The fundamental nature of the log loss function

Vladimir Vovk
v.vovk@rhul.ac.uk

February 24, 2015

Abstract

The standard loss functions used in the literature on probabilistic prediction are the log loss function and the Brier loss function; however, any proper loss function can be used for comparison of prediction algorithms. This note shows that the log loss function is most selective in that any prediction algorithm that is optimal for a given data sequence (in the sense of the algorithmic theory of randomness) under the log loss function will be optimal under any computable proper mixable loss function; on the other hand, there is a data sequence and a prediction algorithm that is optimal for that sequence under the Brier loss function but not under the log loss function.

1 Introduction

In the empirical work on probabilistic prediction (see, e.g., [1]) two main loss functions are usually used: log loss and Brier loss. It is important to understand which of these loss functions is likely to lead to better prediction algorithms. We formalize this question using a generalization of the notion of Kolmogorov complexity called predictive complexity (see, e.g., [2]). Our answer is that the log loss function is likely to lead to better prediction algorithms as it is more selective: if a prediction algorithm is optimal under the log loss function, it will be optimal under the Brier loss function, but the opposite implication is not true in general. This answer, however, is asymptotic and raises further questions.

2 Loss functions

We are interested in the problem of binary probabilistic prediction: the task is to predict a binary label \( y \in \{0, 1\} \) with a number \( p \in [0, 1] \); intuitively, \( p \) is the predicted probability that \( y = 1 \). The quality of the prediction \( p \) is measured by a loss function \( \lambda : [0, 1] \times \{0, 1\} \to \mathbb{R} \cup \{+\infty\} \). Intuitively, \( \lambda(p, y) \) is the loss suffered by a prediction algorithm that outputs a prediction \( p \) while the actual label is \( y \); the value \(+\infty\) (from now on abbreviated to \( \infty \)) is allowed. We will assume that \( \lambda(0, 0) = \lambda(1, 1) = 0 \), that the function \( \lambda(p, 0) \) is increasing
in $p$, that the function $\lambda(p, 0)$ is decreasing in $p$, and that $\lambda(p, y) < \infty$ unless $p \in \{0, 1\}$.

The prime examples of loss functions are the log loss function
\[
\lambda(p, y) := \begin{cases} 
- \log p & \text{if } y = 1 \\
- \log(1 - p) & \text{otherwise}
\end{cases}
\]
and the Brier loss function
\[
\lambda(p, y) := (y - p)^2.
\]
Both loss functions are proper, in that, for all $p, q \in [0, 1]$, \[ \mathbb{E}_p \lambda(p, \cdot) \leq \mathbb{E}_p \lambda(q, \cdot), \] where $\mathbb{E}_p f := pf(1) + (1 - p)f(0)$ for $f : \{0, 1\} \to \mathbb{R}$. They are also strictly proper, in that the inequality in (1) is strict whenever $q \neq p$. In principle, any strictly proper loss function can be used to measure the quality of probabilistic predictions.

A loss function $\lambda$ is called mixable if the set
\[
\left\{ (u, v) \in [0, 1]^2 \mid \exists p \in [0, 1] : u \leq e^{-\eta \lambda(p, 0)}, v \leq e^{-\eta \lambda(p, 1)} \right\}
\]
is convex for some $\eta \in (0, \infty)$. Both the log and Brier loss functions are mixable (see, e.g., [3], Section 2). All proper mixable loss functions are strictly proper.

### 3 Repetitive predictions

Starting from this section we consider the situation, typical in machine learning, where we repeatedly observe data $z_1, z_2, \ldots$ and each observation $z_t = (x_t, y_t) \in \mathbf{Z} = \mathbf{X} \times \{0, 1\}$ consists of an object $x_t \in \mathbf{X}$ and its label $y_t \in \{0, 1\}$. Let us assume, for simplicity, that $\mathbf{X}$ is a finite set, say a set of natural numbers.

A prediction algorithm is a computable function $F : \mathbf{Z} \times \mathbf{X} \to [0, 1]$; intuitively, given a data sequence $\sigma = (z_1, \ldots, z_T)$ and a new object $x$, $F$ outputs a prediction $F(\sigma, x)$ for the label of $x$. For any data sequence $\sigma = (z_1, \ldots, z_T)$ and loss function $\lambda$, we define the loss that $F$ suffers on $\sigma$ as
\[
\text{Loss}_F^\lambda(\sigma) := \sum_{t=1}^T \lambda(F(z_1, \ldots, z_{t-1}, x_t), y_t)
\]
(where $\infty + a$ is defined to be $\infty$ for any $a \in \mathbb{R} \cup \{\infty\}$). Functions $\text{Loss}_F^\lambda : \mathbf{Z}^* \to \mathbb{R}$ that can be defined this way for a given $\lambda$ are called loss processes under $\lambda$. In other words, $L : \mathbf{Z}^* \to \mathbb{R}$ is a loss process under $\lambda$ if and only if
\[
\forall \sigma \in \mathbf{Z}^* \forall x \in \mathbf{X} \exists p \in [0, 1] \forall y \in \{0, 1\} : L(\sigma, x, y) = L(\sigma) + \lambda(p, y). \quad (2)
\]
A function \( L : \mathbb{Z}^* \to \mathbb{R} \) is said to be a loss superprocess under \( \lambda \) if (2) holds with \( \geq \) in place of \( = \). If \( \lambda \) is computable and mixable, there exists a smallest, to within an additive constant, upper semicomputable loss superprocess:

\[
\exists L_1 \forall L_2 \exists c \in \mathbb{R} \forall \sigma \in \mathbb{Z}^* : L_1(\sigma) \leq L_2(\sigma) + c,
\]

where \( L_1 \) and \( L_2 \) range over upper semicomputable loss superprocesses under \( \lambda \). (For a precise statement and proof, see [2], Theorem 1, Lemma 6, and Corollary 3; [2] only consider the case of a trivial one-element \( X \), but the extension to the case of general \( X \) is trivial.) For each computable mixable \( \lambda \) (including the log and Brier loss functions), fix such a smallest upper semicomputable loss superprocess; it will be denoted \( K_\lambda \), and \( K_\lambda(\sigma) \) will be called the predictive complexity of \( \sigma \in \mathbb{Z}^* \) under \( \lambda \). The intuition behind \( K_\lambda(\sigma) \) is that this is the loss of the ideal prediction strategy whose computation is allowed to take an infinite amount of time.

In this note we consider infinite data sequences \( \zeta \in \mathbb{Z}^\infty \), which are idealizations of long finite data sequences. If \( \zeta = (z_1, z_2, \ldots) \in \mathbb{Z}^\infty \) and \( T \) is a nonnegative integer, we let \( \zeta^T \) to stand for the prefix \( z_1 \ldots z_T \) of \( \zeta \) of length \( T \).

The randomness deficiency of \( \sigma \in \mathbb{Z}^* \) with respect to a prediction algorithm \( F \) under a computable mixable loss function \( \lambda \) is defined to be

\[
D^\lambda_F(\sigma) := \text{Loss}^\lambda_F(\sigma) - K^\lambda(\sigma); \tag{3}
\]

since \( \text{Loss}^\lambda_F \) is upper semicomputable ([2], Section 3.1), the function \( D^\lambda_F : \mathbb{Z}^* \to \mathbb{R} \) is bounded below. Notice that the indeterminacy \( \infty - \infty \) never arises in (3) as \( K^\lambda < \infty \). We will sometimes replace the upper index \( \lambda \) in any of the three terms of (3) by “log” (respectively “Brier”) in the case where \( \lambda \) is the log loss function (respectively the Brier loss function).

Let us say that \( \zeta \in \mathbb{Z}^\infty \) is random with respect to \( F \) under \( \lambda \) if

\[
\sup_T D^\lambda_F(\zeta^T) < \infty.
\]

The intuition is that \( F \) is an optimal prediction algorithm for \( \zeta \) under \( \lambda \).

### 4 Statements

In this section, \( \lambda \) is a computable proper mixable loss function. A special case of the following theorem is stated as Proposition 16 in [4].

**Theorem 1.** If a data sequence \( \zeta \in \mathbb{Z}^\infty \) is random under the log loss function with respect to a prediction algorithm \( F \), it is random under \( \lambda \) with respect to \( F \).

The proof, given in the next section, will in fact demonstrate the following quantitative form of Theorem 1 for any \( \lambda \) there exists \( c_\lambda \in (0, \infty) \) such that, for any prediction algorithm \( F \),

\[
c_\lambda \exp(D^\text{log}_F) \geq D^\lambda_F. \tag{4}
\]
Theorem 1 asserts a fundamental character of the log loss function. The following simple statement shows that the Brier loss function does not share this property.

**Theorem 2.** There is a data sequence \( \zeta \in \mathbb{Z}^\infty \) and a prediction algorithm \( F \) such that, with respect to \( F \), \( \zeta \) is random under the Brier loss function but not under the log loss function.

The proof of Theorem 2 given in the next section will show that the randomness deficiency \( D_{log}^F(\zeta^T) \) of \( \zeta^T \) with respect to \( F \) under the log loss function can even grow almost as fast as \( T^{1/2} \) as \( T \to \infty \), while \( \zeta \) is random with respect to \( F \) under the Brier loss function.

**5 Proofs**

Let us say that a function \( f : \mathbb{Z}^* \to \mathbb{R} \) is a supermartingale with respect to a prediction algorithm \( F \) if, for all \( \sigma \in \mathbb{Z}^* \) and \( x \in X \),

\[
f(\sigma) \geq F(\sigma, x)f(\sigma, x, 1) + (1 - F(\sigma, x))f(\sigma, x, 0).
\]

The proof of Theorem 1 is based on the following characterization of positive supermartingales in terms of the log loss function.

**Lemma 1.** A positive function \( f : \mathbb{Z}^* \to (0, \infty) \) is a supermartingale if and only if \( \log f \) can be represented in the form

\[
\log f = \text{Loss}_{log}^F - L,
\]

where \( L \) is a finite loss superprocess under the log loss function.

**Proof.** The statement of the lemma can be checked by a direct calculation; as \( f \) is strictly positive and finite, we need not worry too much about infinities. A function \( L : \mathbb{Z}^* \to \mathbb{R} \) is a loss superprocess under the log loss function if and only if for all \( \sigma \in \mathbb{Z}^* \) and \( x \in X \) there exists \( p \in [0, 1] \) such that

\[
\begin{align*}
L(\sigma, x, 1) - L(\sigma) &\geq -\log p \\
L(\sigma, x, 0) - L(\sigma) &\geq -\log(1 - p).
\end{align*}
\]

Therefore, a function \( \log f : \mathbb{Z}^* \to \mathbb{R} \) can be represented in the form (5) if and only if for all \( \sigma \in \mathbb{Z}^* \) and \( x \in X \) there exists \( p \in [0, 1] \) (and even \( p \in (0, 1) \)) such that

\[
\begin{align*}
-\log F(\sigma, x) + \log f(\sigma) - \log f(\sigma, x, 1) &\geq -\log p \\
-\log(1 - F(\sigma, x)) + \log f(\sigma) - \log f(\sigma, x, 0) &\geq -\log(1 - p),
\end{align*}
\]

i.e.,

\[
\begin{align*}
F(\sigma, x) \frac{f(\sigma, x, 1)}{f(\sigma)} &\leq p \\
(1 - F(\sigma, x)) \frac{f(\sigma, x, 0)}{f(\sigma)} &\leq 1 - p.
\end{align*}
\]
Therefore, a function $\log f : \mathbb{Z}^* \to \mathbb{R}$ can be represented in the form (5) if and only if, for all $\sigma \in \mathbb{Z}^*$ and $x \in X$,

$$F(\sigma, x)\frac{f(\sigma, x, 1)}{f(\sigma)} + (1 - F(\sigma, x))\frac{f(\sigma, x, 0)}{f(\sigma)} \leq 1.$$  

This is equivalent to $f$ being a supermartingale with respect to $F$.  

Proof of Theorem 1

Let us check that the randomness deficiency process (3) is a supermartingale with respect to $F$; namely, we will check that

$$\mathbb{E}_{F(\sigma,x)} \left( \lambda(F(\sigma, x), \cdot) - \mathcal{K}^\lambda(\sigma, x, \cdot) + \mathcal{K}^\lambda(\sigma) \right) \leq 0$$

for all $\sigma \in \mathbb{Z}^*$ and $x \in X$. By the definition of a loss superprocess, it is sufficient to check

$$\mathbb{E}_{F(\sigma,x)} \lambda(F(\sigma, x), \cdot) \leq \mathbb{E}_{F(\sigma,x)} \lambda(p, \cdot)$$

for all $\sigma \in \mathbb{Z}^*$, $x \in X$, and $p \in [0, 1]$, which follows from $\lambda$ being a proper loss function.

We can see that, for some $c_1 \in (0, \infty)$, $D^\lambda_F + c_1$ is a positive supermartingale with respect to $F$; let us assume, for the moment, that this supermartingale does not take value $\infty$. By Lemma 1

$$\log(D^\lambda_F + c_1) = \text{Loss}_{F}^{\log} - \mathcal{L}$$

for some finite loss superprocess $L$ under the log loss function. This equality shows that $L$ is upper semicomputable; therefore,

$$\log(D^\lambda_F + c_1) \leq \text{Loss}_{F}^{\log} - \mathcal{K}^{\log} + c_2 = D^\log_F + c_2$$

for some $c_2 \in \mathbb{R}$, which is equivalent to (6).

It remains to consider the case where $D^\lambda_F + c_1$ takes value $\infty$. Assume, without loss of generality, $c_1 \in (1, \infty)$. Let us say that $\sigma \in \mathbb{Z}^\infty$ is unlucky for $F$ if there exists a prefix $(\sigma', x, y)$ of $\sigma$, where $\sigma' \in \mathbb{Z}^*$, $x \in X$, and $y \in \{0, 1\}$, such that $|F(\sigma', x) - y| = 1$; it is clear that $D^\lambda_F(\sigma) = \infty$ is possible only if $\sigma$ is unlucky for $F$. Redefine, until the end of this proof, the values $D^\lambda_F(\sigma) = \infty$ to $D^\lambda_F(\sigma) = 0$; this does not affect the supermartingale property of $D^\lambda_F$, and so the process $L$ defined by (6) remains a loss superprocess under the log loss function. Therefore, (7) holds for the redefined $D^\lambda_F$, and therefore, it holds for the original $D^\lambda_F$, as $D^\log_F(\sigma) = \infty$ when $\sigma$ is unlucky under $F$.

Proof of Theorem 2

We will partly follow the argument given after Proposition 16 of [4]. Consider any data sequence $\zeta = (x_1, y_1, x_2, y_2, \ldots) \in \mathbb{Z}^\infty$ in which all labels are 0: $y_1 = y_2 = \cdots = 0$. We then have $\sup_T \mathcal{K}^{\text{log}}(\zeta^T) < \infty$ and $\sup_T \mathcal{K}^{\text{brief}}(\zeta^T) < \infty$.
Let $F$ be the prediction algorithm that outputs $p_t := t^{-1/2-\epsilon}$ at step $t$, where $\epsilon \in (0, 1/2)$. Then $\zeta$ is random with respect to $F$ under the Brier loss function since the loss of this prediction algorithm over the first $T$ steps is $\sum_{t=1}^{T} p_t^2$ and the series $\sum_t p_t^2$ is convergent. On the other hand, the randomness deficiency of $\zeta^T$ with respect to $F$ under the log loss function grows as

$$-\sum_{t=1}^{T} \log(1 - p_t) \sim \sum_{t=1}^{T} p_t \sim \frac{2}{1 - 2\epsilon} T^{1/2-\epsilon}.$$ 

References

[1] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In Proceedings of the Twenty Third International Conference on Machine Learning, pages 161–168, New York, 2006. ACM.

[2] Yuri Kalnichkan. Predictive complexity for games with finite outcome spaces. In Vladimir Vovk, Harris Papadopoulos, and Alex Gammerman, editors, Measures of Complexity: Festschrift for Alexey Chervonenkis, pages 119–141. Springer, Berlin, 2015. Available on the web.

[3] Vladimir Vovk. Competitive on-line statistics. International Statistical Review, 69:213–248, 2001.

[4] Vladimir Vovk. Probability theory for the Brier game. Theoretical Computer Science, 261:57–79, 2001.