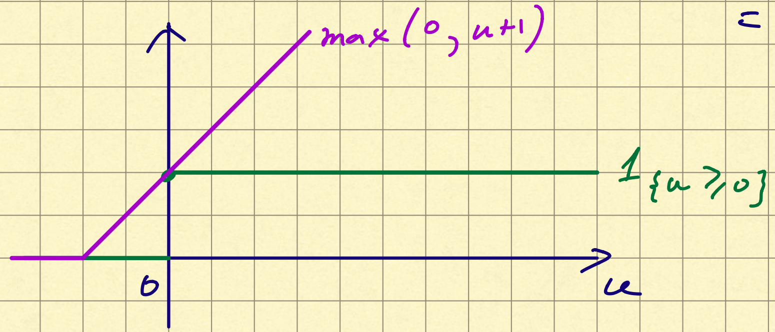
Binary Classification, Part 3

Review: $f : \mathcal{X} \to \mathbb{R}$

$$g_f(x) = \operatorname{sgn} f(x) = \begin{cases} +1, & f(x) \geq 0 \\ -1, & f(x) < 0 \end{cases}$$

(X, Y) in $\mathcal{X} \times \{-1, +1\}$

$$L(g_f) = \mathbb{P}[Y \neq g_f(X)] \leq \mathbb{P}[Y f(X) \leq 0]$$
$$= \mathbb{E}[1_{\{-Y f(X) \geq 0\}}]$$



penalty fcn $\varphi : \mathbb{R} \to \mathbb{R}_+$
continuous
nondecreasing
$\varphi(u) \geq 1_{\{u \geq 0\}}$

$\varphi(\cdot) \longrightarrow$ surrogate loss

$$\ell_\varphi(y, u) := \varphi(-yu)$$

$$\begin{aligned} L(g_f) = L(\operatorname{sgn} f) &\leq \mathbb{E}[1_{\{-Y f(X) \geq 0\}}] \\ &\leq \mathbb{E}[\varphi(-Y f(X))] \\ &= \mathbb{E}[\ell_\varphi(Y, f(X))] \\ &=: A_\varphi(f) \quad : \text{ surrogate loss of } f \end{aligned}$$

$$L(g_f) \leq A_\varphi(f)$$
$$L_n(g_f) \leq A_{\varphi,n}(f) \qquad \text{where, e.g., } A_{\varphi,n}(f) = \frac{1}{n}\sum_{i=1}^n \varphi(-Y_i f(X_i))$$

Thm (Koltchinskii–Panchenko) Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a penalty fcn s.t.:

- $\varphi(-y f(x)) \in [0, 1]$ for all (x, y), all $f \in \mathcal{F}$
- φ is M_φ-Lipschitz: $|\varphi(u) - \varphi(v)| \leq M_\varphi |u - v|$

Let $\hat{f}_n$ be any element of $\mathcal{F}$, based on data.

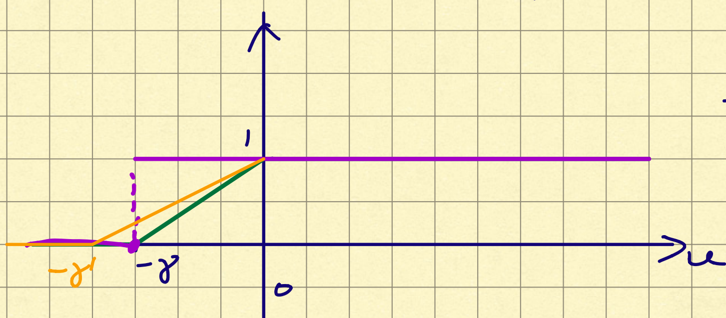
Then, w.p. $\geq 1-e^{-2t^2}$ $(t>0)$,

$$L(\operatorname{sgn}\vec{f}_n) \le \underbrace{A_{\varphi,n}(\vec{f}_n)}_{\text{computable from data}} + 4M_\varphi \underbrace{\mathbb{E}R_n(\mathcal{F}(X^n))}_{\text{typically easy to bound}} + \frac{t}{\sqrt{n}}.$$

<u>Example</u> (ramp penalty + margin)

$\gamma > 0$

$$\varphi(u) := \begin{cases} 0, & u < -\gamma \\ 1 + u/\gamma, & -\gamma \le u < 0 \\ 1, & u \ge 0 \end{cases}$$



$$1_{\{u > -\gamma\}} \ge \varphi(u) \ge 1_{\{u \ge 0\}}$$

- for any $f : \mathcal{X} \to \mathbb{R}$,

$$L(\operatorname{sgn} f) \le A_\varphi(f) \le L^\gamma(f),$$

where $L^\gamma(f) := \mathbb{P}[Yf(X) < \gamma]$

$Yf(X)$: margin of f on (X, Y)

$$L^\gamma(f) = \underbrace{\mathbb{P}[Yf(X) < 0]}_{\mathbb{P}[Y \ne \operatorname{sgn} f(X)]} + \underbrace{\mathbb{P}[0 \le Yf(X) < \gamma]}_{\text{prob. of margin} < \gamma}$$

$\varphi(\cdot)$ bdd between $[0,1]$, $\frac{1}{\gamma}$-Lipschitz

Corollary For any $\hat{f}_n$, w.p. $\geq 1 - e^{-2t^2}$,

$$L(\operatorname{sgn} \hat{f}_n) \leq L_n^{\gamma}(\hat{f}_n) + \frac{4}{\gamma}\mathbb{E}R_n(\mathcal{F}(X^n)) + \frac{t}{\sqrt{n}}$$

increases w. γ decreases w. γ

– would like to make γ data-dependent!

$$L(\operatorname{sgn} \hat{f}_n) \leq \inf_{\gamma \in (0,1]} \left\{ L_n^{\gamma}(\hat{f}_n) + \frac{C}{\gamma} R_n + \dots \right\} + t/\sqrt{n}$$

w.p. $\geq 1 - e^{-O(t^2)}$

Thm (K.–P.) Let φ be a pen fcn, which is:

- bdd between 0 and 1
- 1 – Lipschitz
- monotone : $1 \geq \gamma \geq \gamma' > 0 \Rightarrow \varphi(u/\gamma) \geq \varphi(u/\gamma')$ for all u.

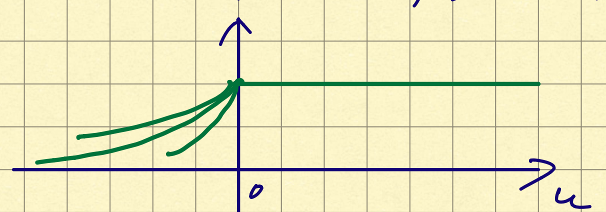
Then, for any $\hat{f}_n \in \mathcal{F}$, w.p. $\geq 1 - 2e^{-2t^2}$,

$$L(\operatorname{sgn} \hat{f}_n) \leq \inf_{0<\gamma\leq 1} \left\{ A_{\varphi(\cdot/\gamma), n}(\hat{f}_n) + \frac{C}{\gamma}\mathbb{E}R_n(\mathcal{F}(X^n)) + C\sqrt{\frac{\log\log(2/\gamma)}{n}} \right\} + \frac{t}{\sqrt{n}}.$$

(cf. lecture notes for exact constants + proof)

Examples of φ:

1) ramp, $\varphi(u) = \min\{1, \max\{0, u+1\}\}$
2) truncated exp, $\varphi(u) = \min\{1, e^u\}$



$$\varphi(u/\gamma) = \min\{1, e^{u/\gamma}\}$$

1) Generalized majority vote

$\mathcal{G}$: fixed collection of classifiers, $g : \mathcal{X} \to \{\pm 1\}$

$v(\mathcal{G}) < \infty$ (base classifiers)

Fix $\lambda > 0$

$$\tilde{\mathcal{F}}_\lambda := \left\{ \sum_{k=1}^{N} c_k g_k(x) : N \in \mathbb{N},\ |c_1| + \dots + |c_N| \le \lambda,\ g_1, \dots, g_N \in \mathcal{G} \right\}$$

- gen. maj. vote: $c_1 = \dots = c_N = \lambda / N$

Note: $\mathcal{F}_\lambda$ may not be a VC class (unlike $\mathcal{G}$)

$$\begin{aligned} R_n(\mathcal{F}_\lambda(X^n)) &= R_n(\lambda \cdot \operatorname{absconv}(\mathcal{G}(X^n))) \\ &= \lambda \cdot R_n(\operatorname{absconv}(\mathcal{G}(X^n))) \\ &= \lambda \cdot R_n(\mathcal{G}(X^n)) \\ &\le C\lambda \cdot \sqrt{\frac{v(\mathcal{G})}{n}} \end{aligned}$$

2) AdaBoost (Y. Freund - R. Schapire, 1997)

$\mathcal{G}$: collection of base classifiers $g : \mathcal{X} \to \{-1, +1\}$
(weak learners)

$\mathcal{F} = \operatorname{conv}(\mathcal{G})$

Data: $(X_1, Y_1), \dots, (X_n, Y_n)$ iid, in $\mathcal{X} \times \{-1, +1\}$

Algo: iterative update of classifiers in $\operatorname{conv}(\mathcal{G})$

$K \ge 1$ iterations

Init: $w^{(1)} = (w_1^{(1)}, \dots, w_n^{(1)})$, $w_i^{(1)} = 1/n \ \forall i$

for $k = 1, \dots, K$:

- $e_k(g) := \sum_{i=1}^{n} w_i^{(k)} \mathbb{1}_{\{Y_i \neq g(X_i)\}}$, $g \in \mathcal{G}$

 – weighted class. error

Note: $k=1$ $\quad e_1(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{Y_i \neq g(X_i)\}} = L_n(g)$

$$g_k := \operatorname*{argmin}_{g \in \mathcal{G}} e_k(g)$$

$$e_k := e_k(g_k) = \min_{g \in \mathcal{G}} \sum_{i=1}^{n} w_i^{(k)} \mathbb{1}_{\{Y_i \neq g(X_i)\}}$$

Key Assumption: $\quad e_k \leq 1/2$

- update $w^{(k)} \longrightarrow w^{(k+1)}$

$$\forall i \in [n]: \quad w_i^{(k+1)} = \frac{w_i^{(k)} \exp(-\alpha_k Y_i g_k(X_i))}{Z_k}$$

where $\alpha_k := \frac{1}{2} \log \frac{1-e_k}{e_k} \quad (\geq 0)$ $\quad 0 \leq e_k \leq \frac{1}{2}$, $1-e_k \geq e_k$

$$Z_k := \sum_{i=1}^{n} w_i^{(k)} \exp(-\alpha_k Y_i g_k(X_i))$$

$$\exp(-\alpha_k Y_i g_k(X_i)) = \begin{cases} e^{\alpha_k} & \text{if } Y_i g_k(X_i) = -1 \\ e^{-\alpha_k} & \text{if } Y_i g_k(X_i) = +1 \end{cases}$$

- After K steps, return

$$\hat{f}_n(x) = \frac{\sum_{k=1}^{K} \alpha_k g_k(x)}{\sum_{k=1}^{K} \alpha_k} \in \operatorname{conv}(\mathcal{G})$$

Note: $\operatorname{sgn} \hat{f}_n(x) = \operatorname{sgn}\left(\sum_{k=1}^{K} \alpha_k g_k(x)\right)$

Preview: $L(\hat{f}_n) \lesssim \prod_{k=1}^{K} 2\sqrt{e_k(1-e_k)}$

— if $2\sqrt{e_k(1-e_k)} < 1$ $\forall k$, then error will decay with K!

— implicit margin minimization!

$$\gamma \sim \frac{1}{\sum_{k=1}^{K} \alpha_k}.$$