1 Loss Functions

1.1 Hinge Loss

Let $g(x)$ be a classifier that defined by a score function $f(x)$

$$g(x) = \begin{cases} 1 & \text{if } f(x) > 0 \\ -1 & \text{if } f(x) \leq 0 \end{cases}$$

Suppose there are N data points $x_1, \ldots, x_N$ with labels $\hat{y}_1, \ldots, \hat{y}_N \in \{-1, 1\}$. The hinge loss of g is defined to be

$$l\big(g(x_n), \hat{y}_n\big) = \max\{0, 1 - \hat{y}_n f(x_n)\}$$

When $f(x_n) \geq 1$ or $f(x_n) \leq -1$, both implies $\hat{y}_n f(x_n) \geq 1$. This means $l\big(g(x_n), \hat{y}_n\big) = 0$.
When $f(x_n) \in (-1, 1)$, we have $\hat{y}_n f(x_n) \in [0, 1)$. Hence $l\big(g(x_n), \hat{y}_n\big) = 1 - \hat{y}_n f(x_n)$.

1.2 Cross Entropy

The cross entropy of the distribution $q(x)$ relative to a distribution $p(x)$ is

$$H(p, q) = -\mathrm{E}_p[\ln q] = -\sum_x p(x) \ln q(x)$$

In deep learning, $p(x)$ refers to the ground truth label, $q(x)$ refers to the output from a deep neural network model. In information theory, minimize cross entropy means
Let the amount of the information carried by $q(x)$ refers to $p(x)$.

1.2.1 Binary Class

$p(x) \in \{0, 1\}$, $q(x) \in [0, 1]$

$$H(p, q) = -\sum_x p(x) \ln q(x) + (1 - p(x)) \ln(1 - q(x))$$

1.2.2 Multi Class

There are several ways to formulate. They are all equivalent.
- One-Hot Ground Truth:
$p(x) \in \mathbb{R}^C$, $p_i(x) \in \{0, 1\}$, $\sum_{i=1}^{C} p_i(x) = 1$, $q(x) \in \mathbb{R}^C$, each $q_i \in [0, 1]$, $\forall i = 1, \ldots, C$

$$H(p, q) = -\sum_x \sum_{i=1}^{C} p_i(x) \ln q_i(x) + (1 - p_i(x)) \ln(1 - q_i(x))$$

- Raw Class Number Ground Truth:

1.3 L^p Norm

Let $\boldsymbol{y}(\boldsymbol{w}, \boldsymbol{b}) : \mathbb{R}^n \to \mathbb{R}^m$ be a model defined by $\boldsymbol{w}$ and $\boldsymbol{b}$ that map $\boldsymbol{x} \in \mathbb{R}^n$ into $\boldsymbol{y} \in \mathbb{R}^m$. Mathematically the L^1 norm is

$$\|\boldsymbol{y}(\boldsymbol{w}, \boldsymbol{b}) - \hat{\boldsymbol{y}}\|_1 = \sum_{k=1}^{m} \left| y_k(\boldsymbol{w}, \boldsymbol{b}) - \hat{y}_k \right|$$

The L^p norm is

$$\|\boldsymbol{y}(\boldsymbol{w}, \boldsymbol{d}) - \hat{\boldsymbol{y}}\|_p = \left(\sum_{k=1}^{m} \left(y_k(\boldsymbol{w}, \boldsymbol{b}) - \hat{y}_k \right)^p \right)^{1/p}$$

The L^∞ norm is

$$\|\boldsymbol{y}(\boldsymbol{w}, \boldsymbol{b}) - \hat{\boldsymbol{y}}\|_\infty = \max_k |y_k(\boldsymbol{w}, \boldsymbol{b}) - \hat{y}_k|$$

1.4 Mean Absolute Error(L^1 Loss)

Suppose there are N instances $\{\boldsymbol{x}_i\}_{i=1}^N$. The MAE is defined by

$$L(\boldsymbol{w}, \boldsymbol{b}) = \frac{\sum_{n=1}^N \|\boldsymbol{y}^n(\boldsymbol{w}, \boldsymbol{b}) - \hat{\boldsymbol{y}^n}\|_1}{N}$$

1.5 Mean Square Error(L^2 Loss)

Suppose there are N instances $\{\boldsymbol{x}_i\}_{i=1}^N$. The MSE is defined by

$$L(\boldsymbol{w}, \boldsymbol{b}) = \frac{\sum_{n=1}^N \|\boldsymbol{y}^n(\boldsymbol{w}, \boldsymbol{b}) - \hat{\boldsymbol{y}^n}\|_2^2}{N}$$