1.1 show that after every round t, we have

$$w_*^T w_{t+1} \geq w_*^T w_t + \gamma$$

Proof: $w_*^T w_{t+1} = w_*^T (w_t + y_i x_i)$

$$= w_*^T w_t + w_*^T y_i x_i$$

$$= w_*^T w_t + y_i w_*^T x_i$$

$\gamma := \min\limits_{i \in \{1,2,\cdots,n\}} \frac{|w_*^T x_i|}{\|w_*\|}$, we assume $\|w_*\|_2^2 = 1$ here.

$\Rightarrow \gamma = \min\limits_i |w_*^T x_i|$

Back to our equality:

$= w_*^T w_t + y_i w_*^T x_i$, where $y_i = \text{sign}(w_*^T x_i)$

$= w_*^T w_t + |w_*^T x_i|$, since $|w_*^T x_i| \geq \min\limits_i |w_*^T x_i| = \gamma$

$\geq w_*^T w_t + \gamma$.

Hence, we showed that $w_*^T w_{t+1} \geq w_*^T w_t + \gamma$

1.2 Show that after every round t, we have

$$\|w_{t+1}\|_2^2 \leq \|w_t\|_2^2 + 3$$

Proof: $\|w_{t+1}\|_2^2 = \|w_t + y_i x_i\|_2^2$ (Assume all 2-norm in this proof).

$$= (w_t + y_i x_i)^T (w_t + y_i x_i)$$

$$= w_t^T w_t + y_i^2 x_i^T x_i + 2 y_i w_t^T x_i$$

$= \|w_t\|^2 + y_i^2 \|x_i\|^2 + 2 y_i w_t^T x_i$ $\longrightarrow$ continue.

continue to 1.2

$= \|W_t\|_2^2 + y_i^2\|X_i\|_2^2 + 2y_iW_t^TX_i$, where $y_i^2\|X_i\|^2 \le 1$ since we band all x_i's in an unit ball

$\le \|W_t\|_2^2 + 1 + 2y_iW_t^TX_i$, since we add a updating condition that $|W_t^TX_i| \le 1$.

There are 3 cases that we need to keep iterating the algorithm:

① $y_i \ne W_t^TX_i$ but $|W_t^TX_i| > 1$.

② $y_i = W_t^TX_i$ but $|W_t^TX_i| \le 1$.

③ $y_i \ne W_t^TX_i$ and $|W_t^TX_i| \le 1$

For ①: Our equality: $\le \|W_t\|^2 + 1 + \underbrace{2y_iW_t^TX_i}_{<-2 \text{ or } =0}$, where $y_iW_t^TX_i \le 0$, $|W_t^TX_i| > 1$

$\le \|W_t\|^2 + 1$.

② ——— : $\le \|W_t\|^2 + 1 + \underbrace{2y_iW_t^TX_i}_{\le 2}$; where $y_iW_t^TX_i > 0$, $|W_t^TX_i| \le 1$

$\le \|W_t\|^2 + 1 + 2 = \|W_t\|^2 + 3$

③ ——— : $\le \|W_t\|^2 + 1 + 2y_iW_t^TX_i$, where $y_iW_t^TX_i \le 0$ and $|W_t^TX_i| \le 1$.

$\le \|W_t\|^2 + 1$

By take the union of 3 cases:

we get $\underline{\|W_{t+1}\|_2^2 \le \|W_t\|^2 + 3}$ always true.

1.3. ① $w_*^T w_{t+1} \geq w_*^T w_t + \gamma$

② $\|w_{t+1}\|_2^2 \leq \|w_t\|_2^2 + 3$

③ Cauchy-Schwarz Inequality: $a^T b \leq \|a\|\|b\|$

We want to show after T rounds,

proof: $\gamma T \leq \|w_{T+1}\|_2 \leq \sqrt{3T}$

②: $\|w_{t+1}\|_2^2 \leq \|w_t\|_2^2 + 3$
$\leq \|w_{t-1}\|_2^2 + 6$
$\leq$
$\vdots$
$\leq \|w_1\|_2^2 + 3T$
$= 0 + 3T$
$= 3T$

$\Rightarrow \|w_{t+1}\|_2^2 \leq 3T$

$\|w_{t+1}\|_2 \leq \sqrt{3T}$

①: $w_*^T w_{t+1} \geq w_*^T w_t + \gamma$
$\geq w_*^T w_{t-1} + 2\gamma$
$\vdots$
$\geq w_*^T w_1 + T\gamma$
$= T\gamma$

$\Rightarrow w_*^T w_{t+1} \geq T\gamma$

By Cauchy Schwarz: $w_*^T w_{t+1} \leq \|w_*^T\|_2 \|w_{t+1}\|_2$
$= 1 \cdot \|w_{t+1}\|_2$
$= \|w_{t+1}\|_2 \geq T\gamma$

Hence, $T\gamma \leq \|w_{t+1}\|_2 \leq \sqrt{3T}$ ✓

1.4 Conclude $T \le 3/\gamma^2$

$$T\gamma \le \|W_{t+1}\|_2 \le \sqrt{3T}$$

$$\Rightarrow T\gamma \le \sqrt{3T}$$

$$T^2\gamma^2 \le 3T$$

$$T \le \frac{3}{\gamma^2} \checkmark$$

1.5 show that the output hyperplane w satisfies:

$$\min_i \frac{|W^T x_i|}{\|W\|_2} \ge \frac{\gamma}{3}$$

The stopping condition of the algorithm:

$$y_i = \text{sign}(W_t^T x_i) \text{ and } |W_t^T x_i| \ge 1$$

$$\begin{cases} \min_i \frac{|W^T x_i|}{\|W\|_2} \ge \frac{1}{\|W\|_2} \\ \gamma T \le \|W\|_2 \le \sqrt{3T} \Rightarrow \frac{1}{\gamma T} \ge \frac{1}{\|W\|_2} \ge \frac{1}{\sqrt{3T}} \end{cases}$$

$\frac{1}{\sqrt{3T}} \le \frac{1}{\|W\|_2} \le \frac{1}{\gamma T}$, and in terms of 1.4: $T \le \frac{3}{\gamma^2}$

$$\Rightarrow \sqrt{T} \le \frac{\sqrt{3}}{\gamma}$$

$$\Rightarrow \quad \frac{1}{\|W\|_2} \ge \frac{1}{\sqrt{3T}} \ge \frac{1}{\sqrt{3}} \cdot \frac{\gamma}{\sqrt{3}} = \frac{\gamma}{3} \qquad \frac{1}{\sqrt{T}} \ge \frac{\gamma}{\sqrt{3}}$$

$$\Rightarrow \min_i \frac{|W^T x_i|}{\|W\|_2} \ge \frac{1}{\|W\|_2} \ge \frac{1}{\sqrt{3T}} \ge \frac{\gamma}{\sqrt{3}\cdot\sqrt{3}} = \frac{\gamma}{3}.$$

Hence, we showed that $\min_i \frac{|W^T x_i|}{\|W\|_2} \ge \frac{\gamma}{3}$

1.6 Because it will make the boundary be more clear, and

⇒ continue 1.6

more robust to the noise since it aims to keep a buffer zone between classes, if the boundary is too close to the data points, small perturbations would lead to misclassification.

2. Bayes Optimal classifier.

Let $y(x)$ denote the conditional probability of label being 1 given a point x under the distribution D. That is,

$$y(x) = Pr(y=+1 \mid x)$$

Under 0/1 loss, for any classifier f is:

$$R(f) = \mathop{E}_{x,y}[\mathbb{1}[f(x) \neq y]]$$

2.1 Show that $R(f) = E_x[y(x)\mathbb{1}[f(x)=-1] + (1-y(x))\mathbb{1}[f(x)=1]]$.

⇒ $E_{x,y}[\cdot] = E_x E_{y|x}[\cdot]$

Proof: The indicator function $\mathbb{1}[f(x) \neq y]$ is defined as.

$$\mathbb{1}[f(x) \neq y] = \begin{cases} 0, & f(x) = y \\ 1, & f(x) \neq y \end{cases}$$

$$\begin{aligned} R(f) &= \mathop{E}_{x,y}[\mathbb{1}[f(x) \neq y]] \\ &= E_{(x,y)\sim D}[\mathbb{1}[f(x) \neq y]] \\ &= E_x[E_{y|x}[\mathbb{1}[f(x) \neq y]]] \\ &= E_x(P(y=+1|x)\mathbb{1}(f(x)=-1) \\ &\quad + P(y=-1|x)\mathbb{1}(f(x)=+1)]. \\ &= \Rightarrow \text{continue.} \end{aligned}$$

$$\Longleftarrow \quad \begin{aligned} &E_{y|x}[\mathbb{1}(f(x) \neq y)] \\ &= \sum_y P(y|x)\mathbb{1}(f(x) \neq y) \\ &= P(y=+1|x)\mathbb{1}(f(x) \neq +1) + \\ &\quad P(y=-1|x)\mathbb{1}(f(x) \neq -1) \end{aligned}$$

⇒ continue 2.2

$$= E_x[p(y=+1|x)1(f(x)=-1)+p(y=-1|x)1(f(x)=+1)]$$

$$= E_x[\eta(x)1(f(x)=-1)+(1-\eta(x))1(f(x)=+1)]$$

Hence, we proved that $E_{(x,y)}[1(f(x)\neq y)]$

2.2 show that the minimum risk possible is

$$R(h_*) = \min_f R(f) = E_x[\min(\eta(x), 1-\eta(x))]$$

$$\min_f E_x[\eta(x)1(f(x)=-1)+(1-\eta(x))\cdot 1(f(x)=+1)]$$

proof: for a fixed x, $f(x)$ can be either -1 or 1.

if $\eta(x) > (1-\eta(x)) \Rightarrow \eta(x) > 0.5$, which means given x, the probability that the label is $+1$ is higher, the prediction more likely to be $+1$.

Otherwise, $\eta(x) \leq 0.5$, means given x, the probability that the label is -1 is higher, the predictor more likely to return -1.

$$\Rightarrow h_*(x) = \begin{cases} +1, & \eta(x) > 0.5 \\ -1, & \text{otherwise} \end{cases}$$

$$R^* = \min_{h: X \to \pm 1} R(f) = E_x[\eta(x)1(f(x)=-1)+(1-\eta(x))1(f(x)=+1)]$$

$$= \underbrace{E_x[\min(\eta(x), (1-\eta(x))]}$$

Since the smaller one will be the error term ⇒ continue.

⇒ continue 2.2

if $\mathbb{1}(f(x)=-1)=1$, which means the true label is +1 but we predict −1,

hence it multiplied by a loss of $\eta(x)$

if $\mathbb{1}(f(x)=+1)=1$, which means the true label is −1 but we predict +1,

hence it multiplied by a loss of $(1-\eta(x))$

And the wrong prediction would be $\min\{\eta(x), (1-\eta(x))\}$:

Hence, $R(h^*) = \min_f R(f) = E_X[\min\{\eta(x), 1-\eta(x)\}]$

2.3. Show that the Bayes Optimal classifier the above loss is

$$f^*(x) = \begin{cases} +1, & \eta(x) \geq \frac{1}{2} \\ -1, & \eta(x) < \frac{1}{2} \end{cases}$$

Proof: like we did in 2.2.

if $\eta(x) \geq (1-\eta(x)) \Rightarrow \eta(x) \geq \frac{1}{2}$, which means the prob of true label is +1 is higher than that of true label is −1 which means the data point more likely to be +1. then we should return +1 here

In contrast, $\eta(x) < (1-\eta(x))$: the prediction more likely to have label −1.

$\Rightarrow \eta(x) < \frac{1}{2}$

2.4 Derive the Bayes Optimal Classifier under the logistic model

$$\eta(x) = \frac{1}{1+\exp(-w^T x)}.$$

$$f^*(x) = \begin{cases} 1, & \text{if } \eta(x) \geq \frac{1}{2} \\ -1, & \text{if } \eta(x) < \frac{1}{2} \end{cases}$$

$$\eta(x) = \frac{1}{1+\exp(-w^T x)} \geq \frac{\exp(-w^T x)}{1+\exp(-w^T x)}$$

$$1 \geq \exp(-w^T x)$$

$$0 \geq -w^T x$$

$$w^T x \geq 0$$

$$\Rightarrow \eta(x) < (1-\eta(x))$$

$$\Rightarrow \frac{1}{1+\exp(-w^T x)} < \frac{\exp(-w^T x)}{1+\exp(-w^T x)}$$

$$\Rightarrow w^T x < 0$$

$$\Rightarrow f^*(x) = \begin{cases} +1, & \text{if } w^T x \geq 0 \text{ (above hyperplane)} \\ -1, & \text{if } w^T x < 0 \text{ (below hyperplane)} \end{cases}$$

2.5 modify loss function from 0/1 to cost-based loss function:

$$l_c(\hat{y}, y) = \begin{cases} c, & \text{if } y=1, \hat{y}=-1 \\ 1-c, & \text{if } y=-1, \hat{y}=1 \\ 0, & \text{if } y=\hat{y} \end{cases}$$

(medical diagnosis)

$$R_c(f) = E_{(x,y)\sim D}[l_c(\overset{\hat{y}}{f(x)}, y)]$$

Find the Bayes Optimal classifier

Proof. $R_c(f) = E_{(x,y)\sim D}[l_c(f(x), y)]$

$= E_x[E_{y|x}[l_c(f(x), y)]]$

$= E_x[\sum_y P(Y=y|X) \cdot l_c(f(x))]$

$= E_x(P(Y=+1|X) \cdot l_c(f(x)) + P(Y=-1|X)\, l_c(f(x)))$

$= E_x[C \cdot \eta(x) + (1-C)(1-\eta(x))]$

$= E_x[C\cdot\eta(x) + 1 - \eta(x) - c + c\eta(x)]$

$= E_x[2c\eta(x) + c\, 1 - \eta(x)]$

Given a fixed x, there are two cases:

⇒ continue.

$\Rightarrow$ Continue 2.5 $f(x)$ can be either $+1$ or -1

① when $f(x) = +1$,

$$E_x[(1-\eta(x))\cdot(1-c)] = (1-\eta(x))\cdot c \quad \text{for a given } x$$

② when $f(x) = -1$

$$E_x[\eta(x)\cdot c] = \eta(x)\cdot c \quad \text{for a given } x$$

Hence $R(h^*) = \min_f R(f) = E_x[\min\{c\eta(x), (1-c)(1-\eta(x))\}]$.

$E_x[\ell(\hat{y}=-1|y)|X=x) \geq E_x[\ell(\hat{y}=+1|y)|X=x)$

① $c\eta(x) \geq (1-c)(1-\eta(x))$

$c\eta(x) \geq [1-\eta(x)] - c[1-\eta(x)] = 1-\eta(x) - c + c\eta(x)$

$0 \geq 1-\eta(x)+c$

$\eta(x) \geq 1-c$

$E_x(\ell(\hat{y}=-1)) > E_x(\ell(\hat{y}=+1))$

$\eta(x) > 1-\eta(x)$

$\eta(x) > 0.5$

$\Rightarrow +1$

预测risk小的

$E_x[\ell(\hat{y}=-1|x)] < E_x[\ell(\hat{y}=+1|x)]$

② $c\eta(x) < (1-c)(1-\eta(x)) = 1-\eta(x) - c + c\eta(x)$

[(Risk) Expected loss of predicting $\hat{y}=-1$] < [Exp loss of predicting $\hat{y}=+1$] $\Rightarrow$ pick lower Expected loss

$\Rightarrow \hat{y} = -1$

$0 < \eta(x) < 1-c$

$$\Rightarrow f^*(x) = \begin{cases} +1, & \eta(x) \geq 1-c \\ -1, & 0 < \eta(x) < 1-c \end{cases}$$

Hence the Bayes Optimal Classifier under loss function

$$\ell_c \text{ is } f^*(x) = \begin{cases} +1, & \eta(x) \geq 1-c \\ -1, & \text{otherwise.} \end{cases}$$