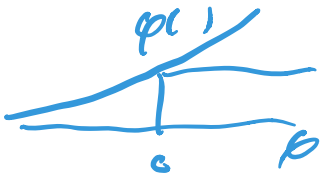
8.2 Binary classification w/surrogate loss

Thm 8.2 $(X, P, \mathcal{F}, \varphi)$ $\ell_\varphi(y, f(x)) = \varphi(-yf(x))$

$\operatorname{sgn}(f(x))$



(single version) w/p $\geq 1 - e^{-2t^2}$

$$L(f) \leq A_\varphi(f) \leq A_{\varphi,n}(f) + 4M_\varphi \mathcal{E}R_n(\mathcal{F}(X^n)) + \frac{Bt}{\sqrt{n}}$$

holds for all functions $f_n \in \mathcal{F}$. $\forall f \in \mathcal{F}$

(double version) For $\hat{f}_n$ – ERM for surrogate loss

$$L(\hat{f}_n) \leq A_\varphi(\hat{f}_n) \leq A_\varphi^*(\mathcal{F}) + 8M_\varphi \mathcal{E}R_n(\mathcal{F}(X^n)) + \frac{2Bt}{\sqrt{n}}$$

8.3 Weighted linear combinations of classifiers

$(X, P, \underset{\varphi}{0\text{-}1 \text{ loss}})$

g - $g : X \to \{-1, 1\}$ "weak set of classifiers"

For $\lambda > 0$, let

$$\mathcal{F}_\lambda = \left\{ f = \sum_{i=1}^{N} c_i g_i : N \geq 1, \sum_{i=1}^{N} |c_i| \leq \lambda, \; g_1, \dots g_N \in g \right\}$$

~~sgn(f)~~

$$= \lambda \, \text{absconv}(g)$$

$$\mathcal{F}_\lambda(x^n) = \lambda \, \text{absconv}(g(x^n))$$

$$\| \quad \{(g(x_1), \dots, g(x_n)) : g \in g\}$$

$$R_n(\mathcal{F}_\lambda(x^n)) \leq \lambda R_n(g(X_n))$$

$$\leq \lambda C \sqrt{\frac{V(g)}{n}}$$

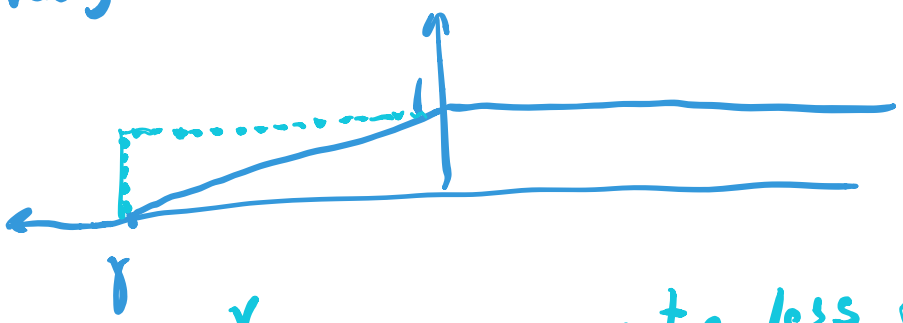
VC dim Dudley version.

Then For any A, w/p $\geq 1-\delta$:

$$L(\hat{f}_n) \leq A_{\varphi,n}(\hat{f}_n) + c\lambda M_\varphi \sqrt{\frac{V(g)}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

for all $\hat{f}_n$

Margin based risk bound. ramp function



$M_\varphi = 1/\gamma$

γ

Let $L_n^\gamma(f)$ = surrogate loss for the upper bound

$\tilde{\varphi}(x) = 1_{x \geq -\gamma}$

If $\overset{\pm 1}{Y} f(x) > \gamma$ - no error, and even $l_\varphi(Y, f(x)) = 0$.

If $0 < Yf(x) \leq \gamma$ - no error, but

$l_\varphi(Y, f(x)) > 0$

$l_{\tilde{\varphi}}(Y, f(x)) = 1$

Thm 8.5 Margin based risk bound

For $\gamma > 0$, with probability $\geq 1-\delta$:

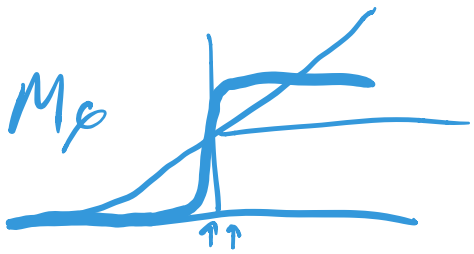
$$L(\hat{f}_n) \leq L_n^{\gamma}(\hat{f}_n) + \frac{C\lambda}{\gamma}\sqrt{\frac{V(\mathcal{G})}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

$L_n^{\gamma}(\hat{f}_n)$: fraction of test samples not strongly correct (i.e. not having $f(x)Y \geq \gamma$)

Let $C_f = \{x : \operatorname{sgn}(f) = 1\} - C$

$\{\operatorname{sgn}(f) : f \in \mathcal{F}\}$ VC dimension?

$$\mathcal{F}_\lambda = \left\{ f = \sum_{i=1}^{N} c_i g_i : N \geq 1, \sum_{i=1}^{N} |c_i| \leq \lambda, g_1, \ldots g_N \in \mathcal{G} \right\}$$



except on event w/p e^{-2t^2} $\varphi : \mathbb{R} \to [0,1]$

$$L(f) \le A_\varphi(f) \le A_{\varphi_k^n}(f) + 4M_{\varphi_k} E R_n(\mathcal{F}(X^n)) + \frac{t + \sqrt{\log k}}{\sqrt{n}}$$

$$e^{-2\left(t+\sqrt{\log k}\right)^2} = e^{-2t^2 - \log k^2 - 2t\sqrt{\log k}}$$

$$\le \frac{1}{k^2} e^{-2t^2}$$

$$\sum_{k=1}^{\infty} \frac{1}{k^2} e^{-2t^2} = \frac{\pi^2}{6} e^{-2t^2} \le 2e^{-2t^2}$$

ADA Boost

$$\mathcal{F}_1 = \left\{ f = \sum_{k=1}^{K} \alpha_k g_k : K \geq 1, \alpha_1, \dots, \alpha_K \in \mathbb{R}, g_1, \dots g_K \in \mathcal{G} \right\}$$

$\varphi(x) = \exp(x)$

Algorithm – Find $g_1, g_2, \dots$ greedily attempting to minimize the surrogate loss.

Once $g_1, \dots g_k$ are found, select g_{k+1} to minimize the empirical loss for a weighted version of training data $x_1, \dots x_n$.

$\varphi(x) = e^x$ $\quad$ $y_i \in \{1, -1\}$

$$A^{k+1}_{\beta,n} = \frac{1}{n}\sum_{i=1}^{n} \exp\left(-Y_i \sum_{k=1}^{k+1} \alpha_k g_k(X_i)\right)$$

Suppose $g_1 \cdots g_k$ found. $\quad \alpha_1 \cdots \alpha_k$

g_{k+1} ?

binary valued

$$A^{k+1}_{\beta,n} = \frac{1}{n}\sum_{i=1}^{n} w_i^{(k)} \exp\left(-Y_{k+1}\alpha_{k+1} g_{k+1}(X_i)\right)$$

$$w_i^{(k)} = \exp\left(-Y_i \sum_{k=1}^{k} \alpha_k g_k(X_i)\right)$$

predictor of Y_i based on $g_1 \cdots g_k(X_i)$

$$\exp\left(-Y_{k+1}\alpha_{k+1} g(X_i)\right) = \exp(-\alpha_{k+1}) \quad \text{if } Y_i = g_{k+1}(X_i)$$
$$= \exp(+\alpha_{k+1}) \quad \text{if } Y_i \neq g_{k+1}(X_i)$$

eg. select g_{k+1} to minimize weighted empirical 0-1 loss using $w_i^{(k)}$ weights.

g is a set of simple binary classifiers