Machine Learning — Statistical Methods for Machine Learning

Surrogate loss functions

Instructor: *Nicolò Cesa-Bianchi* version of June 8, 2021

Surrogate losses $\ell : \{-1,1\} \times \mathbb{R} \to \mathbb{R}$ are convex upper bounds on the zero-one loss function for binary classification. We already encountered two of them:

- Hinge loss $\ell(y,\widehat{y}) = \big[1 - y\,\widehat{y}\big]_+$
- Boosting loss $\ell(y,\widehat{y}) = e^{-y\,\widehat{y}}$

where $y \in \{-1,1\}$ and $\widehat{y} \in \mathbb{R}$.

As many surrogate losses exist, we may wonder whether some should be preferred over the others. We now define an important criterion, called **consistency**, that a surrogate loss may satisfy with respect to the function $\eta(\boldsymbol{x}) = \mathbb{P}\big(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}\big)$ which defines the Bayes optimal predictor f^*.

A surrogate loss function ℓ is consistent if, for all $\boldsymbol{x} \in \mathcal{X}$,

$$\operatorname{sgn}\big(y^*_{\boldsymbol{x}}\big) = f^*(\boldsymbol{x}) \qquad \text{for} \qquad y^*_{\boldsymbol{x}} = \operatorname*{argmin}_{\widehat{y}\in\mathbb{R}} \mathbb{E}\big[\ell(Y,\widehat{y}) \mid \boldsymbol{X} = \boldsymbol{x}\big]$$

In other words, the sign of the prediction minimizing the conditional risk with respect to the surrogate loss must be equal to the Bayes optimal classification for the zero-one loss.

We now verify the consistency of the hinge loss. We have

$$\begin{aligned}
y^*_{\boldsymbol{x}} &= \operatorname*{argmin}_{\widehat{y}\in\mathbb{R}} \Big(\eta(\boldsymbol{x})\big[1-\widehat{y}\big]_+ + \big(1-\eta(\boldsymbol{x})\big)\big[1+\widehat{y}\big]_+\Big) \\
&= \operatorname*{argmin}_{\widehat{y}\in[-1,+1]} \Big(\eta(\boldsymbol{x})\big[1-\widehat{y}\big]_+ + \big(1-\eta(\boldsymbol{x})\big)\big[1+\widehat{y}\big]_+\Big) \\
&= \operatorname*{argmin}_{\widehat{y}\in[-1,+1]} \Big(1 + \big(1-2\eta(\boldsymbol{x})\big)\widehat{y}\Big) \\
&= \begin{cases} -1 & \text{if } \eta(\boldsymbol{x}) \le 1/2, \\ +1 & \text{otherwise} \end{cases} \\
&= f^*(\boldsymbol{x})
\end{aligned}$$

In the second inequality, we could replace $\widehat{y} \in \mathbb{R}$ with $\widehat{y} \in [-1,+1]$ because both functions $[1-\widehat{y}]_+$ and $[1+\widehat{y}]_+$ increase or remain constant outside of the interval $[-1,+1]$.

More generally, the following result holds.

Theorem 1. *If a surrogate loss $\ell : \{-1,1\} \times \mathbb{R} \to \mathbb{R}$ is such that for all $y \in \{-1,1\}$ the function $\ell(y,\cdot)$ is convex, differentiable at zero, and satisfies $\ell'(y,0) < 0$, then ℓ is consistent.*

Besides the hinge loss, the boosting loss, the square loss $\ell(y,\widehat{y}) = \big(1 - y\,\widehat{y}\big)^2$ and the quadratic hinge loss $\ell(y,\widehat{y}) = \big([1 - y\,\widehat{y}]_+\big)^2$ are also consistent.