Logistic Regression

1. Linear Classification

Data Vector $\underline{x} \in \mathbb{R}^{d+1}$

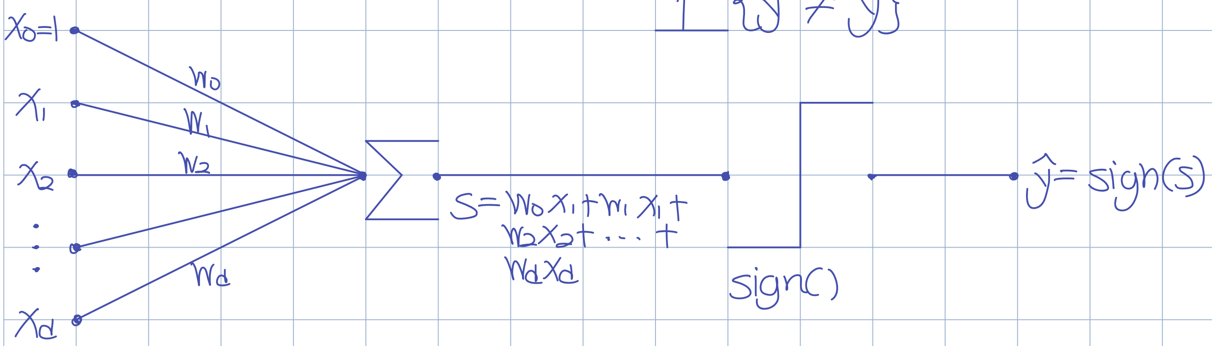
Label $y \in \{-1, +1\}$

Prediction $\hat{y} = \text{sign}\underline{(w^T x)}$

$D = \{(\underline{x_1, y_1}), (\underline{x_2, y_2}), \ldots (\underline{x_N, y_N})\}$

Loss Function

$\underline{1}\{\hat{y} \neq y\}$



2. Linear Regression

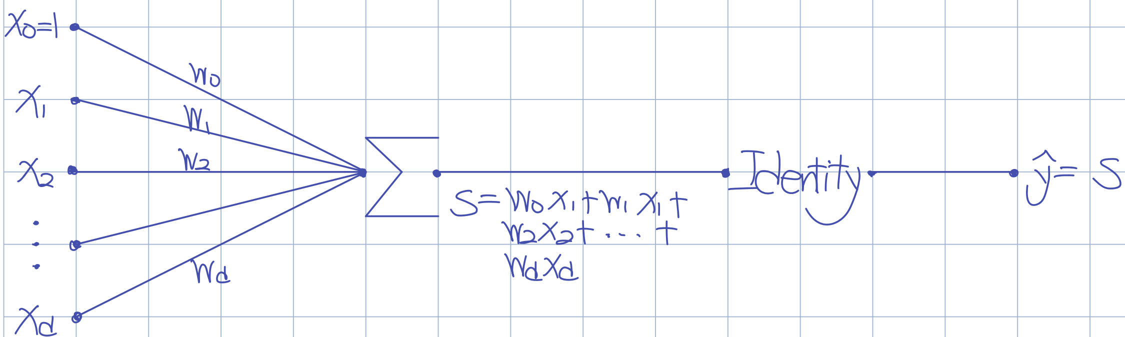
Data Vector $\underline{x} \in \mathbb{R}^{d+1}$

Label $y \in \mathbb{R}$

$\hat{y} = \underline{w^T x}$

Loss Function

$(y - \hat{y})^2$



3. Logistic Regression

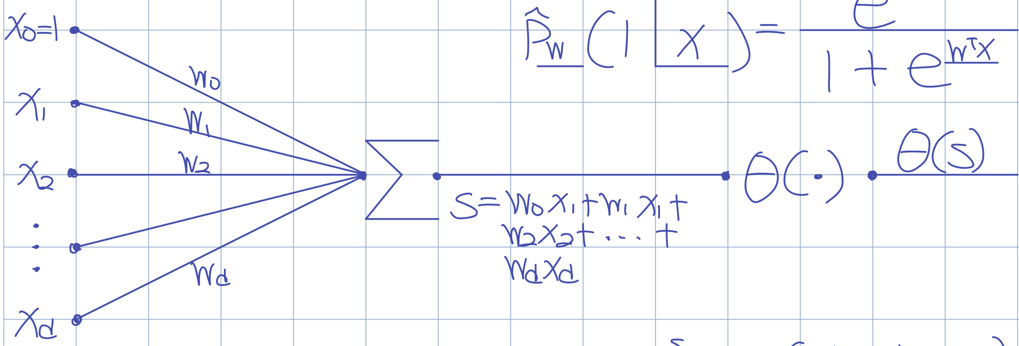
Data Vector $\underline{x} \in \mathbb{R}^{d+1}$

Label $y \in \{+1, -1\}$

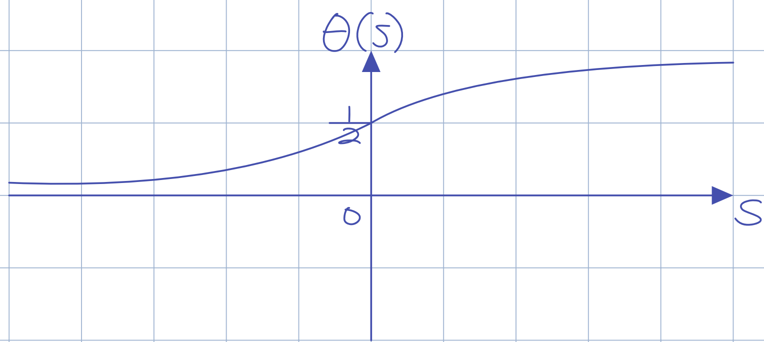
$$\Pr(y = 1 \mid x) = \theta(s) = \theta(w^T x) = \frac{e^{w^T x}}{1 + e^{w^T x}}$$

Same

$$\hat{P}_w(1 \mid x) = \frac{e^{w^T x}}{1 + e^{w^T x}}$$



$$\theta(s) = \frac{e^s}{1 + e^s} \quad \text{(Non Linear)}$$



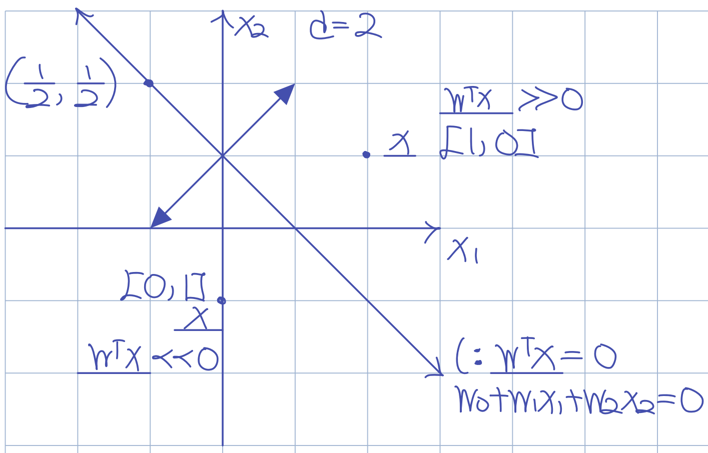
Classification Problem

Sigmoid Function

$$\hat{P}_w(-1 \mid x) = 1 - \hat{P}_w(1 \mid x) = 1 - \frac{e^{w^T x}}{1 + e^{w^T x}}$$

$$= \frac{1}{1 + e^{w^T x}} = \frac{e^{-w^T x}}{1 + e^{-w^T x}}$$

$$\hat{P}_w(y \mid x) = \frac{e^{y w^T x}}{1 + e^{y w^T x}}, \quad y \in \{-1, +1\}$$



Training Set

$D = \{(\underline{x_1}, y_1), \ldots, (\underline{x_n}, y_n)\}$

$\underline{x_i} \in \mathbb{R}^{d+1}, \; y_i \in \{-1, +1\}$

Model Parameter: $\underline{w} \in \mathbb{R}^{d+1}$

Given $(\underline{x}, y)$

Output: $[\hat{P}_{\underline{w}}(1 \mid \underline{x}), \hat{P}_{\underline{w}}(-1 \mid \underline{x})]$

$$\hat{P}_{\underline{w}}(y \mid \underline{x}) = \frac{e^{y\underline{w}^T\underline{x}}}{1+e^{y\underline{w}^T\underline{x}}}, \quad y \in \{-1, +1\}$$

Loss Function

Log Loss Function

$-\log \hat{P}_{\underline{w}}(y \mid \underline{x})$

Example, given $\underline{x}$ and $\underline{w}$ output probability vector of logistic regression is $[0.8, 0.2]$

$= [\hat{P}_{\underline{w}}(1 \mid \underline{x}), \hat{P}_{\underline{w}}(-1 \mid \underline{x})]$

If $y = 1$, loss $= -\log 0.8 \simeq 0.22$

If $y = -1$, loss $= -\log 0.2 \simeq 1.61$

Output $= [0.999, 0.001]$

If $y=+1$, loss $= -\log 0.999 = 10^{-4}$

If $y=-1$, loss $= -\log 0.001 = 10$

$$e_n(\underline{w}) = -\log \hat{P}_{\underline{w}}(y_n \mid \underline{x_n})$$

loss on datapoint $(\underline{x_n}, y_n)$

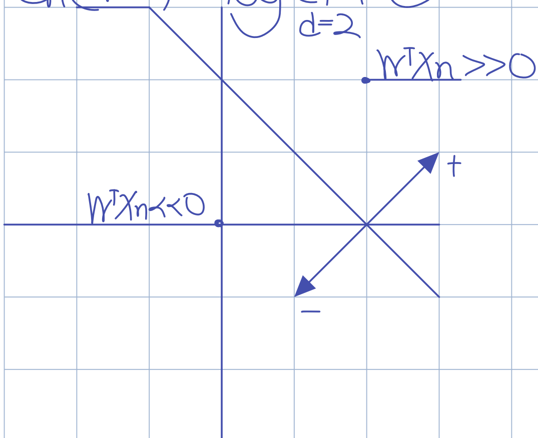
$$= -\log \frac{e^{y_n \underline{w^T x_n}}}{1 + e^{y_n \underline{w^T x_n}}}$$

$$= -\log \frac{1}{1 + e^{-y_n \underline{w^T x_n}}}$$

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

Logarithm base e

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



y_n	$y_n \underline{w^T x_n}$	$e_n(\underline{w})$
+1	$>> 0$	≈ 0
+1	$<< 0$	Large
−1	$>> 0$	≈ 0
−1	$<< 0$	Large

$$E_{in}(\underline{w}) = \frac{1}{N} \sum_{n=1}^{N} e_n(\underline{w}) = \frac{1}{N} \sum_{n=1}^{N} \log(1 + e^{-y_n \underline{w^T x_n}})$$

Training Phase

$\underline{w} = \underset{\underline{w} \in \mathbb{R}^{d+1}}{\operatorname{argmin}} E_{in}(\underline{w})$

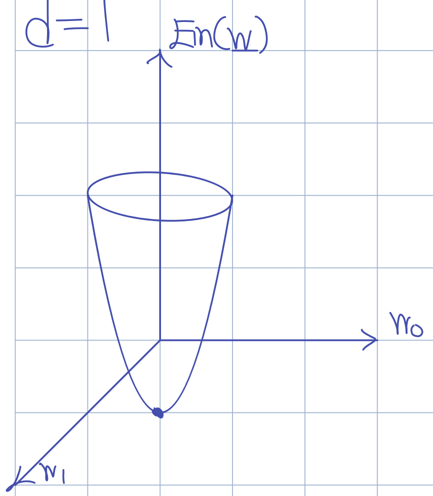
Recall

Linear Regression

$$E_{in}(\underline{w}) = \frac{1}{N}\sum_{n=1}^{N}(y_n - \underline{w^T x_n})^2$$

$E_{in}(\underline{w})$ is convex in $\underline{w}$

$d = 1$

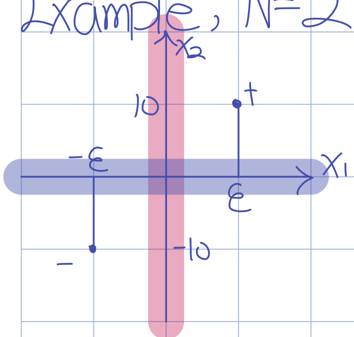


$$\nabla_{\underline{w}} E_{in}(\underline{w}) = 0$$

$\downarrow$

Equation In $\underline{w}$

Example, $N=2$



$\underline{x_1} = (1, \varepsilon, 10)$, $y_1 = +1$

$\underline{x_2} = (1, -\varepsilon, -10)$, $y_2 = -1$

$\varepsilon = 10^{-4}$

Classifier : $x_1 = 0$

$\underline{w} = (0, 1, 0)$

Classifier : $x_2 = 0$ (Preferred)

$\underline{w} = (0, 0, 1)$

Compute $E_{in}(\underline{w})$ for classifier 1 and 2

Classifier 1

$$E_{in}(\underline{w}) = \frac{1}{2}\left(\log(1+e^{-y_1 \underline{w^T x_1}}) + \log(1+e^{-y_n \underline{w^T x_2}})\right)$$

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

$$= \log(1+e^{-\varepsilon})$$

$$\approx 0.693$$

Classifier 2

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

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

$$\approx 5 \times 10^{-4}$$

$\underline{w_1} = (0, 0, 10)$

$\underline{w_1^T x} = 0$

$10x_2 = 0, \; x_2 = 0$

$E_{in}(w_1) < 5 \times 10^{-4}$

$$\underset{\|\underline{w}\|}{\operatorname{argmin}} \left(E_{in}(\underline{w}) + \lambda \|\underline{w}\|^2 \right)$$

Regularized Loss Function

Maximum Likelihood Viewpoint

Training set

$D = \{(\underline{x_1}, y_1), (\underline{x_2}, y_2), \ldots, (\underline{x_N}, y_N)\}$

Pr(label Sequence | Data Vector Sequence)

$= Pr(y_1 y_2 \ldots y_N \mid \underline{x_1 x_2 \ldots x_n})$

Select a model that maximizes probability assigned to label sequence given data vector sequence

Assume that all data samples are generated independently of one another

$Pr(y_1 y_2 \ldots y_n \mid \underline{x_1 x_2 \ldots x_n})$

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

Probability assign to data vector $\underline{x_n}$

Max-Likelihood Objective

Select $\underline{w}$ that maximizes

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

$$\underline{w}^* = \underset{\underline{w} \in \mathbb{R}^{d+1}}{\operatorname{argmax}} \prod_{n=1}^{N} \hat{P}_{\underline{w}}(y_n \mid x_n)$$

$$= \underset{\underline{w} \in \mathbb{R}^{d+1}}{\operatorname{argmax}} \log\left(\prod_{n=1}^{N} \hat{P}_{\underline{w}}(y_n \mid x_n)\right)$$

$$= \underset{\underline{w} \in \mathbb{R}^{d+1}}{\operatorname{argmax}} \sum_{n=1}^{N} \log \hat{P}_{\underline{w}}(y_n \mid x_n)$$

$$= \underset{\underline{w} \in \mathbb{R}^{d+1}}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1}^{N} -\log \hat{P}_{\underline{w}}(y_n \mid x_n)$$

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

$$= \underset{\underline{w} \in \mathbb{R}^{d+1}}{\operatorname{argmin}} E_{in}(\underline{w})$$

Cross Entropy Viewpoint

$S = \{S_1, S_2, S_3, \dots, S_M\}$ be a discrete alphabet

Let $P = (p(S_1), p(S_2), \dots, p(S_M))$

$Q = (q(S_1), q(S_2), \dots, q(S_M))$

be two probability vectors over S

$$CE(P, Q) = -\sum_{i=1}^{M} p(S_i) \log q(S_i)$$

Log Loss Function Can be viewed as a cross entropy

$$e_n(\underline{w}) = -\log \hat{P}_{\underline{w}}(y_n \mid x_n)$$

$$e_n(\underline{w}) = -\Big\{ 1\{y_n = +1\} \log \hat{P}_{\underline{w}}(1 \mid \underline{x_n}) +$$

$$1\{y_n = -1\} \log \hat{P}_{\underline{w}}(-1 \mid \underline{x_n}) \Big\}$$

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

$$Q_n = (\hat{P}_{\underline{w}}(1 \mid \underline{x_n}), \hat{P}_{\underline{w}}(-1 \mid \underline{x_n}))$$

$$\left.\begin{array}{l} y_n = +1 \\ P_n = (1, 0) \\ y_n = -1 \\ P_n = (0, 1) \end{array}\right\} \begin{array}{l} \text{Ideal Output for } (\underline{x_n}, y_n) \\ \text{Atomic } P_n \end{array}$$

$Q \equiv$ Output generated by our model

$e_n(\underline{w}) = CE(P_n, Q_n)$ $\underline{P_n} = (P_n^{(1)}, P_n^{(2)})$

Knowledge Distillation