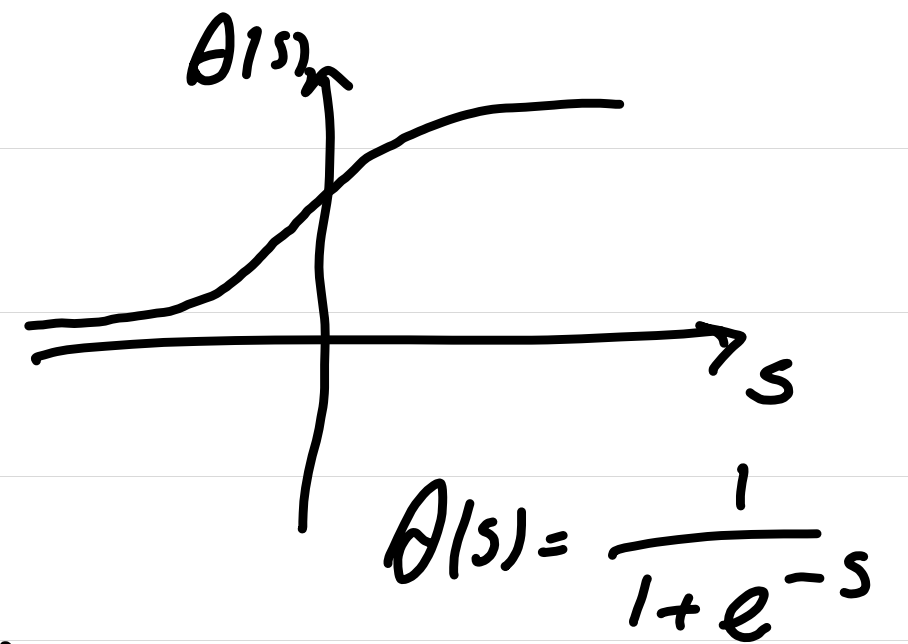


Recap: Logistic Regression

$$\hat{P}_{\underline{w}}(y_n | \underline{x}_n) = A(y_n \underline{w}^T \underline{x}_n)$$



$$e_n(\underline{w}) = -\log \hat{P}_{\underline{w}}(y_n | \underline{x}_n) = \log(1 + e^{-y_n \underline{w}^T \underline{x}_n})$$

Today: We will see why we use this loss function?
(Mathematically speaking)

We want to
predict with
randomness

$$\mathcal{J}(\underline{w}): \hat{P}_{\underline{w}}(y|\underline{x}) = \frac{1}{1 + e^{-y\underline{w}^T \underline{x}}}$$

We saw why it makes
sense

We need a
 $E_{in}(\underline{w})$

We use log-loss
 $\ell_n(\underline{w}) = -\log \hat{P}_{\underline{w}}(y_n|\underline{x}_n)$

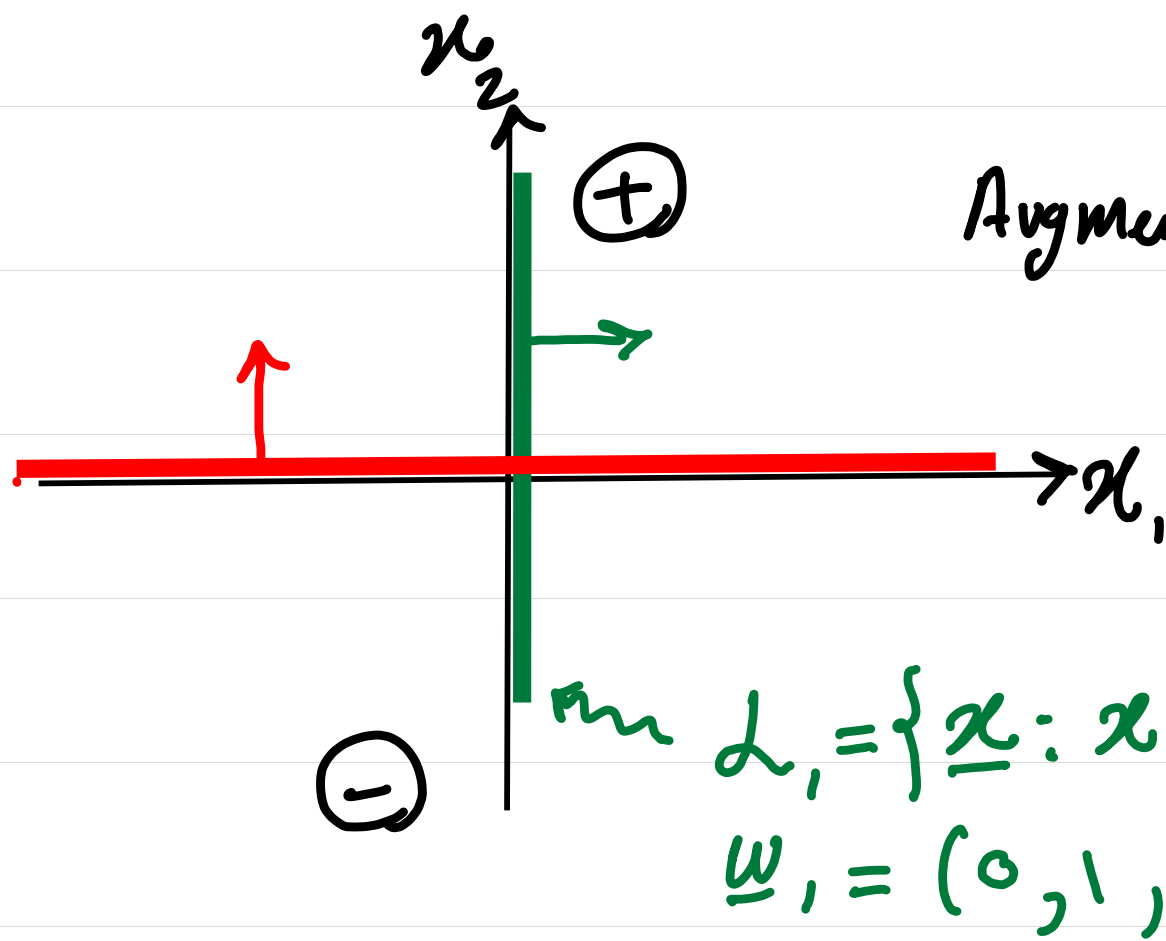
Today: why does
log-loss make
sense?

It can distinguish
better different
hyperplanes:
Numerical examples

Benefits Over Linear Classification

E.g.: $d=2$
 $N=2$

$\mathcal{L}_2 = \{\underline{x} : x_2 = 0\}$
 $\underline{w}_2 = (0, 0, 1)$



■ For linear classification, which line is better?

$(e_n(\underline{w}) = \mathbb{1}(y_n \neq \text{Sign}(\underline{w}^T \underline{x}_n)))$: They are the same

$E_{in}(\underline{w}_1) = E_{in}(\underline{w}_2) = 0$

■ Intuitively, which line is better?

□ \mathcal{L}_1 is preferred. It gives more room

For logistic regression, which line is better?

$$\square E_{in}(\underline{w}_1) = \frac{1}{2} [\log(1 + e^{-y_1 \underline{w}_1^T x_1}) + \log(1 + e^{-y_2 \underline{w}_1^T x_2})]$$

$$= \frac{1}{2} [\log(1 + e^{(-1)(0.001)}) + \log(1 + e^{(+1)(-0.001})] \approx 0.693$$

$$\square E_{in}(\underline{w}_2) = \frac{1}{2} [\log(1 + e^{-y_1 \underline{w}_2^T x_1}) + \log(1 + e^{-y_2 \underline{w}_2^T x_2})]$$

$$= \frac{1}{2} [\log(1 + e^{(-1)(10)}) + \log(1 + e^{(+1)(-10})] \approx 5 \times 10^{-5}$$

So,

$\square \mathcal{L}_1$ is preferred

☒ \mathcal{L}_2 is preferred

■ What about $\underline{w}_3 = (0, 0, 100)$?

□ This is the same line as \underline{w}_2 .

□ What about $E_{in}(\underline{w}_3)$? $\frac{1}{2} (\log(1 + e^{-100}) + \log(1 + e^{100}))$

- $E_{in}(\underline{w}_3)$ is much lower!

- But \underline{w}_3 and \underline{w}_2 are the same lines.

■ We can fix the norm of $\|\underline{w}\|_2$ to be 1.

□ But this constraint makes the optimization more difficult

challenging $\rightsquigarrow \min_{\underline{w}} E_{in}(\underline{w})$
s.t. $\|\underline{w}\| = 1$

■ Typically, we regularize logistic regression

$$\min_{\underline{w}} E_{in}(\underline{w}) + \lambda \|\underline{w}\|_2^2$$

■ So far, we have seen numerical demonstration why log-loss makes sense.

■ Time for more rigorous math.

□ There are two mathematical explanation for log-loss

1. When we minimize log-loss, we maximize the likelihood

2. When we minimize log-loss, we minimize cross-entropy

We want to
predict with
randomness

$$\mathcal{J}(\underline{w}): \hat{P}_{\underline{w}}(y|\underline{x}) = \frac{1}{1 + e^{-y\underline{w}^T \underline{x}}}$$

We saw why it makes
sense

We need a
 $E_{in}(\underline{w})$

We use log-loss
 $\ell_n(\underline{w}) = -\log \hat{P}_{\underline{w}}(y_n|\underline{x}_n)$

Today: why does
log-loss make
sense?

It can distinguish
better different
hyperplanes:
Numerical examples

Maximum likelihood
interpretation:
when we minimize log-loss,
we maximize likelihood

Maximum Likelihood Interpretation:

■ Let $D = \{(\underline{x}_1, y_1), \dots, (\underline{x}_N, y_N)\}$ be the observed datapoint.

■ Consider $P(y_1, y_2, \dots, y_N | \underline{x}_1, \underline{x}_2, \dots, \underline{x}_N)$

$$= P[1^{\text{st}} \text{ label is } y_1, 2^{\text{nd}} \text{ label is } y_2, \dots | 1^{\text{st}} \text{ example} = \underline{x}_1, 2^{\text{nd}} \text{ example} = \underline{x}_2, \dots]$$

= "likelihood of observing this particular labels y_1, \dots, y_N given datapoint $\underline{x}_1, \dots, \underline{x}_N$ "

- This is the joint distribution.
- Assuming having I.I.D. examples, independent and identically distributed

$$P(y_1, y_2, \dots, y_N | x_1, x_2, \dots, x_N) = \prod_{n=1}^N P(y_n | x_n)$$

(we want a $\hat{P}_{\underline{w}} \in \mathcal{H}$ s.t.)

$$\approx \prod_{n=1}^N \hat{P}_{\underline{w}}(y_n | x_n)$$

unknown target probability distribution

We want to find \underline{w} that maximizes $\prod_{n=1}^N \hat{P}_{\underline{w}}(y_n | x_n)$ Likelihood

$$\Leftrightarrow \underset{\underline{w}}{\text{maximize}} \quad \log \prod_{n=1}^N \hat{P}_{\underline{w}}(y_n | x_n)$$

$$\Leftrightarrow \underset{\underline{w}}{\text{max}} \quad \frac{1}{N} \sum_{n=1}^N \log \hat{P}_{\underline{w}}(y_n | x_n) \Leftrightarrow \underset{\underline{w}}{\text{min}} \quad \frac{1}{N} \sum_{n=1}^N \underbrace{-\log \hat{P}_{\underline{w}}(y_n | x_n)}_{e_n(\underline{w})}$$

$$\Leftrightarrow \underset{\underline{w}}{\text{min}} \quad E_{\text{in}}(\underline{w})$$

We want to
predict with
randomness

$$\mathcal{J}(\underline{w}): \hat{P}_{\underline{w}}(y|\underline{x}) = \frac{1}{1 + e^{-y\underline{w}^T \underline{x}}}$$

We saw why it makes
sense

We need a
 $E_{in}(\underline{w})$

We use log-loss
 $\ell_n(\underline{w}) = -\log \hat{P}_{\underline{w}}(y_n|\underline{x}_n)$

Today: why does
log-loss make
sense?

It can distinguish
better different
hyperplanes:
Numerical examples

Maximum likelihood
interpretation:
when we minimize log-loss,
we maximize likelihood

Cross-entropy interpretation:
When we minimize log-loss,
we minimize Cross-entropy

Cross-Entropy Interpretation

Defn: Suppose P and Q are two distributions over the same sample space S

Sample space S : The set of all values a r.v. can take

E.g.: Poisson R.V. with mean λ :
 $S = \{0, 1, \dots\}$
distribution: $P(k) = \frac{\lambda^k}{k!} e^{-\lambda}$

The cross-entropy b/w P and Q is

$$CE(P, Q) = - \sum_{k \in S} P(k) \log Q(k)$$

(measures the difference b/w P and Q)

■ For the n^{th} example, Consider the "true" distribution
 $P_n = (P[y_n = +1], P[y_n = -1])$ (distribution of the true labels)

$$= \begin{cases} (1, 0), & \text{if } y_n = 1 \\ (0, 1), & \text{if } y_n = -1 \end{cases}$$

$$= (\mathbb{1}(y_n = +1), \mathbb{1}(y_n = -1))$$

■ For n^{th} example, Consider the estimated distribution
 $Q_n = (\hat{P}_{\underline{w}}(1|x_n), \hat{P}_{\underline{w}}(-1|x_n))$ "estimated distribution of y_n
given example x_n "

■ The "closer" Q_n 's are to P_n 's, the better it is.

■ what does "closeness" mean? Cross-Entropy

$$\square CE(P_n, Q_n) = - \left(\mathbb{1}(y_n=+1) \log \hat{P}_{\underline{w}}(1|x_n) + \mathbb{1}(y_n=-1) \log \hat{P}_{\underline{w}}(-1|x_n) \right)$$
$$= - \log \hat{P}_{\underline{w}}(y_n|x_n) = e_n(\underline{w})$$

$\min E_{in}(\underline{w}) \iff \min$ average "distance" b/w P_n 's and Q_n 's.
 $\frac{1}{N} \sum_{n=1}^N CE(P_n, Q_n)$

■ Hopefully we are all convinced that log-loss is a reasonable loss function.

■ But, what is our learning algorithm to find the \underline{w} that minimizes $E_{in}(\underline{w})$?

□ Next time. Gradient descent.