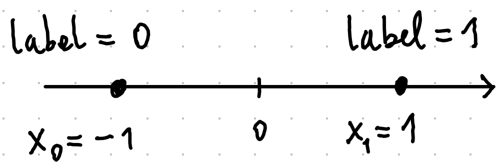
QUESTION: Does a minimum always exist for the cross-entropy loss in classification?

ANSWER: In general, no.

<u>Example</u>: Consider a dataset of $N=2$ points.

features: $\begin{cases} x_0 = -1 \\ x_1 = 1 \end{cases}$ labels: $\begin{cases} y_0 = 0 \\ y_1 = 1 \end{cases}$



$w \in \mathbb{R}$ weight

$b \in \mathbb{R}$ bias

The cost for this binary classification is:

$$L(w,b) = -\sum_{i=0}^{N-1}\Big(y_i \log(\sigma(wx_i+b)) + (1-y_i)\log(1-\sigma(wx_i+b))\Big)$$

$$= \underbrace{-\log(1-\sigma(-w+b))}_{i=0} \underbrace{-\log(\sigma(w+b))}_{i=1}$$

$$= -\log\big((1-\sigma(-w+b))\,\sigma(w+b)\big)$$

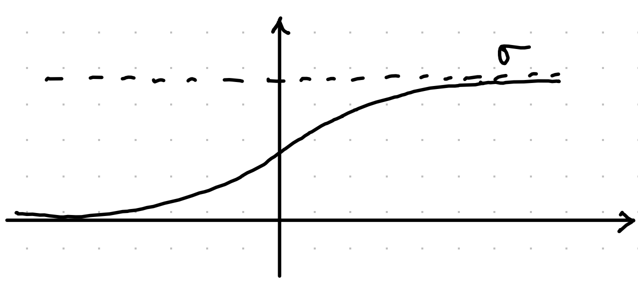
Let us show that $L(w,b)$ does not have a minimum.

Since $0 < \sigma < 1$, we have
$-\log((1-\sigma)\sigma) > 0$.

Hence, the value 0 is not reached for any w and b.

However, choosing $b=0$ and letting $w \to +\infty$, we have

$$\lim_{w\to+\infty} \sigma(w) = 1 \quad , \quad \lim_{w\to+\infty} \sigma(-w) = 0$$



This implies

$$\lim_{w \to +\infty} (1-\sigma(-w))\sigma(w) = 1$$

and thus

$$\lim_{w \to +\infty} L(w,0) = 0.$$

It means that $\inf_{w,b} L(w,b) = 0$,

but it is not reached.

Intuition: The dataset is very discriminatory. From the dataset, it seems that the probability distribution behind data is

$$P(-1,0) = 1 \qquad P(-1,1) = 0$$

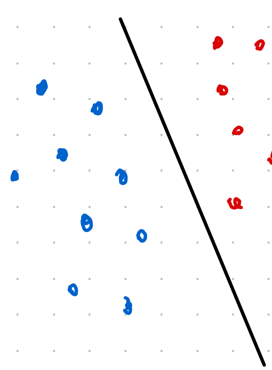
$$P(+1,1) = 1 \qquad P(+1,0) = 0$$

This can be approximated using the logistic regression.

$$Q(-1, 0; w, b) = \sigma(-w+b)\,\mathbb{P}(X=-1)$$
$$Q(-1, 1; w, b) = (1-\sigma(-w+b))\,\mathbb{P}(X=-1)$$
$$Q(+1, 0; w, b) = (1-\sigma(w+b))\,\mathbb{P}(X=1)$$
$$Q(+1, 1; w, b) = \sigma(w+b)\,\mathbb{P}(X=1)$$

We can approximate the probabilities $\mathbb{P}$ with $w \longrightarrow +\infty$.

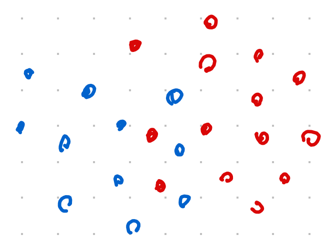
The same reasoning works in a more general case when data can be linearly separated.



In some sense, the logistic regression model can "potentially" model perfectly the actual probability distribution, but in a limit $w \to +\infty$.

<u>Remark</u>: Even if the minimum does not exist, a numerical optimization algorithm may allow to approach the infimum value.

In the next example we show that, when the dataset is not linearly separated, the minimum exists.

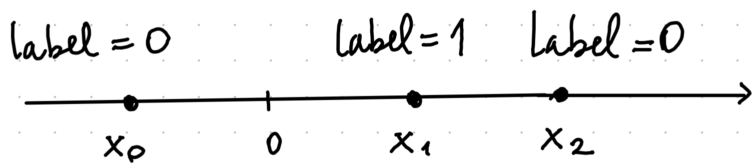


Example:

$x_0 = -1 \qquad y_0 = 0$

$x_1 = 1 \qquad y_1 = 1 \qquad N = 3$

$x_2 = 2 \qquad y_2 = 0$



$$L(w, b) = -\sum_{i=0}^{N-1} \Big(y_i \log(\sigma(wx_i + b)) + (1-y_i)\log(1-\sigma(wx_i+b)) \Big)$$

$$= \underbrace{-\log(1-\sigma(-w+b))}_{i=0} \underbrace{-\log(\sigma(w+b))}_{i=1} + \underbrace{-\log(1-\sigma(2w+b))}_{i=2} =$$

$$= -\log\left(1 - \frac{1}{e^{w-b}+1}\right) - \log\left(\frac{1}{e^{-w-b}+1}\right) +$$

$$-\log\left(1 - \frac{1}{e^{-2w-b}+1}\right) =$$

$$= -\log\left(\frac{e^{w-b}}{e^{w-b}+1}\right) - \log\left(\frac{1}{e^{-w-b}+1}\right) +$$

$$-\log\left(\frac{e^{-2w-b}}{e^{-2w-b}+1}\right) =$$

$$= -\log\left(\frac{1}{1+e^{-w+b}}\right) - \log\left(\frac{1}{1+e^{-w-b}}\right) +$$

$$-\log\left(\frac{1}{1+e^{2w+b}}\right) =$$

$$= \log\left(1+e^{-w+b}\right) + \log\left(1+e^{-w-b}\right) + \log\left(1+e^{2w+b}\right)$$

We exploit the "misclassified" point.

Consider a sequence (w_k, b_k) such that $|w_k| + |b_k| \to +\infty$.

In the following cases, we assume $|w_k|+|b_k|$ is large.

We want to show that $L(w_k, b_k)$ is large. This allows us to search for a minimum in a bounded region, which exists by Weierstrass' theorem.

Assume that

$$\log(1+e^{-w_k+b_k}) + \log(1+e^{-w_k-b_k}) + \log(1+e^{2w_k+b_k})$$

is bounded from above. (We know $L > 0$)

Since $t \mapsto \log(1+e^t)$ is increasing, it means that

$$\begin{cases} -w_k + b_k \leq M & (1) \\ -w_k - b_k \leq M & (2) \\ 2w_k + b_k \leq M & (3) \end{cases}$$

for some $M > 0$ independent of k.

Summing $(1)+(2)$:

$$-2w_k \leq 2M \Rightarrow -w_k \leq M \Rightarrow -M \leq w_k$$

Summing $(2)+(3)$

$$w_k \leq 2M$$

Hence $|w_k| \leq 2M$.

From (1), $b_k \leq M + w_k \leq 3M$

From (2), $-b_k \leq M + w_k \leq 3M$

It follows that $|w_k|+|b_k| \leq 4M$ is bounded, contradicting the fact that it is large.

In fact, we can say more precisely that

$$\lim_{\|(w_k,b_k)\|\to+\infty} L(w_k,b_k) = +\infty.$$

To show this, assume that

$$|w_k|+|b_k| > 4M$$

Then one of the three inequalities

$$-w_k+b_k \leq M \quad (1)$$
$$-w_k-b_k \leq M \quad (2)$$
$$2w_k+b_k \leq M \quad (3)$$

must be violated (for, otherwise, we have shown that $|w_k|+|b_k| \leq 4M$)

If (1) is violated, $-w_k+b_k > M \Rightarrow$
$\Rightarrow \log(1+e^{-w_k+b_k}) > \log(1+e^M)$

If (2) is violated,

$$\log(1+e^{-w_k - b_k}) > \log(1+e^{M})$$

If (3) is violated,

$$\log(1+e^{2w_k + b_k}) > \log(1+e^{M})$$

In either case,

$$L(w_k, b_k) > \underbrace{\log(1+e^{M})}$$

Since this term $\longrightarrow +\infty$ as $M \to +\infty$, we have shown coercivity of L.