x_i) < 0\}$$

$\boxed{\text{No further changes are required}}$

Generalization II

Slide Can Be Skipped

Can replace exp-loss with log-loss (logistic-loss)

$$\mathcal{L}_S^{log}(f) = \frac{1}{m}\sum_{i=1}^{m} \log\left(1+e^{-y_i f(x_i)}\right)$$

$$q_i^{t+1} \sim e^{-y_i f_{t+1}(x_i)} \sim q_i^t e^{-y_i \alpha_t h_t(x_i)}$$

$$q_i^{t+1} = \frac{1}{1+\exp\left(y_i f_{t+1}(x_i)\right)} = \frac{q_i^t}{q_i^t + (1-q_i^t)e^{y_i \alpha_t h_t(x_i)}}$$

Altenatively: $z_i^{t+1} \triangleq y_i f_{t+1}(x_i) = z_i^t + y_i \alpha_t h_t(x_i)$

$$q_i^{t+1} = \frac{e^{-z_i^{t+1}}}{Z}$$ Exp Loss

$$q_i^{t+1} = \left(1+\exp(z_i^{t+1})\right)^{-1}$$ Log Loss

Boosting

Slide Can Be Skipped

Log-loss
Confidence-rated
W. H.

Initialize: $\forall i \quad q_i^1 = \cancel{\frac{1}{m}} \to \frac{1}{2} \qquad f_0(x) \equiv 0$

For $t = 1, \ldots, T$:

Find $h_t(x)$ s.t. $\sum_i q_i^t y_i h_t(x_i)$ is large

Partition: $S \mapsto S_+^t, \; S_-^t, \; S_0^t$

Calculate: $\alpha_t = \frac{1}{2} \log\left(\frac{\mu_+^t}{\mu_-^t}\right)$

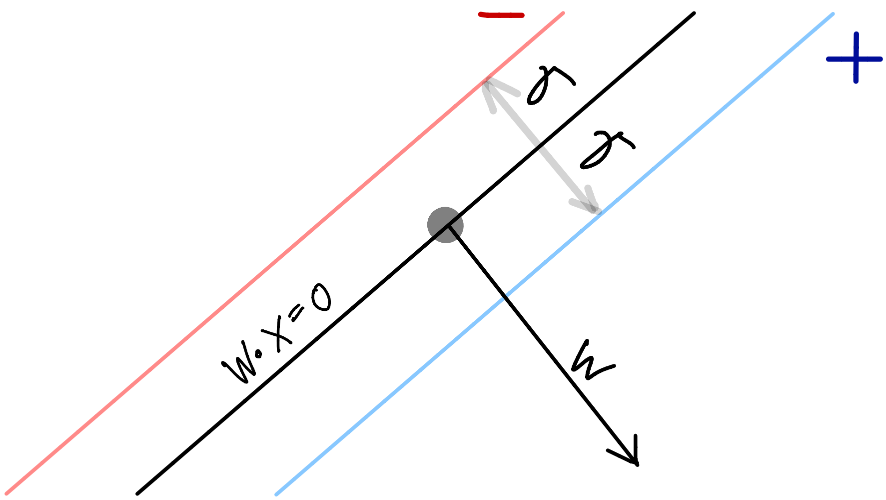
Update:

1. $f_{t+1}(x) = f_t(x) + \alpha_t h_t(x)$

2. $q_i^{t+1} = q_i^t e^{-y_i \alpha_t h_t(x_i)} \to$ 2. $q_i^{t+1} = q_i^t \left(q_i^t + (1 - q_i^t) e^{y_i \alpha_t h_t(x_i)}\right)^{-1}$

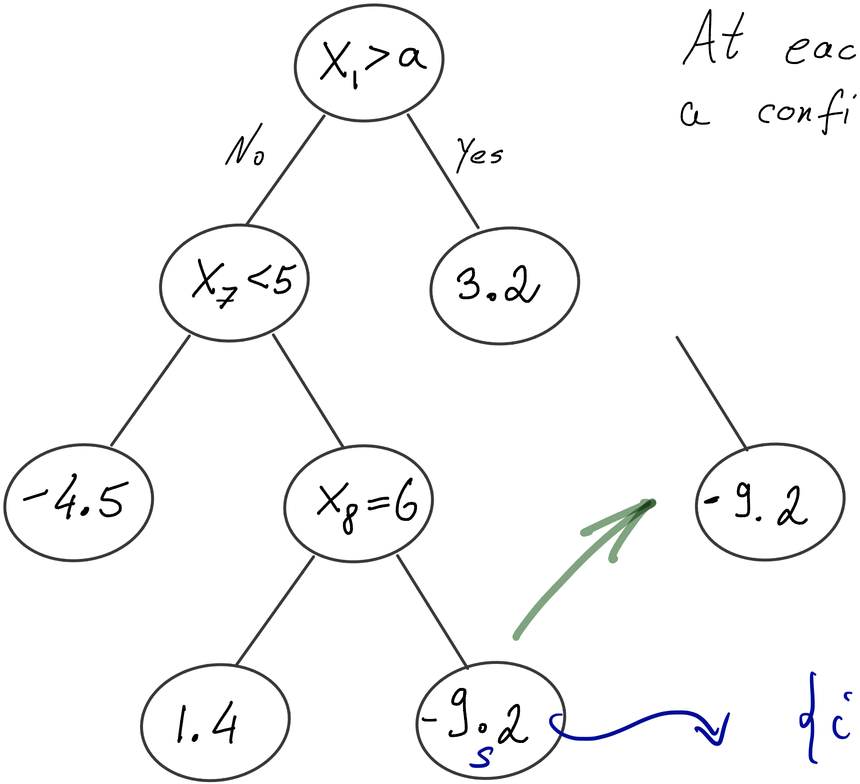
Detour - Linear Predictors w/ Abstantion

$f(x) = w^T x \Rightarrow \tilde{f}(x)$ w/ abstantion & continous



$$w^T x = z \qquad \tilde{f}(z) = \mathrm{sign}(z)\left[|z| - \tau\right]_+ = \begin{cases} z - \tau & z > \tau \\ 0 & |z| \leq \tau \\ z + \tau & z < -\tau \end{cases}$$

Decision Trees



At each leaf there exists a confidence-rated predictor

$$h_t(x) = \begin{cases} -9.2 & \text{if } x_1 \le a \wedge x_7 \ge 5 \wedge x_8 = 6 \\ 0 & \text{o.w.} \end{cases}$$

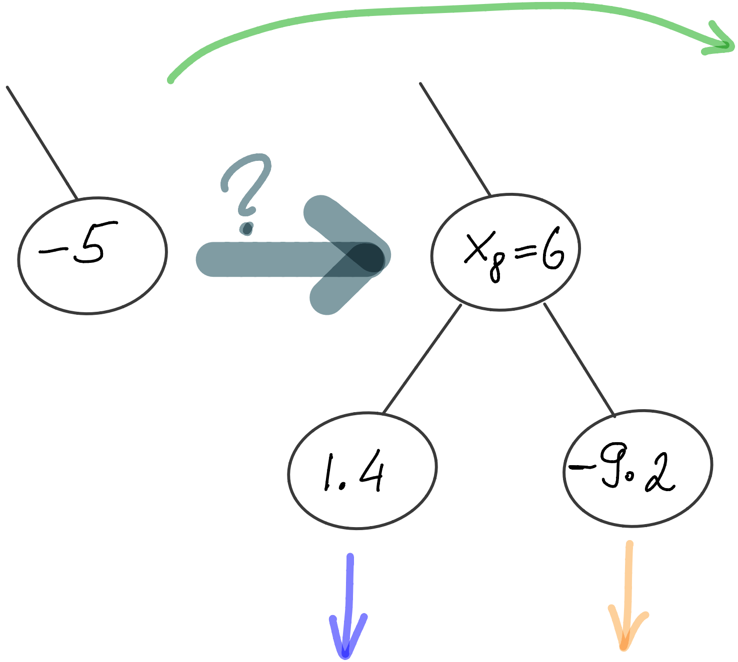
$\{i : y_i = +1\} = \mu_s^+$

$\{i : y_i = -1\} = \mu_s^-$

$\frac{1}{2}\log\left(\frac{\mu_s^+}{\mu_s^-}\right)$

Growing Decision Trees

μ_+, μ_-



$$\boxed{\begin{aligned} \mu_+ &= \mu_+^L + \mu_+^R \\ \mu_- &= \mu_-^L + \mu_-^R \\ \mu_0^R &= \mu_+^L + \mu_-^L \\ \mu_0^L &= \mu_+^R + \mu_-^R \end{aligned}}$$

$\mu_+^L, \mu_-^L, \mu_0^L$ $\qquad$ $\mu_+^R, \mu_-^R, \mu_0^R$

Improvement in surrogate loss?

Growing ~~Pain~~ Gain

Assume $\mu_+ + \mu_- = 1$ for parent node (normalization)

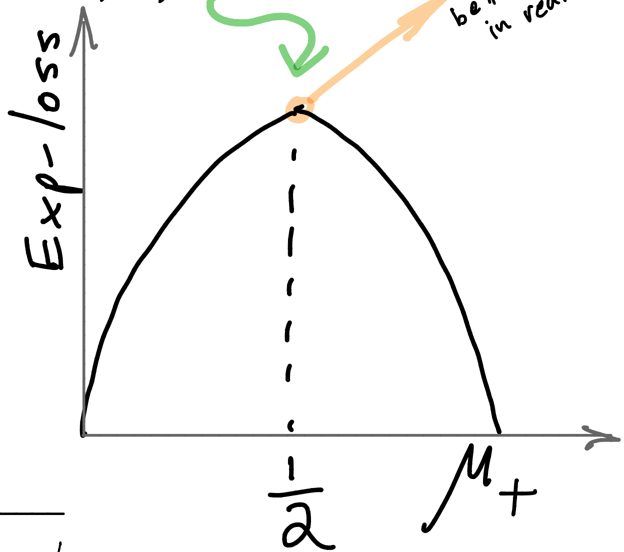
Exp-loss for parent node : $2\sqrt{\mu_+ \mu_-}$

Exp-loss for children:

$$2\left(\sqrt{\mu_+^L \mu_-^L} + \sqrt{\mu_+^R \mu_-^R}\right)$$

Gain ~

$$\sqrt{(\mu_+^R + \mu_+^L)(\mu_-^R + \mu_-^L)} - \sqrt{\mu_+^R \mu_-^R} - \sqrt{\mu_+^L \mu_-^L}$$



I am better looking in real life

Towards Neural Networks

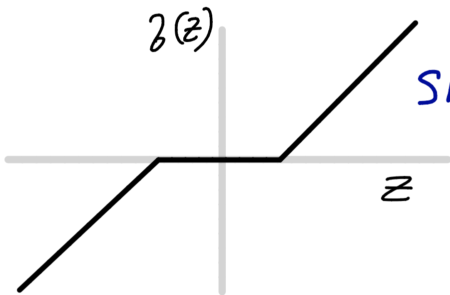
$$f_t(x) = \sum_{i=1}^{t} \alpha_i h_t(x) \quad \longrightarrow \quad \text{"Strong" hypothesis}$$

Suppose each $h_t(x)$ is of the form:

$$\sigma(w^t \cdot x) \quad \text{where } \sigma : \mathbb{R} \to \mathbb{R} \text{ is non-linear}$$

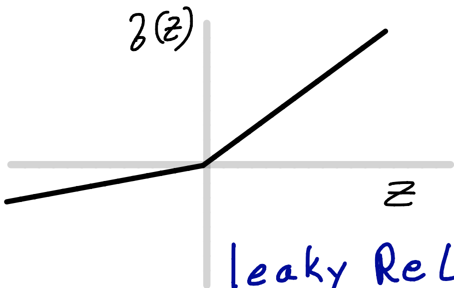
Examples:

$$\sigma(z) = \begin{cases} z - \tau & z > \tau \\ 0 & |z| \leq \tau \\ z + \tau & z < -\tau \end{cases}$$



Shrinkage

$$\sigma(z) = \begin{cases} z & z \geq 0 \\ az & z < 0 \end{cases}$$

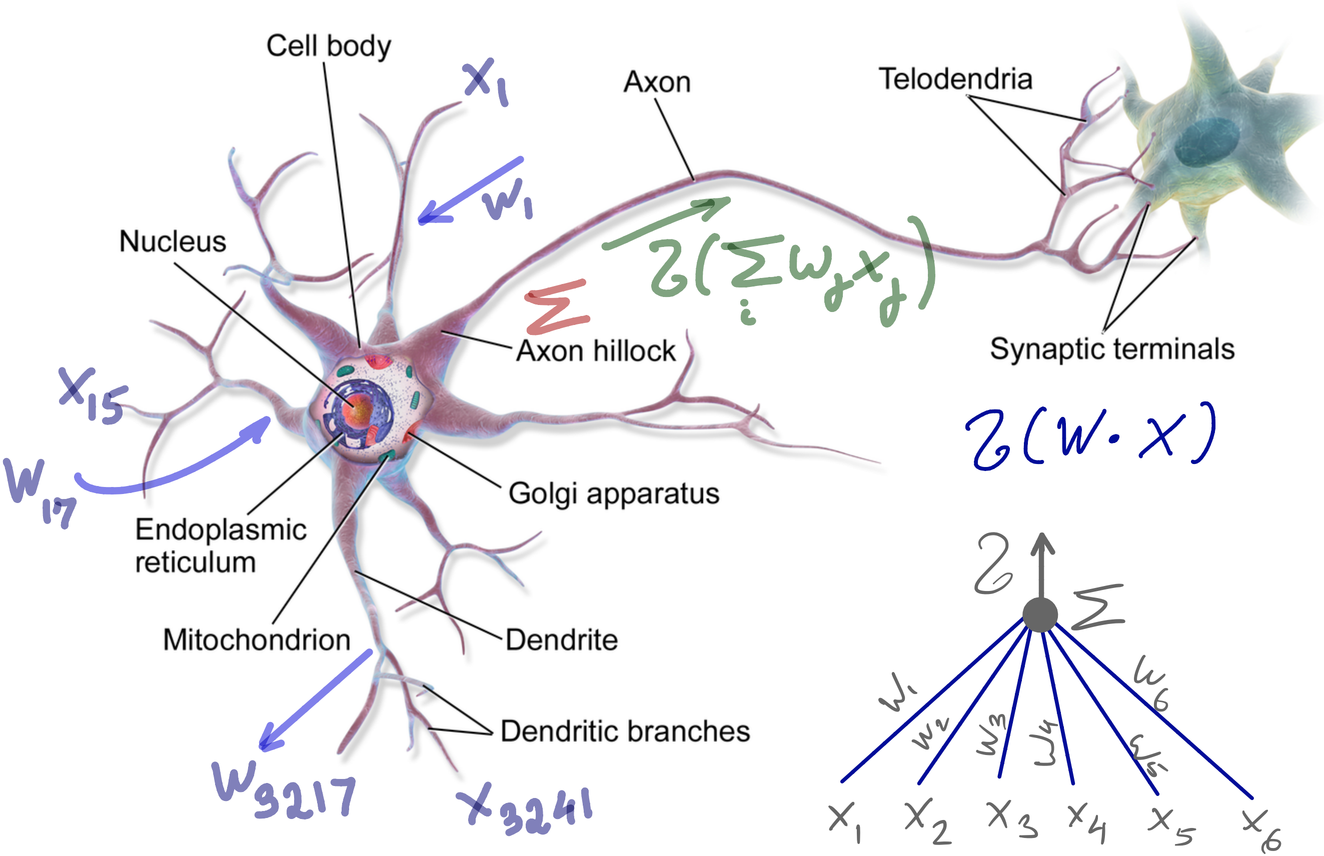


leaky ReLU

$$\sigma(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$$

Sigmoid

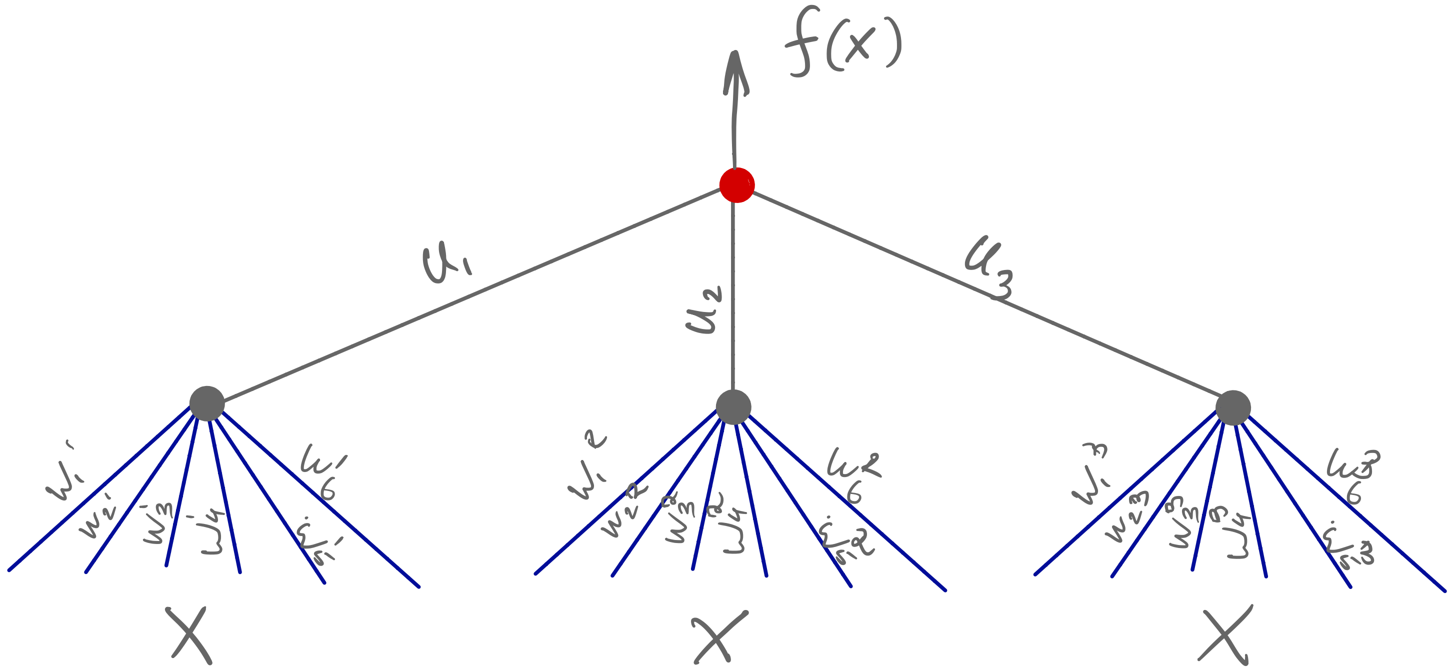
"Neuron"



Cell body
x1
Axon
Telodendria
w1
Nucleus
Σ
σ(Σ_i w_j x_j)
Axon hillock
Synaptic terminals
x15
σ(W·X)
W17
Golgi apparatus
Endoplasmic reticulum
σ
Σ
Mitochondrion
Dendrite
w1
w6
w2
w3
w4
w5
Dendritic branches
W3217
x3241
x1
x2
x3
x4
x5
x6

Multi layer perceptron (NN)

$$f(x) = \sigma\left(u_1\,\sigma(w^1\cdot x) + u_2\,\sigma(w^2\cdot x) + u_3\,\sigma(w^3\cdot x)\right)$$



- You are likely to have many questions at this point...
- A few will be addressed@IML
- Many will be left unanswered
- Since we don't know... the answer