Binary Classification, Part 2

(Surrogate losses)

- (X, Y) in $\mathcal{X} \times \{-1, +1\}$ (instead of $\{0, 1\}$)

Sign-based classifiers:

$\mathcal{F}$ - class of fcns $f: \mathcal{X} \to \mathbb{R}$

$f \in \mathcal{F} \longrightarrow g_f(x) := \operatorname{sgn} f(x)$

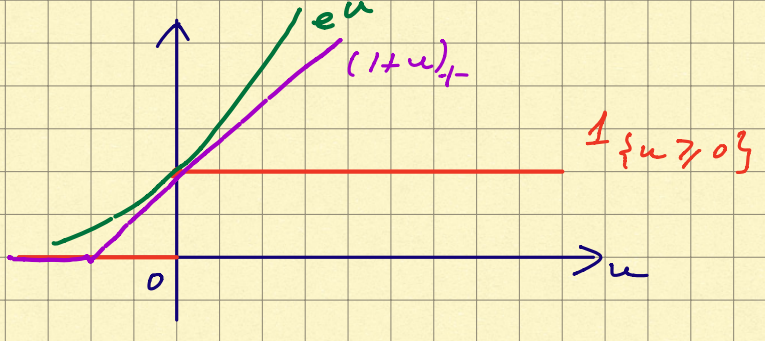
$$= \begin{cases} 1, & f(x) \geq 0 \\ -1, & f(x) < 0 \end{cases}$$

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

$Y \operatorname{sgn} f(X) \leq 0 \Rightarrow Y f(X) \leq 0$

Replace 0-1 loss $\ell(y, \hat{y}) = 1_{\{y \neq \hat{y}\}}$ with a tractable bound.

- **Penalty fcns / Surrogate losses**



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

$\varphi(u) = e^u$

$\varphi(u) = \log_2(1 + e^u)$

$\varphi(u) = (1+u)_+ = \max\{0, 1+u\}$

$\varphi(u) = \min\{1, (1+u)_+\}$

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

$$L(g_f) = \mathbb{P}[Y \neq \operatorname{sgn} f(X)]$$
$$\leq \mathbb{P}[Y f(X) \leq 0]$$
$$= \mathbb{E}[1_{\{-Y f(X) \geq 0\}}]$$
$$\leq \mathbb{E}[\varphi(-Y f(X))] \quad \text{for any pen fcn } \varphi$$
$$=: A_\varphi(f) \quad \text{— surrogate loss of } f \text{ (for pen fcn } \varphi)$$

$$\ell(y, \hat{y}) = 1_{\{y \neq \hat{y}\}} \longrightarrow \ell_\varphi(y, u) := \varphi(-yu)$$
$$A_\varphi(f) = \mathbb{E}[\ell_\varphi(Y, f(X))]$$

Empirical surrogate loss:

$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$

$$A_{\varphi,n}(f) := \frac{1}{n} \sum_{i=1}^n \varphi(-Y_i f(X_i))$$
$$\geq \frac{1}{n} \sum_{i=1}^n 1_{\{-Y_i f(X_i) \geq 0\}}$$
$$\geq \frac{1}{n} \sum_{i=1}^n 1_{\{Y_i \neq \operatorname{sgn} f(X_i)\}} = L_n(g_f)$$

easier to min. over f, easier to analyze

Analysis plan:

1) ERM for surrogate loss
   — excess risk bounds, symmetrization, Rademacher
2) for any $\hat{f}_n \in \mathcal{F}$,
   $L(\operatorname{sgn} \hat{f}_n) \leq A_\varphi(\hat{f}_n) \leq \ldots$

e.g. $L(\operatorname{sgn} \hat{f}_n) \le \inf_{f \in \mathcal{F}} A_\varphi(f) + \mathrm{Rad}(\cdots)$

4) Under additional mild assumptions on φ,

$\mathrm{Rad}(\cdots) = \mathrm{Rad}(\mathcal{F})$

ERM: $\hat{f}_n = \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \varphi(-Y_i f(X_i))$

$L(\operatorname{sgn} \hat{f}_n) \le A_\varphi(\hat{f}_n)$

$\le \inf_{f \in \mathcal{F}} A_\varphi(f) + 2 \sup_{f \in \mathcal{F}} |A_\varphi(f) - A_{\varphi,n}(f)|$

$A_\varphi(\hat{f}_n) - \inf_{f \in \mathcal{F}} A_\varphi(f)$

assume $\exists f^* \in \mathcal{F}$ s.t. $A_\varphi(f^*) \le A_\varphi(f), \forall f \in \mathcal{F}$

$= A_\varphi(\hat{f}_n) - A_{\varphi,n}(\hat{f}_n) + \underbrace{A_{\varphi,n}(\hat{f}_n) - A_{\varphi,n}(f^*)}_{\le 0} + A_{\varphi,n}(f^*) - A_\varphi(f^*)$

$\le 2 \sup_{f \in \mathcal{F}} |A_\varphi(f) - A_{\varphi,n}(f)|$

$\Delta_n(Z^n) := \sup_{f \in \mathcal{F}} |A_\varphi(f) - A_{\varphi,n}(f)|$

$Z_i := (X_i, Y_i) \quad i \in [n]$

Symmetrization:

$\mathbb{E}\, \Delta_n(Z^n) \le 2 \mathrm{Rad}_n(\cdots)$

$(x, y) \mapsto \varphi(-y f(x)), \quad f \in \mathcal{F}$

$\mathrm{Rad}_n(\cdots)$:

$\{ (\varphi(-Y_1 f(X_1)), \ldots, \varphi(-Y_n f(X_n))) : f \in \mathcal{F} \}$

$\hookrightarrow \{ (f(X_1), \ldots, f(X_n)) : f \in \mathcal{F} \}$ — contraction principle

Lemma Suppose that φ is Lipschitz-cont.:
$\exists M_\varphi > 0$ s.t.

$$|\varphi(u) - \varphi(v)| \le M|u-v| \quad \text{for all } u, v \in \mathbb{R}.$$

Let $\mathcal{H}_\varphi$ be the class of fcns

$$(x, y) \mapsto \varphi(-y f(x)) - \varphi(0), \quad f \in \mathcal{F}.$$

Then $$R_n(\mathcal{H}_\varphi(z^n)) \le 2 M_\varphi R_n(\mathcal{F}(x^n)).$$

Proof

0) Contraction principle (Prop. 6.2 in Lec. notes)

Let $\mathcal{H}$ be a class of fcns $h: \mathcal{Z} \to \mathbb{R}$, and let $F: \mathbb{R} \to \mathbb{R}$ be a Lip. cont. fcn, s.t. $F(0) = 0$.

Let $F \circ \mathcal{H}$ be the class of fcns $z \mapsto F \circ h(z)$, $h \in \mathcal{H}$

$$\mathcal{Z} \xrightarrow{h \in \mathcal{H}} \mathbb{R} \xrightarrow{F} \mathbb{R}$$

Then $$R_n(F \circ \mathcal{H}) \le 2 \|F\|_{Lip} R_n(\mathcal{H})$$

where $\|F\|_{Lip}$ is the Lipschitz const. of F.

1) $$A_{\varphi,n}(f) = \frac{1}{n} \sum_{i=1}^n \varphi(-Y_i f(X_i))$$

$$A_\varphi(f) = \mathbb{E}[\varphi(-Y f(X))]$$

$$\Delta_n(z^n) = \sup_{f \in \mathcal{F}} |A_{\varphi,n}(f) - A_\varphi(f)|$$

$$= \sup_{f \in \mathcal{F}} |P_n(\ell_{\varphi,f}) - P(\ell_{\varphi,f})|$$

$$\ell_{\varphi,f}(x, y) := \varphi(-y f(x))$$

$$P_n(\ell_{\varphi,f}) - P(\ell_{\varphi,f}) = P_n(\ell_{\varphi,f} - \varphi(0)) - P(\ell_{\varphi,f} - \varphi(0))$$

$$\therefore \Delta_n(z^n) = \sup_{h \in \mathcal{H}_\varphi} |P_n(h) - P(h)|$$

where $\mathcal{H}_\varphi = \{ l_{\varphi,f} - \varphi(0) : \quad f \in \mathcal{F} \}$

if $f(x) = 0$, then $l_{\varphi,f}(x,y) - \varphi(0) = 0$

$\mathcal{H} := \{ (x,y) \mapsto -y f(x) : \quad f \in \mathcal{F} \}$

in fact, $\mathcal{H}_\varphi = F \circ \mathcal{H}$, where

$$F(v) := \varphi(v) - \varphi(0)$$

$|F(u) - F(v)| = |\varphi(u) - \varphi(v)| \le M_\varphi |u - v|$

$F(0) = \varphi(0) - \varphi(0) = 0$

$$\Rightarrow \boxed{R_n(\mathcal{H}_\varphi(Z^n)) \le 2 M_\varphi R_n(\mathcal{H}(Z^n))}$$

by the contraction principle.

2) $R_n(\mathcal{H}(Z^n))$ condition on Z^n!!!

$$= \frac{1}{n} \mathbb{E}_\varepsilon \left\{ \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n - \varepsilon_i Y_i f(X_i) \right| \right\}$$

$$= \frac{1}{n} \mathbb{E}_\varepsilon \left\{ \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right\}$$

$$= R_n(\mathcal{F}(X^n))$$

$\varepsilon_1, \dots, \varepsilon_n$ iid $\sim$ Rad

$\varepsilon_1 Y_1, \dots, \varepsilon_n Y_n$ iid $\sim$ Rad for any fixed $Y_1, \dots, Y_n \in \{\pm 1\}$

■

Key takeaway: $\mathbb{E} \left\{ \sup_{f \in \mathcal{F}} \left| A_{\varphi,n}(f) - A_\varphi(f) \right| \right\}$

$$\le 2 \mathbb{E} R_n(\mathcal{H}_\varphi(Z^n)) \}$$

$$\le 4 M_\varphi \mathbb{E} R_n(\mathcal{F}(X^n)).$$

Thm (Koltchinskii-Panchenko, 2002)

Let φ be a penalty fcn which is M_φ-Lipschitz, and assume $\exists\, B < \infty$ s.t.

$$\varphi(-yf(x)) \in [0, B) \qquad \text{for all } (x,y) \in \mathcal{X} \times \{\pm 1\} \text{ and all } f \in \mathcal{F}.$$

Then, $\forall\, t > 0$, the surrogate ERM classifier $\hat{f}_n$ satisfies

$$L(\operatorname{sgn} \hat{f}_n) \le A_\varphi(\hat{f}_n) \le \inf_{f \in \mathcal{F}} A_\varphi(f) + 8 M_\varphi \mathbb{E} R_n(\mathcal{F}(Z^n)) + \frac{2Bt}{\sqrt{n}}$$

w.p. $\ge 1 - e^{-2t^2}$.

NB: $e^{-2t^2} \le \delta$

$e^{2t^2} \ge 1/\delta$

$2t^2 \ge \log(1/\delta)$

$t = \sqrt{\frac{1}{2}\log(1/\delta)} \Rightarrow \frac{2Bt}{\sqrt{n}} = B\sqrt{\frac{2\log(1/\delta)}{n}}$

Proof

$$L(\operatorname{sgn} \hat{f}_n) \le A_\varphi(\hat{f}_n) \le \inf_{f \in \mathcal{F}} A_\varphi(f) + 2\Delta_n(Z^n)$$

$$\mathbb{E}\Delta_n(Z^n) \le 4 M_\varphi \mathbb{E} R_n|\mathcal{F}(X^n)|$$

$$\mathbb{P}\left\{\Delta_n(Z^n) \ge \mathbb{E}\Delta_n(Z^n) + \frac{Bt}{\sqrt{n}}\right\} \le e^{-2t^2}$$

– McDiarmid: $0 \le \varphi(-yf(x)) \le B$

$\Delta_n(Z^n)$ has bdd dif. $c_1 = \dots = c_n = B/n$

$$\mathbb{P}\left\{ \Delta_n(z^n) \geq \mathbb{E}\Delta_n(z^n) + \frac{Bt}{\sqrt{n}} \right\}$$

$$\leq \exp\left(- \frac{2B^2t^2}{n \cdot \frac{B^2}{n}} \right)$$; by McDiarmid.