Connection between classification and probability estimation

Miaoyan Wang, July 16, 2020

Since the hinge loss $L(t) = (1-t)_+$ is 1-Lipschitz function,

$$|L(yf(\boldsymbol{X})) - L(y\bar{f}_\pi(\boldsymbol{X}))| \leq |yf(\boldsymbol{X}) - y\bar{f}_\pi(\boldsymbol{X})| = |f(\boldsymbol{X}) - \bar{f}_\pi(\boldsymbol{X})|,$$

holds for any $(\boldsymbol{X}, y) \in \mathcal{X} \times \{-1, 1\}$. Let $S(y) = (1-\pi)\mathbb{1}\{y = 1\} + \pi\mathbb{1}\{y = -1\}$ be a positive function. Then

$$\mathbb{E}[S(y)L(yf(\boldsymbol{X})) - L(y\bar{f}_\pi(\boldsymbol{X}))] \leq \mathbb{E}S(y)|f(\boldsymbol{X}) - \bar{f}_\pi(\boldsymbol{X})|.$$

The left hand side is $e_V(f, \bar{f}_\pi)$ by definition, and the right hand side is upper bounded by $\mathbb{E}|f(\boldsymbol{X}) - \bar{f}_\pi(\boldsymbol{X})|$. Therefore,

$$e_V(f, \bar{f}_\pi) \leq \left\| f - \bar{f}_\pi \right\|_1.$$

Suppose f is

Proposition 1. *Let $\bar{f}_\pi \colon \mathcal{X} \to \{0, 1\}$ be the Bayes rule. For any sufficiently small $\delta > 0$,*

$$\sup_{\{f \in \mathcal{F}\colon e_V(f, \bar{f}_\pi) \leq \delta\}} \left\| \text{sign}(f) - \text{sign}(\bar{f}_\pi) \right\|_1 \geq \delta.$$

Proof. We proof by construction. Define $\text{sign}(f) = -\text{sign}(\bar{f}_\pi)$ at a interval with measure δ, and $\text{sign}(f) = \text{sign}(\bar{f}_\pi)$ otherwise. We aim to find $f \in \mathcal{F}$ such that

$$\begin{cases} e_V(f, \bar{f}_\pi) = \mathbb{E}\left[(1 - yf(\boldsymbol{X}))_+ - (1 - y\bar{f}_\pi(\boldsymbol{X}))_+\right] \leq \delta, \\ \mu\{\boldsymbol{X} \colon \text{sign}[f(\boldsymbol{X})] \neq \text{sign}[\bar{f}_\pi(\boldsymbol{X})]\} \geq \delta. \end{cases}$$

$$e_V(f, \bar{f}_\pi) = \mathbb{P}(y = 1)\mathbb{E}[(1 - f(\boldsymbol{X}))_+ - (1 - \bar{f}_\pi(\boldsymbol{X}))_+ | y = 1] + \mathbb{P}(y = -1)\mathbb{E}[(1 + f(\boldsymbol{X}))_+ - (1 + \bar{f}_\pi(\boldsymbol{X}))_+ | y = -1]$$

$$\begin{aligned} \text{True Positive (TP)} &= \{(\boldsymbol{X}, y) \colon \bar{f}_\pi(\boldsymbol{X}) = y = 1\}, \\ \text{True Negative (TN)} &= \{(\boldsymbol{X}, y) \colon \bar{f}_\pi(\boldsymbol{X}) = y = -1\}, \\ \text{False Positive (FP)} &= \{(\boldsymbol{X}, y) \colon \bar{f}_\pi(\boldsymbol{X}) = 1, y = -1\}, \\ \text{False Negative (FN)} &= \{(\boldsymbol{X}, y) \colon \bar{f}_\pi(\boldsymbol{X}) = -1, y = 1\}. \end{aligned}$$

Then $\bar{f}_\pi$

$$e_V(f, \bar{f}_\pi) = \mathbb{E}[(1 - f(\boldsymbol{X}))_+ | \text{TP}]\mathbb{P}[(\boldsymbol{X}, y) \in \text{TP}] + \mathbb{E}[(1 + f(\boldsymbol{X}))_+ | \text{TN}]\mathbb{P}[(\boldsymbol{X}, y) \in \text{TN}]$$

$$
\begin{aligned}
&+ \mathbb{E}[(1+f(\boldsymbol{X}))_+ - 2|\mathrm{FP}]\mathbb{P}[(\boldsymbol{X},y)\in \mathrm{FP}] + \mathbb{E}[(1-f(\boldsymbol{X}))_+ - 2|\mathrm{FN}]\mathbb{P}[(\boldsymbol{X},y)\in \mathrm{FN}] \\
= {}& \mathbb{E}[(1-f(\boldsymbol{X}))_+|\mathrm{TP}]\mathbb{P}[(\boldsymbol{X},y)\in \mathrm{TP}] + \mathbb{E}[(1+f(\boldsymbol{X}))_+|\mathrm{TN}]\mathbb{P}[(\boldsymbol{X},y)\in \mathrm{TN}] \\
&+ \mathbb{E}[(1+f(\boldsymbol{X}))_+|\mathrm{FP}]\mathbb{P}[(\boldsymbol{X},y)\in \mathrm{FP}] + \mathbb{E}[(1-f(\boldsymbol{X}))_+|\mathrm{FN}]\mathbb{P}[(\boldsymbol{X},y)\in \mathrm{FN}] \\
&- 2\mathbb{P}[(\boldsymbol{X},y)\in \mathrm{FP}] - 2\mathbb{P}[(\boldsymbol{X},y)\in \mathrm{FN}]
\end{aligned}
$$

The Bayes rule $\bar{f}_\pi(\boldsymbol{X}) \in \{-1, 1, 0\}$ divides the input space into three disjoint regions (but possibly with 0 measure). Without loss of generality, assume $\mathbb{P}[\boldsymbol{X} : \bar{f}_\pi(\boldsymbol{X}) = -1] = c > 0$. For any sufficiently small $\delta \in (0, c)$, define

$$
f(\boldsymbol{X}) = \begin{cases} 0 & \boldsymbol{X} \in \mathcal{A} \subset \{\boldsymbol{X} \colon \bar{f}_\pi(\boldsymbol{X}) = -1\}, \\ \bar{f}_\pi(\boldsymbol{X}) & \text{otherwise.} \end{cases}
$$

Choose set $\mathcal{A}$ such that $\mathbb{P}(\boldsymbol{X} \in \mathcal{A}) = \delta$. Then $f(\boldsymbol{X})$ incurs error at most

$$
e_V(f, \bar{f}_\pi) \le \mathbb{E}[1|TN \cap \mathcal{A}]\mathbb{P}(\boldsymbol{X} \in \mathcal{A})\mathbb{P}(\boldsymbol{X}, y) \le \delta.
$$

Assume changing a measurable set

□

Proposition 2. *Let $\bar{f}_\pi \colon \mathcal{X} \to \{0, 1\}$ be the Bayes rule. Define weighted 0-1 excess risk*

$$
\mathrm{Risk}(f, \bar{f}_\pi) := \mathbb{E}\left[|\mathbb{P}(y = 1|\boldsymbol{X}) - \pi)||\mathrm{sign}(f) - \mathrm{sign}(\bar{f}_\pi)|\right].
$$

By proof below Remake 3, we have

$$
\mathrm{Risk}(f, \bar{f}_\pi) \le e_V(f, \bar{f}_\pi) \Rightarrow \{f \in \mathcal{F} \colon e_V(f, \bar{f}_\pi) \le \delta\} \subset \{f \in \mathcal{F} \colon \mathrm{Risk}(f, \bar{f}_\pi) \le \delta\}
$$

Therefore,

$$
\sup_{\{f \in \mathcal{F} \colon e_V(f, \bar{f}_\pi) \le \delta\}} \left\|\mathrm{sign}(f) - \mathrm{sign}(\bar{f}_\pi)\right\|_1 \le \sup_{\{f \in \mathcal{F} \colon \mathrm{Risk}(f, \bar{f}_\pi) \le \delta\}} \left\|\mathrm{sign}(f) - \mathrm{sign}(\bar{f}_\pi)\right\|_1.
$$

Proof.

$$
\begin{aligned}
e_V(f, \bar{f}_\pi) &= \mathbb{E}\left\{S(y)L(yf(\boldsymbol{X})) - S(y)L(y\bar{f}_\pi(\boldsymbol{X}))\right\} \\
&\ge \mathbb{E}\{S(y)\mathbb{1}\{y \ne \mathrm{sign} f(\boldsymbol{X})\} - S(y)\mathbb{1}\{y \ne \mathrm{sign}(\bar{f}(\boldsymbol{X}) - \pi)\} \\
&= \mathbb{E}\{S(y)y(\mathrm{sign}(f(\boldsymbol{X})) - \mathrm{sign}(\bar{f}(\boldsymbol{X}) - \pi)\} \\
&\ge 01\mathrm{Risk}(f) - 01\mathrm{Risk}(\bar{f}_\pi)
\end{aligned}
$$

□

$$\text{01Risk}(\bar{f}_\pi) = \mathbb{E}|\bar{f}(\boldsymbol{X}) - \pi|$$

$$\text{01Risk}(f) - \text{01Risk}(\bar{f}_\pi) = \mathbb{E}\left\{|\bar{f}(\boldsymbol{X}) - \pi|\mathbb{1}\{f(\boldsymbol{X}) \overset{\text{sign}}{\neq} (\bar{f}(\boldsymbol{X}) - \pi)\}\right\}$$

Definition 1 (Weighted Risk)**.**

$$\begin{aligned}\text{01Risk}(f) &:= (1-\pi) \times \text{False negative}(f) + \pi \times \text{False positive}(f)\\ &= (1-\pi)P\{(\boldsymbol{X}, y)\colon \text{sign}(f(\boldsymbol{X})) = -1,\ y = 1\} + \pi P\{(\boldsymbol{X}, y)\colon \text{sign}(f(\boldsymbol{X})) = 1, y = -1\}\\ &= \mathbb{E}\left\{S(y)\mathbb{1}\{\text{sign}(f(\boldsymbol{X})) \neq y\}\right\}\end{aligned}$$

$$e_V(f, \bar{f}_\pi) \geq R(yf(\boldsymbol{X})) - R(yf_\pi(\boldsymbol{X}))$$

1 Previous results

- Convergence result from probability estimation:

  $$\mathbb{P}\{e_V(\hat{f}_\pi, f_\pi^*) \geq \delta_n\} \leq C\exp\left\{-an(\lambda J_\pi^*)^{2-\beta}\right\},$$

  where

  $$e_V(f_1, f_2) := \mathbb{E}\{V(f_1, \boldsymbol{Z}) - V(f_2, \boldsymbol{Z})\}$$

  is the excess risk, and

  $$V(f, \boldsymbol{Z}) = V(f, (\boldsymbol{X}, y)) = \begin{cases}(1-\pi)(1 - f(\boldsymbol{X}))_+, & \text{if } y = 1,\\ \pi(1 + f(\boldsymbol{X}))_+, & \text{if } y = -1.\end{cases}$$

  is the weighted margin loss.

- Convergence result from classification:

  With probability at least $1 - \delta$, we have

  $$\underbrace{\mathbb{P}[Y \neq \text{sign } f^*(\boldsymbol{X})] - \mathbb{P}[Y \neq \text{sign } \hat{f}(\boldsymbol{X})]}_{\text{statistical error for estimating } f} \leq 4MG\sqrt{\frac{r}{n}} + \sqrt{\frac{\log(\frac{1}{\delta})}{2n}}.$$