Regularization and Convex Surrogate Loss

§1 Motivation

① $\mathcal{H} = \{\mathcal{H}_k\}_{k \in \mathbb{N}}$ s.t. $\mathcal{H}_k \subseteq \mathcal{H}_{k+1}$.

$\hat{R}_S(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{h(x_i) \neq y_i\}}$.

② • Extension to $\mathcal{H}_r$, where $r \in \mathbb{R}_{\geq 0}$: (→ Regularization)

$\{\mathcal{H}_r\}_{r \in \mathbb{R}_{\geq 0}}$ s.t. $\mathcal{H}_r$ varies "Continuously" w.r.t. r.

• Alternative of $\hat{R}_S(h)$ that is easier to optimize.

↳ Convex surrogate loss.

§2 Regularization

① $\{\mathcal{H}_r\}_{r \geq 0}$

• $\mathcal{H}_r = \{ x \mapsto \underbrace{\operatorname{sgn}(\langle w, \Phi(x) \rangle)}_{h_w(x)} \mid w \in \mathbb{R}^k, \|w_p\| \leq r \}$.

$\Phi : \mathcal{X} \to \mathbb{R}^k$ ... feature map.

$\|w\|_p := \left(\sum_{i=1}^{k} |w_i|^p\right)^{\frac{1}{p}}$.

• $\lambda > 0$: Hyperparameter.

$w \in \operatorname{argmin}_w \hat{R}_S(h_w) + \lambda \|w\|_p$

§3 Convex Surrogate Loss

① Assumption

$\forall h \in \mathcal{H}, \exists f : \mathcal{X} \to \mathbb{R}$ s.t. $h(x) = \operatorname{sgn}(f(x))$.

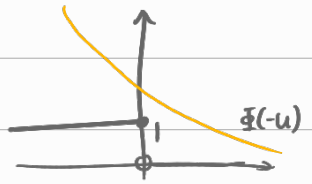
Denote $h_f := \operatorname{sgn} \circ f$.

$\operatorname{sgn}(s) = \begin{cases} +1 & \text{if } s \geq 0 \\ -1 & \text{otherwise} \end{cases}$

/ /

② Convex Surrogate Loss: $\Phi : \mathbb{R} \to \mathbb{R}$ s.t.

- Φ is convex
- Φ is nondecreasing
- $\forall u \in \mathbb{R}, \mathbb{1}_{\{u \le 0\}} \le \Phi(-u)$.



$\Phi(u) = \max(0, 1+u)$ ··· hinge loss

$\Phi(u) = \exp(u)$ ··· exponential loss.

$I(u) = \log_2(1+e^u)$ ··· logistic loss.

③ Why $\mathbb{1}_{\{u \le 0\}} \le \Phi(-u)$?

- $\hat{R}_S(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{h(x_i) \neq y_i\}}$
- $\mathbb{1}_{\{h_f(x_i) \neq y_i\}} = \mathbb{1}_{\{\mathrm{sgn}(f(x_i)) \neq y_i\}}$
  $\le \mathbb{1}_{\{f(x_i)y_i \le 0\}}$
  $\le \Phi(-f(x_i)y_i)$

Thus, $\hat{R}_S(h) \le \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{h_f(x_i) \neq y_i\}} \le \frac{1}{m}\sum_{i=1}^{m} \Phi(-f(x_i)y_i) =: \hat{R}_S^{\Phi}(f)$.

Lemma $\hat{R}_S^{\Phi}$ is convex.

(pf) $f, g \in [\mathcal{X} \to \mathbb{R}]$, $\alpha \in (0,1)$, then $\hat{R}_S^{\Phi}(\alpha f + (1-\alpha)g) \le \alpha \hat{R}_S^{\Phi}(f) + (1-\alpha)\hat{R}_S^{\Phi}(g)$.

$$\begin{aligned}
\hat{R}_S^{\Phi}(\alpha f + (1-\alpha)g) &= \frac{1}{m}\sum_{i=1}^{m} \Phi(-(\alpha f(x_i) + (1-\alpha)g(x_i))y_i) \\
&= \frac{1}{m}\sum_{i=1}^{m} \Phi(\alpha(-f(x_i)y_i) + (1-\alpha)(-g(x_i)y_i)) \\
&\le \frac{1}{m}\sum_{i=1}^{m} [\alpha\Phi(-f(x_i)y_i) + (1-\alpha)\Phi(-g(x_i)y_i)] \\
&= \alpha\hat{R}_S^{\Phi}(f) + (1-\alpha)\hat{R}_S^{\Phi}(g).
\end{aligned}$$

④ Bayes' Scoring Function $f^* \cdots R$ $\quad R(h_{f^*}) = R(f^*)$ minimal

Let $\mathop{\mathbb{E}}_{S\sim D^m} \hat{R}_S^{\Phi}(f) = R^{\Phi}(f)$. $\quad R^{\Phi}(f_{\Phi}^*)$ minimal

$R(f^*)$
=

$R(h_f) - R(h_{Bayes})$ (Excess Error)] Relation?
$R^\Phi(h_f) - R^\Phi(f_\Phi^*)$

Proposition

$f^*(x) = \eta(x) - \frac{1}{2}$, where $\eta(x) = \mathbb{P}(y=1|x)$.

$f_\Phi^*(x) = \underset{u\in[-\infty,\infty]}{\operatorname{argmin}}\ \eta(x)\Phi(-u) + (1-\eta(x))\Phi(u)$

(pf)

① $R(h_f) \geq R(h_{f^*})$ $\forall f$.

$\because) R(h_f) = \mathbb{E}[\mathbb{1}_{\{h_f(x)\neq y\}}]$

$= \mathbb{E}[\mathbb{E}[\mathbb{1}_{h_f(x)\neq y}|x]]$

$= \mathbb{E}[\mathbb{P}(\operatorname{sgn}(f(x))\neq y)|x]$

$\geq \mathbb{E}[\mathbb{P}(\operatorname{sgn}(f^*(x))\neq y)|x]$

$= \mathbb{E}[\mathbb{E}[\mathbb{1}_{h_{f^*}(x)\neq y}|x]]$

$= \mathbb{E}[\mathbb{1}_{\{h_{f^*}(x)\neq y\}}]$

$= R(h_{f^*})$

$\to \mathbb{E}[\mathbb{1}_{h_f(x)=1}\times\mathbb{1}_{y=-1}|x] + \mathbb{E}[\mathbb{1}_{h_f(x)=-1}\times\mathbb{1}_{y=1}|x]$

$(1-\eta(x))\mathbb{1}_{\{h(x)=1\}} + \eta(x)\mathbb{1}_{\{h(x)=-1\}}$

$2\min(1-\eta(x), \eta(x))$.

$(\because) \eta(x) > \frac{1}{2} \to \operatorname{sgn}(f^*(x)) = 1$

$\mathbb{P}(y=1|x) > \frac{1}{2}$.

$\Rightarrow \mathbb{P}(y=1|x) > \mathbb{P}(y=-1|x)$

$\Rightarrow \mathbb{P}(\operatorname{sgn}(f^*(x)) = y|x)$

$\geq \mathbb{P}(\operatorname{sgn}(f(x)) = y|x)$,

given $\operatorname{sgn}(f(x)) \in \{\pm 1\}$.

② $R^\Phi(f) \geq R^\Phi(f_\Phi^*)$ $\forall f$.

$\because) R^\Phi(f) = \mathbb{E}[\Phi(-yf(x))]$

$= \mathbb{E}[\mathbb{E}[\Phi(-yf(x))|x]]$

It suffices to prove $\mathbb{E}[\Phi(-yf(x))|x] \geq \mathbb{E}[\Phi(-yf_\Phi^*(x))|x]$ for $\forall x$.

For any u, $\eta(x)\Phi(-f_\Phi^*(x)) + (1-\eta(x))\Phi(f_\Phi^*(x)) \leq \eta(x)\Phi(-u) + (1-\eta(x))\Phi(u)$.

$\mathbb{E}[\Phi(-yf(x))|x] = \mathbb{E}[\Phi(-yf(x))\mathbb{1}_{\{y=1\}}|x] + \mathbb{E}[\Phi(-yf(x))\mathbb{1}_{\{y=-1\}}|x]$

$= \mathbb{E}[\Phi(-f(x))\mathbb{1}_{\{y=1\}}|x] + \mathbb{E}[\Phi(f(x))\mathbb{1}_{\{y=-1\}}|x]$

$= \Phi(-f(x))\eta(x) + \Phi(f(x))(1-\eta(x))$

$\geq \eta(x)\Phi(-f_\Phi^*(x)) + (1-\eta(x))\Phi(f_\Phi^*(x))$.

$= \mathbb{E}[\Phi(-yf_\Phi^*(x))|x]$.

③ $\Phi_{\pm}(x) = \max(0, 1 \pm x)$, $\Phi(x) = \exp(x)$.

$$\min_u \ \eta(x)\max(0, 1-u) + (1-\eta(x))\max(0, 1+u).$$

$u = 1 \Rightarrow 2(1-\eta(x))$

$u = -1 \Rightarrow 2\eta(x)$.

$$\Rightarrow f_{\pm}^*(x) = \begin{cases} +1 & \text{if } \eta(x) \geq \frac{1}{2} \\ -1 & \text{if } \eta(x) < \frac{1}{2}. \end{cases}$$

$\Phi(u) = \exp(u)$. $\eta(x)e^{-u} + (1-\eta(x))e^{u}$.

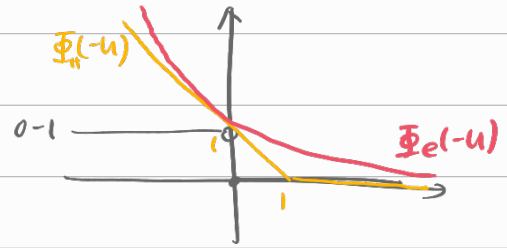
$\frac{\partial}{\partial u}(\ ) = 0 \Rightarrow f_{\exp}^*(x) = \frac{1}{2}\log\frac{\eta(x)}{1-\eta(x)}$

Oct / 24 / 2023

Examples

① $\Phi_H(u) = \max\{0, 1+u\}$ (Hinge loss)

② $\Phi_e(u) = \exp(u)$ (Exponential loss)



(Relation Between SVM and Hinge Loss)

$\mathcal{H} = \{x \mapsto \text{sgn}(\langle w, x\rangle + b) \mid w \in \mathbb{R}^n, b \in \mathbb{R}\}$.

$\underset{w,b}{\text{argmin}}\ \hat{R}_S^{\Phi_H}(x \mapsto \langle w, x\rangle + b) + \frac{1}{2}\|w\|^2$.

(3)

Risk

$$R(f) = R(h_f) = \mathbb{E}_S[\hat{R}_S(h_f)] = \mathbb{E}_{(x,y)\sim D}[\mathbb{1}_{\text{sgn}(f(x)) \neq y}]$$

Bayes' scoring fn f^* s.t. $h_{f^*} = h^*$.

$f^*(x) = \eta(x) - \frac{1}{2}$, where $\eta(x) = \mathbb{P}[y=1 \mid x]$

Bayes' error $R^* = R(h^*) = R(h_{f^*})$

Φ-Risk

$$R^\Phi(f) = \mathbb{E}_{(x,y)\sim D}[\Phi(-f(x)y)]$$

$$f_\Phi^*(x) = \underset{u \in [-\alpha, \alpha]}{\text{argmin}}\ \eta(x)\Phi(-u) + (1-\eta(x))\Phi(u).$$

Then, $R^\Phi(f_\Phi^*) \leq R^\Phi(g)$ $\forall g$.

$\| $

$R^{\Phi,*}$

§ Excess Error

$R(f) - R^*$ v.s. $R^{\Phi}(f) - R^{\Phi,*}$.

Theorem 4.7 $\exists\, s \geq 1,\ c > 0$ s.t. $\forall x \in \mathcal{X}$ $|f^*(x)|^s \leq c^s \left[\Phi(0) - \mathbb{E}_{P[y|x]}\left[\Phi(-y f^{\Phi,*}(x))\right]\right]$

Then for $\forall f \in [\mathcal{X} \to \mathbb{R}]$,
$(R(f) - R^*)^s \leq (2c)^s \left[R^{\Phi}(f) - R^{\Phi,*}\right]$

example) Hinge loss $s=1,\ c=\frac{1}{2}$
Exponential loss $s=2,\ c=\frac{1}{\sqrt{2}}$.

Remark
$f^*(x) = \eta(x) - \frac{1}{2}$
$\eta(x) = P[y=1|x]$
$f^{\Phi,*}(x) = \operatorname{argmin}_u \eta(x)\Phi(-u) + (1-\eta(x))\Phi(u)$.

Lemma 4.5 $R(f) - R^* \leq 2 \mathbb{E}_{x \sim D_{\mathcal{X}}}\left[|f^*(x)| \mathbb{1}_{[f(x) f^*(x) \leq 0]}\right]$

(pf of lemma)
$R(f) - R^* = \mathbb{E}_{(x,y)\sim D}\left[\mathbb{1}_{[f(x)\neq y]} - \mathbb{1}_{[f^*(x)\neq y]}\right]$
$\leq \eta(x)\, \mathbb{E}[$
$+ (1-\eta(x))\, \mathbb{E}$

See below

(pf of thm 4.7) ① $\Phi(-2f(x)f^*(x)) \leq \mathbb{E}_{P[y|x]}\left[\Phi(-yf(x))\right]$
$\|$

°°) $\Phi\left(-2\left(\eta(x) - \frac{1}{2}\right) f(x)\right) = \Phi\left((1-2\eta(x)) f(x)\right)$
$= \Phi\left(\eta(x)(-f(x)) + (1-\eta(x))(f(x))\right)$
$\leq \eta(x)\Phi(-f(x)) + (1-\eta(x))\Phi(-(-f(x)))$

$$= \underset{\mathbb{P}[y|x]}{\mathbb{E}}[\Phi(-yf(x))]$$

② $R(f) - R^* \le 2 \mathbb{E}_x[|f^*(x)| \mathbb{1}_{\{f(x)f^*(x) \le 0\}}]$ (lemma 4.5)

$= 2 \mathbb{E}_x[(|f^*(x)|^s \mathbb{1}_{\{\cdot,\cdot\}})^{1/s}]$ $(\mathbb{1}_{\{\cdot,\cdot\}} \in \{0,1\})$.

$\le 2 (\mathbb{E}_x[|f^*(x)|^s \mathbb{1}_{\{\cdot,\cdot\}}])^{1/s}$ (Jensen for concave).

$\le 2 (\mathbb{E}_x[c^s(\Phi(0) - \underset{y|x}{\mathbb{E}}[\Phi(-yf^{\Phi,*}(x))]) \mathbb{1}_{\{\cdot,\cdot\}}])^{1/s}$

(By assumptions)

$\le 2 (\mathbb{E}_x[c^s(\Phi(-f(x)f^*(x)) - \underset{y|x}{\mathbb{E}}[\Phi(-yf^{\Phi,*}(x))]) \mathbb{1}_{\{\cdot,\cdot\}}])^{1/s}$

(Φ is nondecreasing)

$\le 2 (\mathbb{E}_x[c^s(\Phi(-f(x)f^*(x)) - \underset{y|x}{\mathbb{E}}[\Phi(-yf^{\Phi,*}(x))])])^{1/s}$

$\le 2 (\mathbb{E}_x[c^s \underset{y|x}{\mathbb{E}}[\Phi(-yf(x)) - \Phi(-yf^{\Phi,*}(x))]])^{1/s}$ (By ①)

$= 2c (\underset{(x,y)\sim D}{\mathbb{E}}[\Phi(-yf(x)) - \Phi(-yf^{\Phi,*}(x))])^{1/s}$

$= 2c (R^{\Phi}(f) - R^{\Phi,*})^{1/s}$

(pf of lemma4.3) ① $g \in [\mathcal{X} \to \mathbb{R}]$, $R(g) = \underset{x\sim D_x}{\mathbb{E}}[2f^* \mathbb{1}_{\{g(x)<0\}} + (1-\eta(x))]$

$\because) R(g) = \underset{x\sim D_x}{\mathbb{E}}[\underset{y|x}{\mathbb{E}}[\mathbb{1}_{\{y \ne \mathrm{sgn}(g(x))\}}]]$

$= \underset{x}{\mathbb{E}}[\underset{y|x}{\mathbb{E}}[\mathbb{1}_{\{y=1\}} \mathbb{1}_{\{g(x)<0\}} + \mathbb{1}_{\{y=0\}} \mathbb{1}_{\{g(x)\ge 0\}}]]$

$= \underset{x}{\mathbb{E}}[\eta(x) \mathbb{1}_{\{g(x)<0\}} + (1-\eta(x))(1 - \mathbb{1}_{\{g(x)<0\}})]$

$= \mathbb{E}_x[2\eta(x) - \frac{1}{2}) \mathbb{1}_{\{g(x)<0\}} + (1-\eta(x))]$

$= \mathbb{E}_x[2f^*(x) \mathbb{1}_{\{g(x)<0\}} + (1-\eta(x))]$.

② $R(f) - R^* = \mathop{\mathbb{E}}_{x \sim D_x} [2f^*(x) \mathbb{1}_{\{f(x) < 0\}} - 2f^*(x) \mathbb{1}_{\{f^*(x) < 0\}}]$ by ①

$= 2\mathbb{E}[\, f^*(x) (\mathbb{1}_{\{f(x) < 0\}} - \mathbb{1}_{\{f^*(x) < 0\}})]$

$\le 2\mathbb{E}[\mathbb{1}_{\{f^*(x) \ge 0\}} f^*(x) (\mathbb{1}_{\{f(x) < 0\}} - \mathbb{1}_{\{f^*(x) < 0\}})$

$+ \mathbb{1}_{\{f^*(x) < 0\}} f^*(x) (\quad " \quad)]$

$= 2\mathbb{E}[\mathbb{1}_{\{f^*(x) \ge 0\}} |f^*(x)| (\mathbb{1}_{\{f(x) < 0\}} - \mathbb{1}_{\{f^*(x) < 0\}})$

$+ \mathbb{1}_{\{f^*(x) < 0\}} |f^*(x)| (\mathbb{1}_{\{f^*(x) < 0\}} - \mathbb{1}_{\{f(x) < 0\}})]$

$= 2\mathbb{E}[\mathbb{1}_{\{f^*(x) \ge 0\}} |f^*(x)| \mathbb{1}_{\{f(x) f^*(x) < 0\}}$

$+ \mathbb{1}_{\{f^*(x) < 0\}} |f^*(x)| \mathbb{1}_{\{f(x) f^*(x) < 0\}}]$

$= 2\mathbb{E}[|f^*(x)| \mathbb{1}_{\{f(x) f^*(x) < 0\}}]$.