Lec 8 Log-loss function

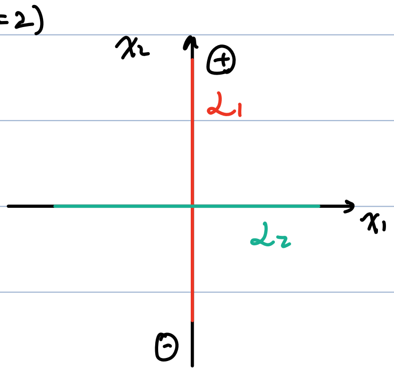
↳ Main goal:

Why use $e_n(\underline{w}) = -\log \hat{P}_{\underline{w}}(y_n \mid \underline{x}_n)$

$= \log(1 + e^{-y_n \underline{w}^T \underline{x}_n})$

1. Benefit over linear classification:

↳ e.g. $(d=2)$



$N = 2$

$\underline{x}_1 = (0.001, 10)$ $\quad y_1 = +1$

$\underline{x}_2 = (-0.001, -10)$ $\quad y_2 = -1$

$L_1 : x_1 = 0$ $\quad \underline{w}_1 = (0, 1, 0)$

$L_2 : x_2 = 0$ $\quad \underline{w}_2 = (0, 0, 1)$

w_{x_0} $\quad w_{x_1}$ $\quad w_{x_2}$

• Linear classification:

$e_n(\underline{w}) = 1(y_n \neq \text{sign}(\underline{w}^T \underline{x}_n))$

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

It does not tell us which line is better.

• Logistic Regression:

y_1 $\qquad = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} [0 \quad 0.001 \quad 10]$ $\qquad y_2$

$E_{in}(\underline{w}_1) = \frac{1}{2}\left[\log(1 + e^{-(1)\cdot \underline{w}_1^T \underline{x}_1}) + \log(1 + e^{-(-1)\cdot \underline{w}_1^T \underline{x}_2})\right]$

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

≈ 0.693

$E_{in}(\underline{w}_2) = \frac{1}{2}\left[\log(1 + e^{-\underline{w}_2^T \underline{x}_1}) + \log(1 + e^{+\underline{w}_2^T \underline{x}_2})\right]$

$= \frac{1}{2}\left[\log(1 + e^{-10}) + \log(1 + e^{-10})\right]$

$\approx 5 \times 10^{-5}$

$\Rightarrow L_2$ is clearly preferred.

• Note: what if $\underline{w}_2 = (0, 0, 100)$?

i.e. $0(x_0) + 0(x_1) + 100(x_2) = 0.$

It's the same line as L_2

But the loss $E_{in}(\underline{w})$ is much lower

$\Rightarrow$ Logistic regression typically requires some regularization.

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

2. Maximum likelihood viewpoint

↳ set up:

Let $\{(x_1, y_1), (x_2, y_2) \cdots (x_N, y_N)\}$ be the trainig dataset

consider $P(y_1, y_2 \cdots y_N \mid \underline{x}_1, \underline{x}_2, \cdots \underline{x}_N) \triangleq \Pr[1^{st}$ label is $y_1 \cdots \mid 1^{st}$ example is $x_1 \cdots]$

$$= \prod_{n=1}^{N} P(y_n \mid \underline{x}_n)$$

by Hypothesis

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

↳ Maximum Likelihood approach:

Want to find $\underline{w} \in \mathbb{R}^{d+1}$ that maximize $P(y_1, y_2 \cdots y_N \mid x_1, x_2 \cdots x_N)$

$$\Leftrightarrow \text{Maximize } \frac{1}{N} \log \prod_{n=1}^{N} \hat{P}_{\underline{w}}(y_n \mid \underline{x}_n)$$

$$= \frac{1}{N} \sum_{n=1}^{N} \log \hat{P}_{\underline{w}}(y_n \mid \underline{x}_n)$$

$$\Leftrightarrow \text{Minimize } \frac{1}{N} \sum_{n=1}^{N} -\log \hat{P}_{\underline{w}}(y_n \mid \underline{x}_n)$$

$$= \frac{1}{N} \sum_{n=1}^{N} e_n(\underline{w})$$

$$= E_{in}(\underline{w})$$

3. Cross-Entropy Viewpoint:

↳ If we write out the training error:

$$E_{in}(\underline{w}) = -\frac{1}{N}\sum_{n=1}^{N}\Big[\, 1(y_n=+1)\log \hat{P}_{\underline{w}}(1\,|\,\underline{x}_n) \qquad \text{// +ve case}$$

$$+\, 1(y_n=-1)\log \hat{P}_{\underline{w}}(-1\,|\,\underline{x}_n)\Big] \qquad \text{// -ve case}$$

This is exactly the form of cross-entropy

↳ Def:

Suppose P and Q are two prob. distributions over $\mathcal{X} = \{x_1, \cdots, x_m\}$

// e.g. Poission with mean α.

The set of possible value $\mathcal{X} = \{0, 1, 2, \cdots\}$

$$P(x) = \frac{\alpha^x}{x!} e^{-\alpha}$$

The cross-entropy between P and Q (measurement of distance) is:

$$\boxed{CE(P, Q) = -\sum_{i=1}^{M} P(x_i)\log Q(x_i)}$$

↳ Relation between CE and $E_{in}(\underline{w})$:

For the nth example, consider the distribution

$$P_n = (\Pr\{y_n=+1\},\ \Pr\{y_n=-1\}) \qquad \text{// true prob. distribution for } y_n$$

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

$$= (1(y_n=+1),\ 1(y_n=-1))$$

Let $Q_n = (\hat{P}_{\underline{w}}(+1\,|\,\underline{x}_n),\ \hat{P}_{\underline{w}}(-1\,|\,\underline{x}_n))$ // Our estimate for prob. distrb. of y_n given $\underline{x}_n$

$$\Rightarrow \boxed{E_{in}(\underline{w}) = \frac{1}{N}\sum_{n=1}^{N} CE(P_n, Q_n)}$$

∴ Minimizing $E_{in}(\underline{w})$ = Minimizing distance between P_n and Q_n

true distrb of y_n estimate distrb. of y_n