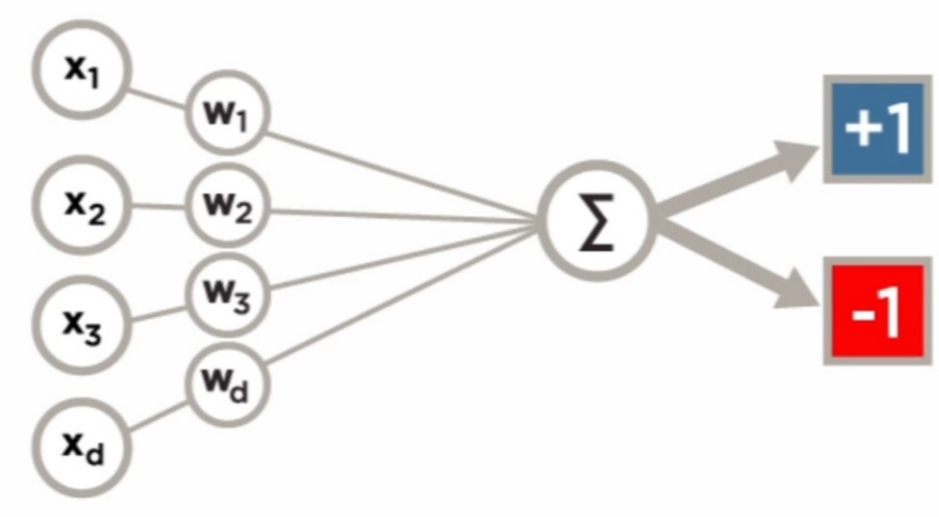
Office Hours: Learning with Linear Classifiers -- Perceptron

By Abraham Kang

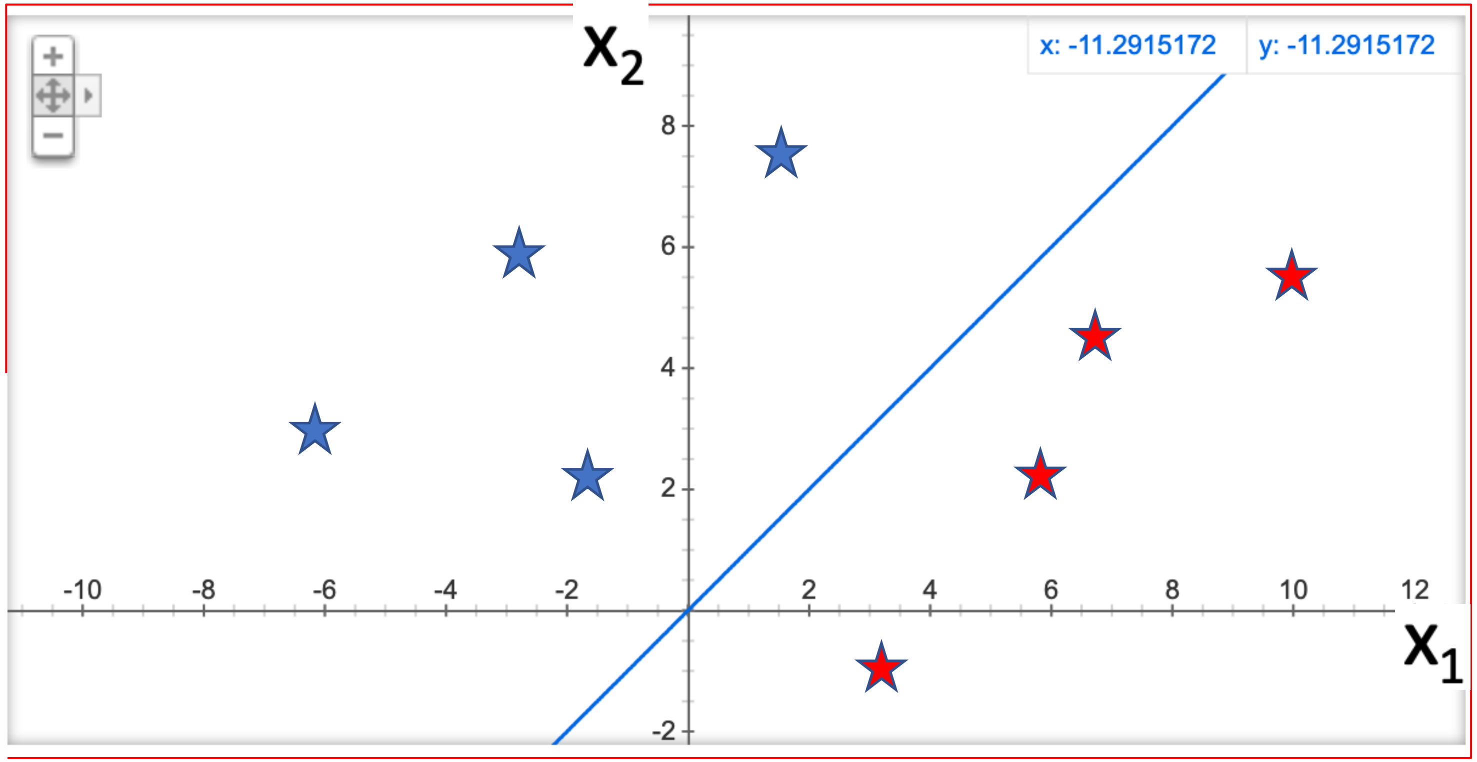
Before starting your work, please review eCornell's policy regarding plagiarism (the presentation of someone else's work as your own without source credit - https://s3.amazonaws.com/ecornell/global/eCornellPlagiarismPolicy.pdf).

What are we doing with Linear Classifiers?

- Given that x_1 and x_2 are our axis what are our w's and b's?



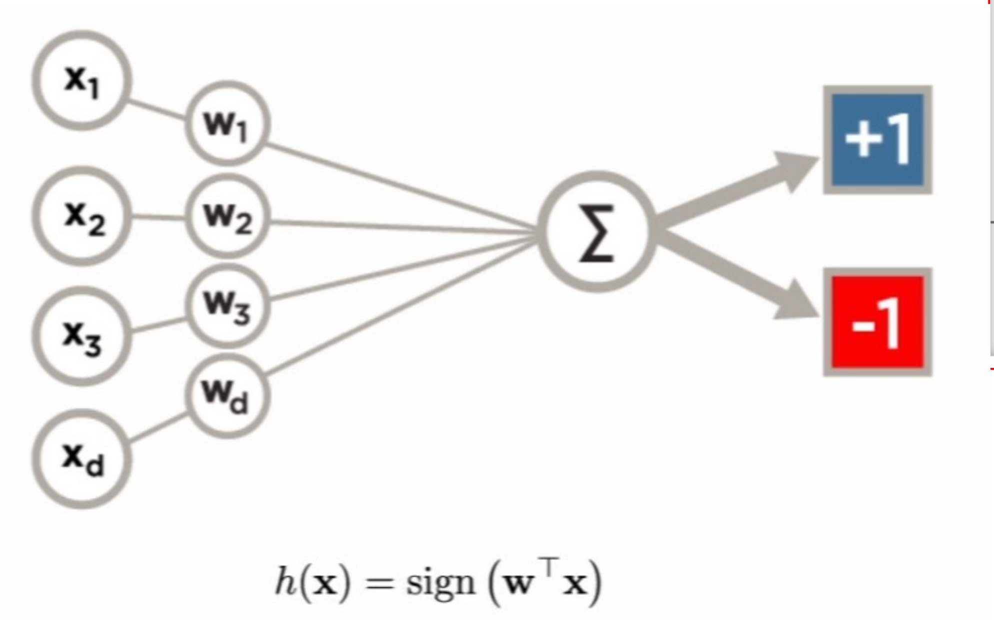
$h(\mathbf{x}) = \text{sign}\left(\mathbf{w}^\top \mathbf{x}\right)$

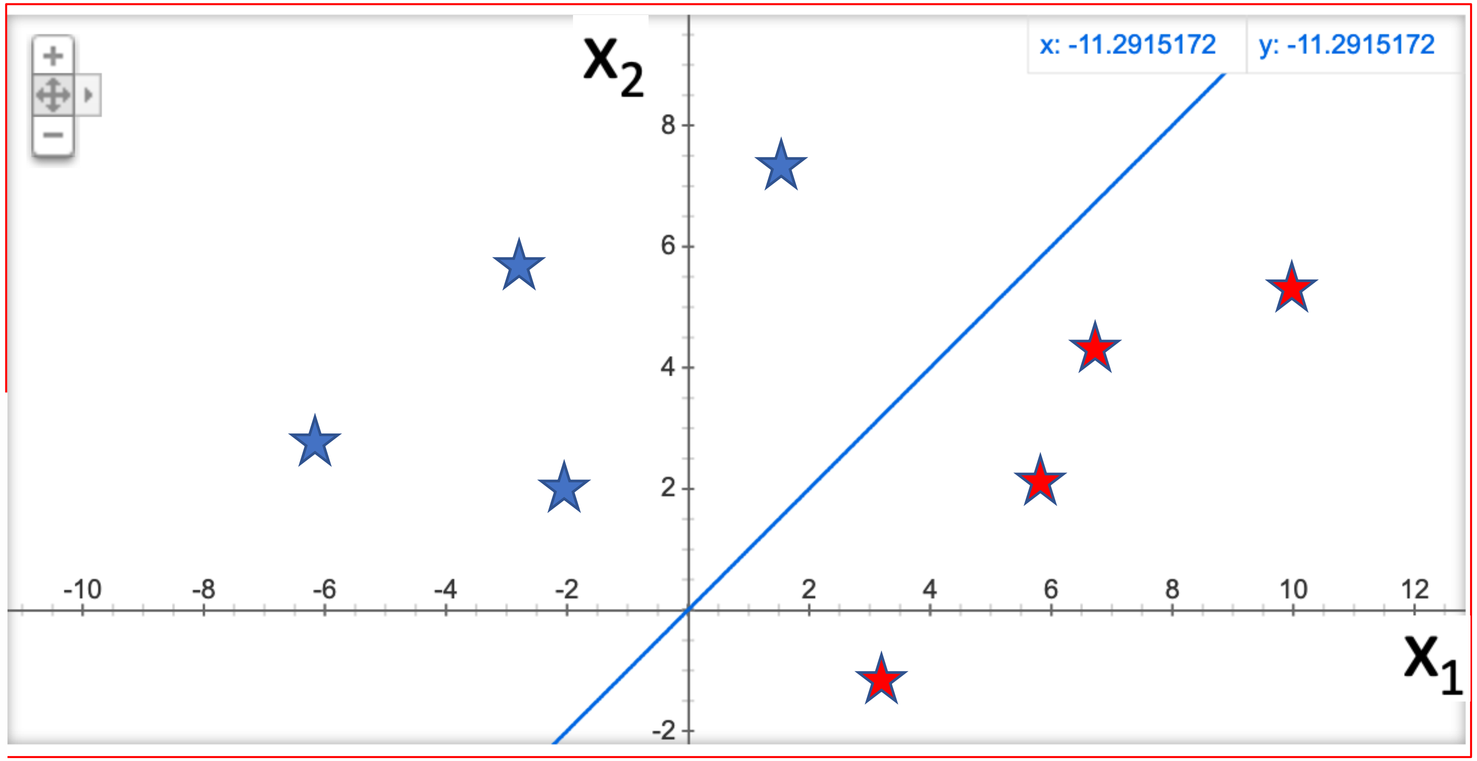


$0 = x_1 * w_1 + x_2 * w_2 + b_1$

What are we doing with Linear Classifiers?

- Given the x_1 and x_2 are our axis what are our w's and b's?
- w_1= -1, w_2=1
- b = 0





$$0 = -1x_1 + 1x_2$$

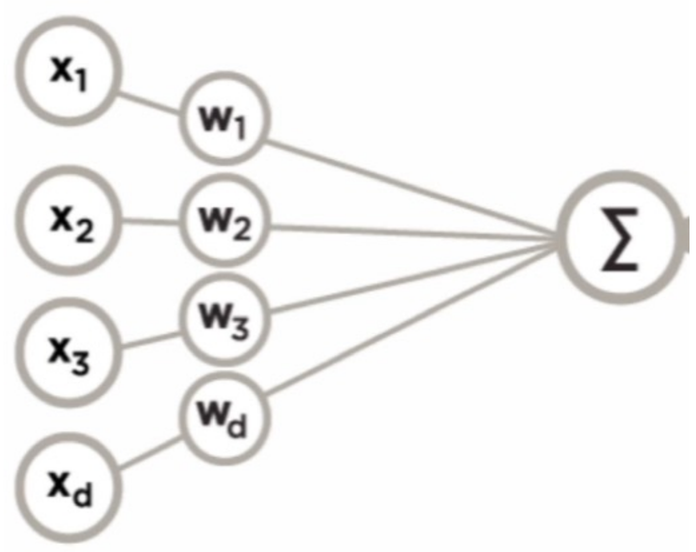
np.sign($-x_1 + x_2$)

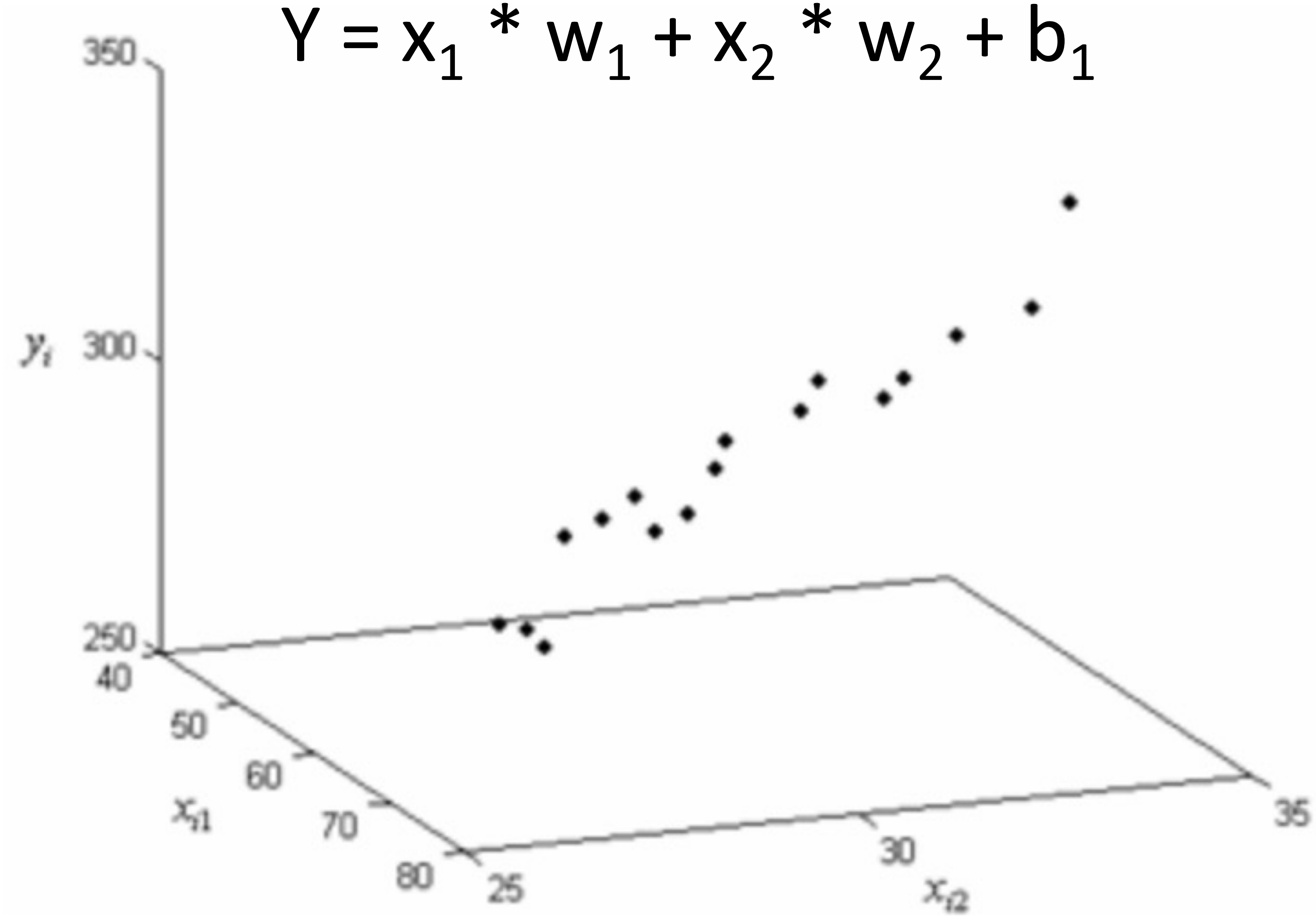
Scaling the line with W's

What are we doing with Linear Regression?

- Given the x_1, x_2 and y as our regression value, what are our w's and b's?

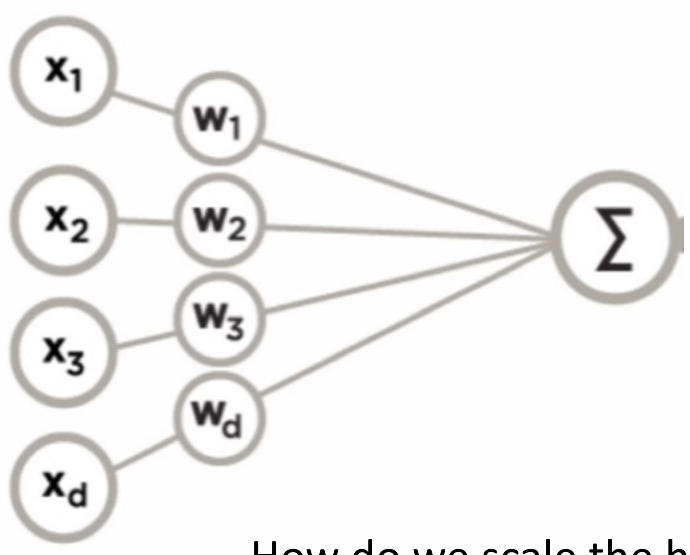
$$Y = x_1 * w_1 + x_2 * w_2 + b_1$$

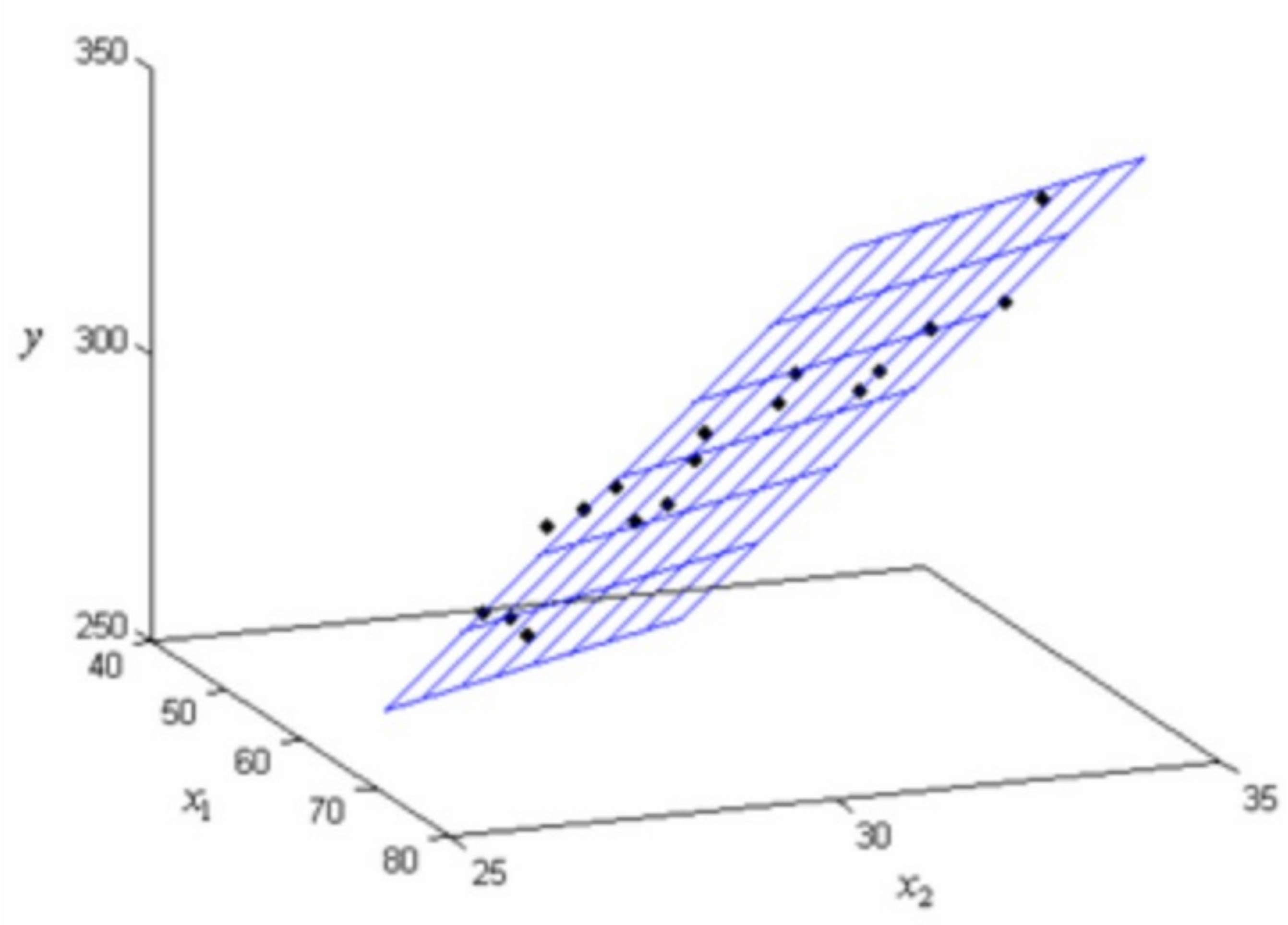




What are we doing with Linear Regression?

- Given the x_1, x_2 and y as our regression value, what are our w's and b's?

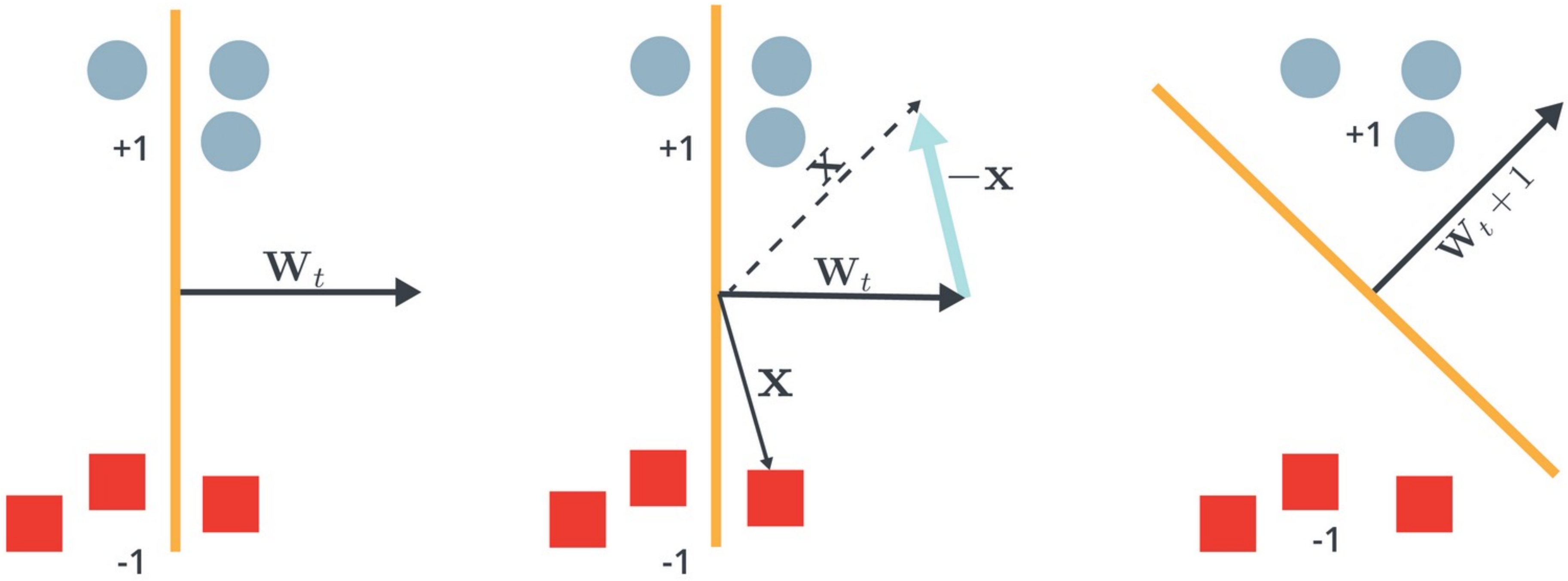




How do we scale the hyperplane?

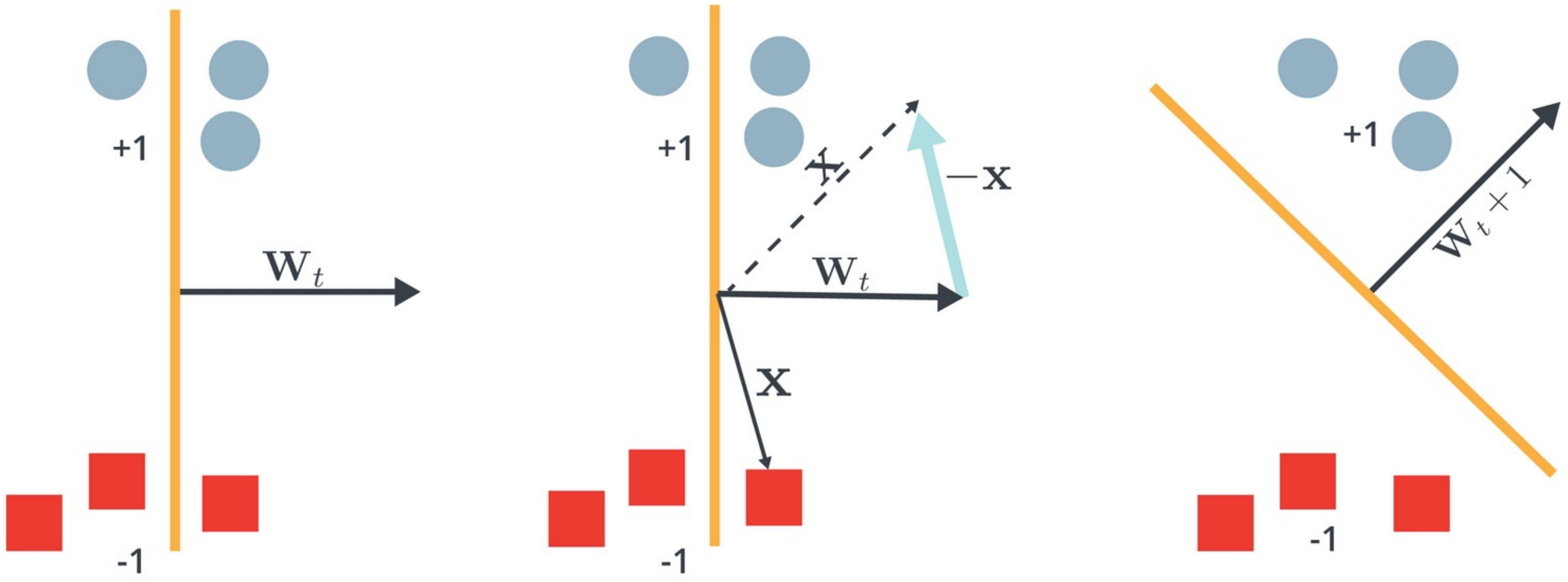
$$Y = x_1 * w_1 + x_2 * w_2 + b_1$$

Perceptron Update – Part One



How would you generalize the picture above into a formula?

Perceptron Update – Part One



$W_{t+1} = W_t + (x * y)$

If we have a scalar and multiply it into an array, what do we get? $(X * y)$

If we have two arrays and add them together what do we get? $W_t + (X * y)$

What is important about the size of the arrays when you are doing an element-wise addition? " + "

Knowledge Check (Do you understand?)

$? = 1x_1 + 3x_2$

$? = 4x_1 - 2x_2$

$? = 5x_1 + 1x_2$

Exercise 1

Consider the following two-point 2D data set:

- Positive class (+1): $(1, 3)$
- Negative class (-1): $(-1, 4)$

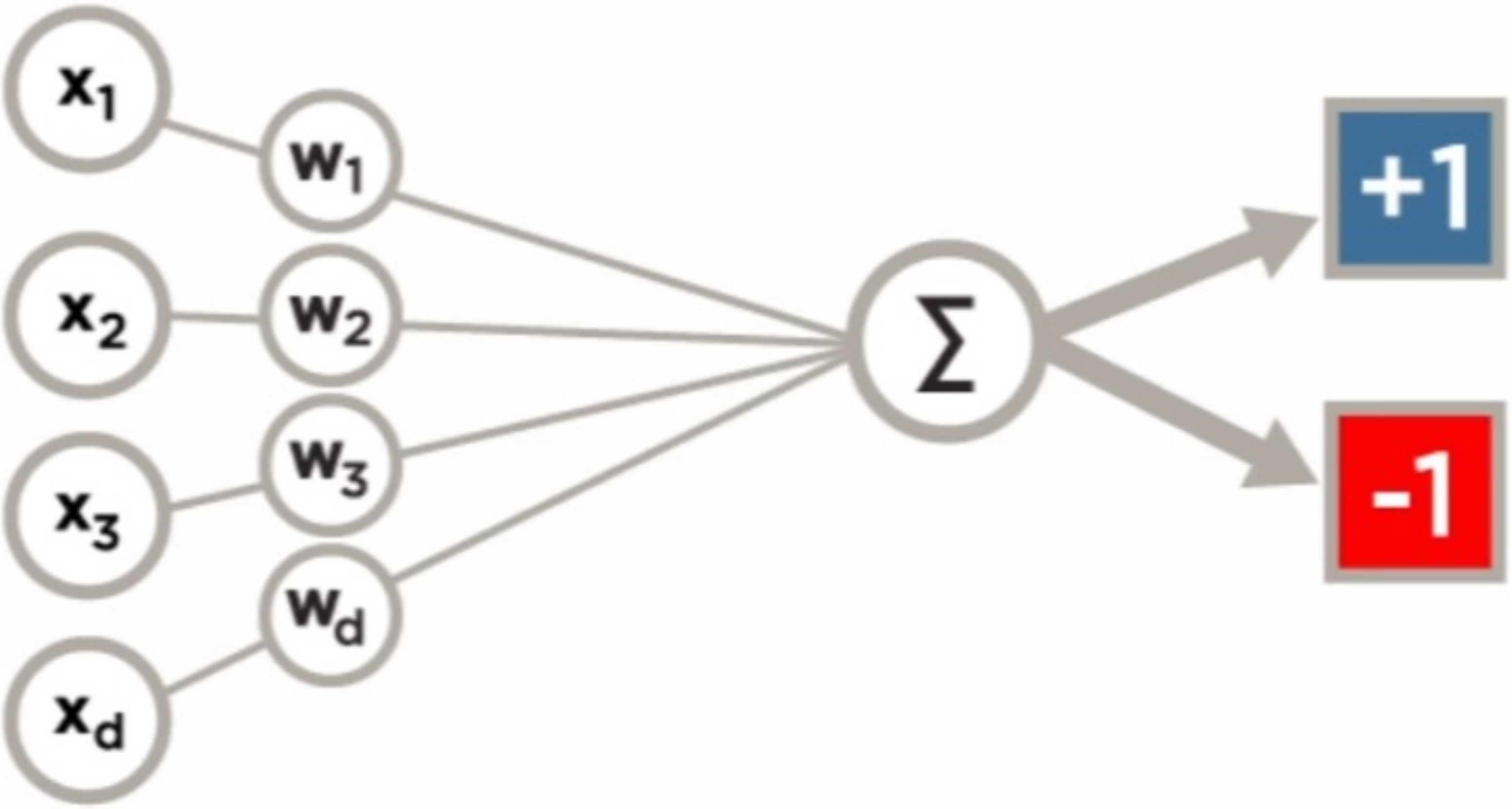
Starting with $\mathbf{w}_0 = (0, 0)$, which is equivalent to a vertical hyperplane as on the previous page, how many updates will you have to perform to $\mathbf{w}$ until convergence? Write down the sequence of each updated $\mathbf{w}_i([\mathbf{w}_1, \mathbf{w}_2 \ldots \mathbf{w}_n])$ by iterating the data points in the order:$[(1,3), (-1,4), (1,3), (-1,4), \ldots]$.

When you think you know the answers, click the images to reveal the solutions

Click to reveal the answer

Answer: There are 5 updates as follows: $[(0,0), (1,3), (2,-1), (3,2), (4,-2), (5,1)]$

Perceptron Foundation



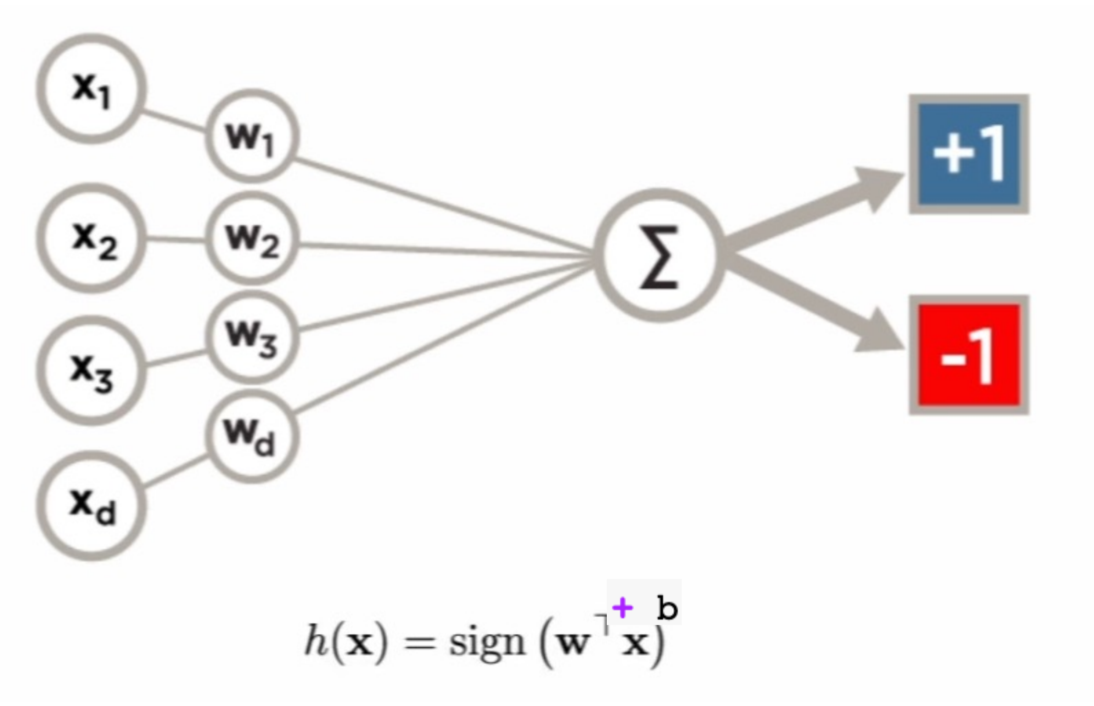
$$h(\mathbf{x}) = \mathrm{sign}\left(\mathbf{w}^\top \mathbf{x} + b\right)$$

Part 2: Perceptron (in Notebook)

Create a random array of row index values (from xs) using np.random.permutation(???).
Then pick the rows out of xs[???] and ys[???] using “for i in random_indexes:”

Initialize $\vec{w} = \vec{0}$	// Initialize $\vec{w}$. $\vec{w} = \vec{0}$ misclassifies everything.
while TRUE **do**	// Keep looping
$m = 0$	// Count the number of misclassifications, m
for $(x_i, y_i) \in D$ **do**	// Loop over each (data, label) pair in the dataset, D
if $y_i(\vec{w}^T \cdot \vec{x_i}$ + b$) \leq 0$ **then**	// If the pair $(\vec{x_i}, y_i)$ is misclassified
$\vec{w} \leftarrow \vec{w} + y\vec{x}$ b += ys[i]	// Update the weight vector $\vec{w}$
$m \leftarrow m + 1$	// Counter the number of misclassification
end if	
end for	
if $m = 0$ **then**	// If the most recent $\vec{w}$ gave 0 misclassifications
break	// Break out of the while-loop
end if	
end while	// Otherwise, keep looping!

Perceptron Part 3



$$\mathbf{w}^\top \mathbf{x}_i + b$$

```
X@w
xs.dot(w)
```

I don’t like the picture on the left.

What are we doing and what do we get?

Neural Networks (Deep Learning) are built on Perceptrons and Restricted Boltzmann Machines

Where Are We Stuck

- What is a Loss Function?
- What is the purpose of a Loss Function?
- Why is Convexity Important to a Loss Function?
- How do you determine the min/max points in a function?
- How do you tell if a min/max point is in fact a minimum or maximum point?
- How do you tell if there is one or multiple minimum or maximum points in a Loss Function?

Where Are We Stuck

Ethan Sasiela – Former Student

$$h\left(w\right) = \log\left(1 + e^{-\left(y_i w^T x_i\right)}\right)$$

to simplify things, let $s = y_i w^T x_i$

restate h in terms of s:

$$h\left(s\right) = \log(1 + e^{-s})$$

Refresher, the derivative of the log of a composite function:

$\frac{d(\ln(g(x)))}{dx} = \frac{1}{g(x)} \cdot g'\left(x\right)$ and the chain rule: $\frac{df(g(x))}{dx} = \frac{df(g)}{dg} \cdot \frac{dg(x)}{dx}$

applying these to $h\left(s\right)$:

$$\frac{dh}{dw} = \frac{1}{1+e^{-s}} \cdot \frac{d\left(1+e^{-s}\right)}{ds} \cdot \frac{ds}{dw}$$

Where Are We Stuck

1) To show loss function $L(w) = \sum_{i=1}^{n} \log\left(1 + exp\left(-y_i w^T x_i\right)\right)$ is strictly convex

- step 1: show that the scaler function $g(s) = \log(1 + e^{-s})$ is strictly convex.

Ethan Sasiela – Former Student

continuing very slowly, expand out the middle term using the sum rule for differentiation:

$$\frac{dh}{dw} = \frac{1}{1+e^{-s}} \cdot \left(\frac{d(1)}{ds} + \frac{d\left(e^{-s}\right)}{ds}\right) \cdot \frac{ds}{dw}$$

derivative of constant is zero and apply the rule for derivative of $exp()$function:

$$\frac{dh}{dw} = \frac{1}{1+e^{-s}} \cdot \left(0 + \frac{d(-s)}{ds} \cdot e^{-s}\right) \cdot \frac{ds}{dw}$$

even I remember this part from AP calc forever ago:

$$\frac{dh}{dw} = \frac{1}{1+e^{-s}} \cdot \left(-1 \cdot e^{-s}\right) \cdot \frac{ds}{dw}$$

simplifying:

$$\frac{dh}{dw} = -\frac{e^{-s}}{1+e^{-s}} \cdot \frac{ds}{dw}$$

Where Are We Stuck

1) To show loss function $L(w) = \sum_{i=1}^{n} \log\left(1 + exp\left(-y_i w^T x_i\right)\right)$ is strictly convex

- step 1: show that the scaler function $g(s) = \log(1 + e^{-s})$ is strictly convex.

Ethan Sasiela – Former Student

to simplify further, multiply by $\frac{e^s}{e^s} = 1$

$$\frac{dh}{dw} = -\frac{e^{-s}}{1+e^{-s}} \cdot \frac{e^s}{e^s} \cdot \frac{ds}{dw} = -\frac{1}{e^s+1} \cdot \frac{ds}{dw} = -\frac{1}{1+e^s} \cdot \frac{ds}{dw}$$

Now, take derivative of s with respect to w:

$$\frac{ds}{dw} = y_i \cdot \left(1w^{T^0}\right) \cdot x_i = y_i x_i$$

Substituting...

$$\frac{dh}{dw} = -\frac{1}{1+e^s} \cdot y_i x_i$$

Where Are We Stuck

1) To show loss function $L(w) = \sum_{i=1}^{n} \log\left(1 + exp\left(-y_i w^T x_i\right)\right)$ is strictly convex

- step 1: show that the scaler function $g(s) = \log(1 + e^{-s})$ is strictly convex.

Lorne Jensen – Former Student

$$\frac{dh}{dw} = \frac{-e^{-s}}{1+e^{-s}} \frac{ds}{dw} = -\frac{1}{1+e^{s}} \frac{ds}{dw} = -\sigma(-s) \frac{ds}{dw} = (\sigma(s) - 1)\, y_i x_i$$

$$\frac{d^2h}{dw^2} = \frac{d}{dw}\left(\frac{dh}{dw}\right) = \frac{ds}{dw}\frac{d}{ds}((\sigma - 1) \cdot y_i \cdot x_i) = (y_i x_i)^2 \cdot \left(\frac{d\sigma}{ds}\right) = (y_i \cdot x_i)^2 \cdot (\sigma \cdot (1 - \sigma))$$

Since $(y_i \cdot x_i)^2 \geq 0$, we just need to show that $\sigma \cdot (1 - \sigma) \geq 0$

for all s to have it be convex:

$$\sigma \cdot (1 - \sigma) = \sigma - \sigma^2 \geq 0$$

$$\sigma \geq \sigma^2$$

$$\frac{1}{1+e^{-s}} \geq \frac{1}{(1+e^{-s})^2}$$

$$1 + e^{-s} \geq 1$$

Where Are We Stuck

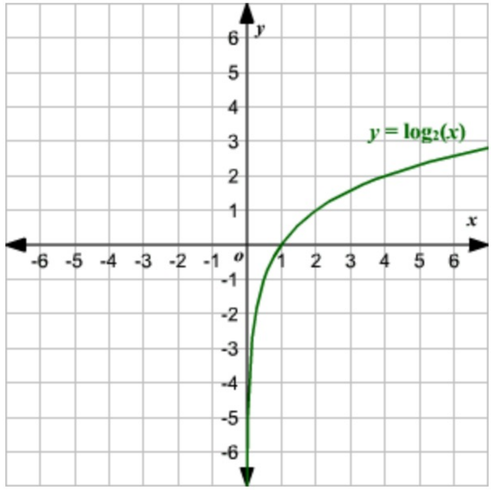
$$\mathbf{w}_{MLE} = \arg\min_{\mathbf{w}} \sum_{i=1}^{n} \log(1 + e^{-y_i \mathbf{w}^T \mathbf{x}_i})$$

Discussion topic:

- Explain why the loss is convex. (Take a look at the second derivative.)
- What function does the loss approximate as $\mathbf{w}^\top \mathbf{x}$ becomes large?

Haiyan Weng
– Former Student

$if\ the\ i^{th}\ label\ is\ correctly\ classified,\ i.e.,\ y_i\ and\ w^T x_i\ have\ the\ same\ sign$
$(both\ positive\ or\ both\ negative),\ then\ \log\left(1 + exp\left(-y_i w^T x_i\right)\right) \sim \log(1) = 0\ as\ w^T x_i\ becomes\ large$



$if\ the\ i^{th}\ label\ is\ misclassified,\ i.e.,\ y_i\ and\ w^T x_i\ have\ opposite\ signs$
$(one\ is\ postive\ and\ the\ other\ one\ is\ negative),\ then\ \log\left(1 + exp\left(-y_i w^T x_i\right)\right) \sim \log$
$\left(exp\left(-y_i w^T x_i\right)\right) = -y_i w^T x_i = abs\left(w^T x_i\right)\ as\ w^T x_i\ becomes\ large$

$so\ as\ w^T x_i\ becomes\ large,\ the\ loss\ \sum_{i=1}^{n} \log\left(1 + exp\left(-y_i w^T x_i\right)\right) \sim \sum_j abs\left(w^T x_j\right),\ where\ j\ are$
$indices\ of\ misclassified\ labels.\ Since\ w^T x_i\ is\ large,\ the\ loss\ function\ will\ be\ large\ when\ there\ are$
$misclassified\ labels\ and\ the\ loss\ function\ will\ be\ close\ to\ 0\ if\ all\ labels\ are\ classified\ correctly.$

Questions