

Logistic regression and surrogate loss functions

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 $\eta(\mathbf{x}) = \mathbb{P}(Y = 1 \mid X = \mathbf{x})$ in a binary classification problem. A popular approach to do that is known as **logistic regression**: we train a predictor $g : \mathcal{X} \rightarrow \mathbb{R}$ and then use $\sigma(g(\mathbf{x}))$ to predict $\eta(\mathbf{x})$. The function $\sigma : \mathbb{R} \rightarrow \mathbb{R}$, 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),

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

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

$$\mathbb{I}\{y = +1\} \log_2 \frac{1}{\hat{y}} + \mathbb{I}\{y = -1\} \log_2 \frac{1}{1 - \hat{y}} = \log_2 \left(1 + e^{-y\hat{y}} \right)$$

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

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

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

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

The gradient descent update can then be written as

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

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,

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

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

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

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

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

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 $\eta(\mathbf{x}) = \mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x})$ which defines the Bayes optimal predictor f^* .

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

$$\text{sgn}(g^*) = f^* \quad \text{for} \quad g^*(\mathbf{x}) = \underset{\hat{y} \in \mathbb{R}}{\text{argmin}} \mathbb{E}[\ell(Y, \hat{y}) \mid \mathbf{X} = \mathbf{x}]$$

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

$$g^*(\mathbf{x}) = \underset{\hat{y} \in \mathbb{R}}{\text{argmin}} \left(\eta(\mathbf{x}) \log_2(1 + e^{-\hat{y}}) + (1 - \eta(\mathbf{x})) \log_2(1 + e^{\hat{y}}) \right) = \ln \frac{\eta(\mathbf{x})}{1 - \eta(\mathbf{x})}$$

which implies

$$\text{sgn}(g^*(\mathbf{x})) = \text{sgn} \left(\ln \frac{\eta(\mathbf{x})}{1 - \eta(\mathbf{x})} \right) = \text{sgn}(\eta(\mathbf{x}) - \frac{1}{2}) = f^*(\mathbf{x})$$

The Bayes optimal prediction $g^*(\mathbf{x}) = \ln \frac{\eta(\mathbf{x})}{1 - \eta(\mathbf{x})}$ 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

$$\mathbb{E} \left[\log_2 \left(1 + e^{-Y g^*(\mathbf{x})} \right) \mid \mathbf{X} = \mathbf{x} \right] = -\eta(\mathbf{x}) \log_2 \eta(\mathbf{x}) - (1 - \eta(\mathbf{x})) \log_2 (1 - \eta(\mathbf{x}))$$

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

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

The quantity on the right-hand side is now the conditional entropy $H(Y \mid \mathbf{X})$ of the label Y given \mathbf{X} , which corresponds the Bayes risk for the logistic loss.

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

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

In the second inequality, we could replace $\hat{y} \in \mathbb{R}$ with $\hat{y} \in [-1, +1]$ because both functions $[1 - \hat{y}]_+$ and $[1 + \hat{y}]_+$ 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 $\ell : \{-1, 1\} \times \mathbb{R} \rightarrow \mathbb{R}$ 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 $\ell(y, \hat{y}) = (1 - y\hat{y})^2$ and the quadratic hinge loss $\ell(y, \hat{y}) = ([1 - y\hat{y}]_+)^2$ are all consistent.