Machine Learning — Statistical Methods for Machine Learning

Logistic regression and surrogate loss functions

Instructor: *Nicolò Cesa-Bianchi* version of June 11, 2022

In certain application domains, such as weather prediction, one typically prefers to output a probability (e.g., the chance of rain) instead of a binary prediction (e.g., it will rain). This task corresponds to the problem of learning the function $\eta(\boldsymbol{x}) = \mathbb{P}(Y = 1 \mid X = \boldsymbol{x})$ in a binary classification problem. A popular approach to do that is known as **logistic regression**: we train a predictor $g : \mathcal{X} \to \mathbb{R}$ and then use $\sigma\big(g(\boldsymbol{x})\big)$ to predict $\eta(\boldsymbol{x})$. The function $\sigma : \mathbb{R} \to \mathbb{R}$, called logistic, is defined by

$$\sigma(z) = \frac{1}{1 + e^{-z}} \in (0, 1)$$

Because we estimate a probability, an appropriate loss function is the logarithmic loss (here we use logarithms in base 2 for convenience),

$$\ell(y, \widehat{y}) = \mathbb{I}\{y = +1\} \log_2 \frac{1}{\widehat{y}} + \mathbb{I}\{y = -1\} \log_2 \frac{1}{1 - \widehat{y}}$$

Noting that $1 - \sigma(z) = \sigma(-z)$, we can write the identity

$$\mathbb{I}\{y = +1\} \log_2 \frac{1}{\widehat{y}} + \mathbb{I}\{y = -1\} \log_2 \frac{1}{1 - \widehat{y}} = \log_2 \left(1 + e^{-yg(\boldsymbol{x})}\right)$$

where $\widehat{y} = \sigma\big(g(\boldsymbol{x})\big)$. The right-hand side of the above identity is a function known as **logistic loss**, and is typically defined using $\widehat{y} = g(\boldsymbol{x})$ as follows,

$$\ell(y, \widehat{y}) = \log_2 \left(1 + e^{-y\,\widehat{y}}\right)$$

We now describe the important case of logistic regression when $g(\boldsymbol{x})$ is a linear model $\boldsymbol{w}^\top \boldsymbol{x}$. Given a training set $S = \big\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\big\}$, let $\ell_t(\boldsymbol{w}) = \log_2 \left(1 + e^{-y_t \boldsymbol{w}^\top \boldsymbol{x}_t}\right)$, we show how to compute $\nabla \ell_t(\boldsymbol{w})$. Let $s_t = \boldsymbol{w}^\top \boldsymbol{x}_t$. First, observe that

$$\frac{d}{ds_t} \log_2 \left(1 + e^{-y_t s_t}\right) = \frac{1}{\ln 2} \frac{-y_t e^{-y_t s_t}}{1 + e^{-y_t s_t}} = \frac{1}{\ln 2} \frac{-y_t}{1 + e^{y_t s_t}} = \frac{-y_t\, \sigma(-y_t s_t)}{\ln 2}$$

Therefore,

$$\nabla \ell_t(\boldsymbol{w}) = \left(\frac{d}{ds_t} \log_2 \left(1 + e^{-y_t s_t}\right) \Big|_{s_t = \boldsymbol{w}^\top \boldsymbol{x}_t} \right) \boldsymbol{x}_t = \frac{-\sigma\big(-y_t \boldsymbol{w}^\top \boldsymbol{x}_t\big)}{\ln 2} y_t \boldsymbol{x}_t \ .$$

The gradient descent update can then be written as

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \eta_t \sigma\big(-y_t \boldsymbol{w}^\top \boldsymbol{x}_t\big) y_t \boldsymbol{x}_t$$

where we hid the $\ln 2$ factor in the learning rate η_t.

To avoid overfitting, logistic regression is often used with a regularization term that enforces stability,

$$\ell_t(\boldsymbol{w}) = \log_2\left(1 + e^{-y_t \boldsymbol{w}^\top \boldsymbol{x}_t}\right) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \ .$$

If we run stochastic gradient descent using regularized logistic regression we get an algorithm similar to Pegasos for regularized hinge loss.

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 three of them:

- Hinge loss $\ell(y, \widehat{y}) = \left[1 - y\,\widehat{y}\right]_+$
- Boosting loss $\ell(y, \widehat{y}) = e^{-y\,\widehat{y}}$
- Logistic loss $\ell(y, \widehat{y}) = \log_2\left(1 + e^{-y\,\widehat{y}}\right)$

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

As many surrogate losses exist, we may wonder whether some of them 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 $\boldsymbol{x} \in \mathcal{X}$,

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

In other words, the sign of the Bayes optimal predictor for the surrogate loss must be the Bayes optimal classifier for the zero-one loss.

We now verify the consistency of the logistic loss. By taking derivatives, it is easy to check that

$$g^*(\boldsymbol{x}) = \operatorname*{argmin}_{\widehat{y} \in \mathbb{R}} \Big(\eta(\boldsymbol{x}) \log_2\left(1 + e^{-\widehat{y}}\right) + \big(1 - \eta(\boldsymbol{x})\big) \log_2\left(1 + e^{\widehat{y}}\right)\Big) = \ln \frac{\eta(\boldsymbol{x})}{1 - \eta(\boldsymbol{x})}$$

which implies

$$\operatorname{sgn}\big(g^*(\boldsymbol{x})\big) = \operatorname{sgn}\left(\ln \frac{\eta(\boldsymbol{x})}{1 - \eta(\boldsymbol{x})}\right) = \operatorname{sgn}\big(\eta(\boldsymbol{x}) - \tfrac{1}{2}\big) = f^*(\boldsymbol{x})$$

The Bayes optimal prediction $g^*(\boldsymbol{x}) = \ln \frac{\eta(\boldsymbol{x})}{1-\eta(\boldsymbol{x})}$ for the logistic loss is known as *log-odds ratio*. If we compute the conditional Bayes risk of g^* with respect to the logistic loss we get

$$\mathbb{E}\Big[\log_2\left(1 + e^{-Y\,g^*(\boldsymbol{x})}\right) \,\Big|\, \boldsymbol{X} = \boldsymbol{x}\Big] = -\eta(\boldsymbol{x}) \log_2 \eta(\boldsymbol{x}) + \big(1 - \eta(\boldsymbol{x})\big) \log_2\big(1 - \eta(\boldsymbol{x})\big)$$

The quantity on the right-hand side is the conditional entropy $H\big(Y \mid \boldsymbol{X} = \boldsymbol{x}\big)$ of Y given $\boldsymbol{X}$. This corresponds to the expected number of bits that we receive by observing Y when X is already known. From the conditional Bayes risk, we can easily obtain the Bayes risk,

$$\ell_{\mathcal{D}}(g^*) = \mathbb{E}\Big[\log_2\left(1 + e^{-Y\,g^*(\boldsymbol{X})}\right)\Big] = H(Y)$$

In other words, the entropy $H(Y)$ of the label Y is the Bayes risk for the logistic loss.

Next, we verify the consistency of the hinge loss. We have

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

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 (Sufficient condition for consistency of a surrogate loss)**.** *If a surrogate loss $\ell : \{-1,1\} \times \mathbb{R} \to \mathbb{R}$ is such that for all $y \in \{-1,1\}$ the derivative $\ell'(y,0)$ exists and satisfies $\ell'(y,0) < 0$, then ℓ is consistent.*

Besides the hinge loss and the logistic loss, also 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 all consistent.