1 Calssification Loss Functions

1.1 Hinge Loss

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

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

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

Let $p(x)$ be an unknown distribution and we use $q(x)$ to esitmate it. The cross entropy of $q(x)$ relative to $p(x)$ is

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

Roughly speaking, this stands for the average amount of information to transimit when we use $q(x)$ as $p(x)$. Compare this with only use $p(x)$, the additional information will be transimitted is called the KL divergenc between $p(x)$ and $q(x)$

$$\mathrm{KL}(p \| q) = H(p, q) - H(p) = \left(-\sum_x p(x) \ln q(x)\right) - \left(-\sum_x p(x) \ln p(x)\right)$$

where $H(p) = -\sum_x p(x) \ln p(x)$ is the entropy of $p(x)$ alone.