Loss functions and their risks

Scribes: Heng Zhao and Yaofeng Su

1 Introduction

In this section, we discuss the properties of loss functions and their risks following the material from [1, Ch.2]. To motivate this discussion, let us recall that the goal of supervised learning methods is to find a solution function f^* that (approximately) minimizes the risk $R_{L,P}(f)$

$$R^*_{L,P}(f) = \inf_{f:X\to\mathbb{R}} R_{L,P}(f). \tag{1}$$

In practice, the probability P is unknown, so we examine the empirical risk

$$R_{L,D} = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)), \tag{2}$$

where D is the set of training samples $D = \{(x_i, y_i) : i = 1, ..., n\}$.

In general, equation (1) may have non-unique solution and, in any case, computing the solution may be unfeasible. For example, the $0/1$ loss function used in binary classification is non-convex and, consequently, solving equation 1 is NP-hard as Hoffgen *et al.* (1995) showed.

Let us explain how SVMs make the optimization problem computationally feasible. The first step is to replace the 0/1 classification loss by a convex surrogate. The most common choice in this regard is the hinge loss, which is defined by

$$L_{hinge}(y,t) := max\{0, 1-yt\},\ y \in \{-1,+1\},\ t \in \mathbb{R}$$

To show that hinge loss is a convex surrogate of the $0/1$ classification, we make some observations below.

Explanation. Let us consider the classical SVM setting where input data $x_1, \ldots x_n \in X$ are mapped into a possibly infinite dimensional Hilbert space H_0 by a feature map

$\Phi : X \to H_0$. Thea soft-margin SVM problem requires to solve the following constrained minimization problem:

$$\begin{array}{l} \min\limits_{w,b,\xi} \ \frac{1}{2}\langle w,w\rangle + C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \ \ y_i(\langle w,\Phi(x_i)\rangle + b) \geq 1-\xi_i, \end{array} \tag{3}$$

where $w \in H_0, b \in \mathbb{R}, \xi_i \geq 0, i = 1,...,n$. We can write the constraint inequality as

$$\xi_i \geq 1 - y_i(\langle w,\Phi(x_i)\rangle + b),\ \xi_i \geq 0$$

which is equivalent to

$$\xi_i = \begin{cases} 0, & if\ y_i(\langle w,\Phi(x_i)\rangle + b) > 1 \\ 1-y_i(\langle w,\Phi(x_i)\rangle + b) & if\ y_i(\langle w,\Phi(x_i)\rangle + b) \leq 1 \end{cases} \tag{4}$$

and this, in turn, is equivalent to

$$\xi_i = max\{0, 1-y_i(\langle w,\Phi(x_i)\rangle + b)\} \tag{5}$$

$$= L_{hinge}(y_i, f_{w,b}(x_i)) \tag{6}$$

where $f_{w,b}(\cdot) = \langle w,\Phi(\cdot)\rangle + b$. Thus, the soft-margin SVM problem can be state as:

$$\inf_{f\in H} \lambda \|f\|_H^2 + R_{L,D}(f)$$

where

$$\|f\|_H = \inf\left\{\|w\|_{H_0} : w \in H_0,\ f_{w,b}(\cdot) = \langle w,\Phi(\cdot)\rangle + b\right\}$$

$$R_{L,D}(f) = \frac{1}{N}\sum_{i=1}^{N} L_{hinge}(y_i, f_{w,b}(x_i)).$$

Here the regularization term $\lambda \|f\|_H^2$ penalizes solution functions with a large RKHS norm.

Remark. Due to the convexity of the hinge loss function, the minimization problem is convex hence it has unique solution.

In the next section, we analyze in details the basic properties of a loss function.

2 Loss Functions and Risks

2.1 Loss Functions

Definition 1. Let $(X,\mathbb{R})$ be a measurable space and $Y \subset \mathbb{R}$ be a closed subset. Then a function $f : X \times Y \times \mathbb{R} \to [0,\infty)$ is called a *loss function* if it is a measurable function.

In the following, we will interpret $L(x,y,f(x))$ as the *cost* or *loss* of predicting the value y using $f(x)$ when x is observed, i.e., the smaller the value $L(x,y,f(x))$ is, the better $f(x)$ predicts y with respect to the loss L. From this, it is clear that constant loss functions, such as $L := 0$, are rather meaningless for our purposes, since they do not distinguish between good and bad predictions.

Let us now recall that our major goal is to have a small *average* loss for future unseen observations (x,y). This leads to the following definition.

Definition 2. Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss function and P be a probability measure on $X \times Y$. Then, for a measurable function $f : X \to \mathbb{R}$, the *L-risk* is defined by

$$R_{L,P} = \int_{X\times Y} L(x,y,f(x))dP(x,y) = \int_X \int_Y L(x,y,f(x))\, dP(y|x)\, dP_X(x),$$

where $P(y|x)$ is a conditional probability.

Let $D = \{(x_i,y_i) : i = 1,\dots n\}$ be *i.i.d.* points in $X \times Y$. The empirical risk is

$$R_{L,D} = \frac{1}{n}\sum_{i=1}^{n} L(x_i,y_i,f(x_i)).$$

By the law of large number, $R_{L,D}(f)$ gets closer to $R_{L,P}(f)$ as n is large. In this sense, the empirical risk can be seen as an approximation of L-risk of f (for f a fixed funstion).

Now recall that $L(x,y,f(x))$ was interpreted as a cost that we wish to keep small and, hence, it is natural to look for functions f whose risks are as small as possible. Since the smallest possible risk plays an important role, we give the follwoing definition.

Definition 3. Let $L : X \times Y \times \mathbb{R} \to [0,\infty)$ be a loss function and P be a probability measure on $X \times Y$. Then the minimal L-risk

$$R^*_{L,P} := \inf\{R_{L,P}(f)|f : X \to \mathbb{R}\ \textit{measurable}\}$$

is called the *Bayes risk* with respect to P and L. In addition, a measurable $f^*_{L,P} : X \to \mathbb{R}$ with $R_{L,P}(f^*_{L,P}) = R^*_{L,P}$ is called a *Bayes decision function*.

Example 1. (Standard binary classification). Let $Y := \{-1,1\}$ and P be an unknown data-generating distribution on $X \times Y$. Then the goal of binary classification is to predict y for a pair $(x,y) \in X \times Y$ when x is observed. The most common loss function describing this learning goal is the *classification loss* $L_{class} : Y \times \mathbb{R} \to [0,\infty)$, which is defined by

$$L_{class}(y,t) := I_{(-\infty,0]}(y\ \mathrm{sgn}(t)), \quad y \in Y, t \in \mathbb{R}$$

or, equivalently,

$$L_{class} : Y \times \mathbb{R} \to [0,\infty) := \begin{cases} 0, \textit{ if } y = \mathrm{sgn}(t) \\ 1, \textit{ if } y \neq \mathrm{sgn}(t) \end{cases} \tag{7}$$

Note that L_{class} only penalizes predictions t whose signs disagree with that of y, so it indeed reflects our informal learning goal. Now, for a measurable function $f : X \to \mathbb{R}$, an elementary calculation shows

$$R_{L_{class},P}(f) = \int_X \int_Y L_{class}(y, f(x))\, dP(y|x)\, dP_X(x) \tag{8}$$

$$= \int_X \eta(x) 1_{(-\infty,0)}(f(x)) + (1-\eta(x)) 1_{(0,\infty)}(f(x))\, \mathrm{d}P_X(x) \tag{9}$$

$$= P(\{(x,y) \in X \times Y : \mathrm{sign} f(x) \neq y\} \tag{10}$$

where $\eta(x) := P(y = 1|x), x \in X$. From this we conclude that f is a Bayes decision function if and only if $(2\eta(x) - 1)\,\mathrm{sgn} f(x) \geq 0$ for P_X-almost all $x \in X$. In addition, this consideration yields

$$R^*_{L_{class},P} = \int_X \min\{\eta, 1-\eta\}\, \mathrm{d}P_X$$

Example 2. (Weighted binary classification). Let $Y := \{-1, 1\}$ and $\alpha \in (0,1)$. Then the α-weighted classification loss $L_{\alpha-class} : Y \times \mathbb{R} \to [0,\infty)$ is defined by

$$L_{\alpha-class}(y,t) := \begin{cases} 1-\alpha & \text{if } y = 1 \text{ and } t < 0 \\ \alpha & \text{if } y = -1 \text{ and } t \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

for all $y \in Y, t \in \mathbb{R}$. Obviously we have $2L_{\frac{1}{2}-class} = L_{class}$, i.e., the standard binary classification scenario is a special case of the general weighted classification scenario. Given a probability measure P on $X \times Y$ and a measurable $f : X \to \mathbb{R}$, the $L_{\alpha-class}$-risk can be computed by

$$R_{L_{\alpha-class},P}(f) = (1-\alpha) \int_{f<0} \eta\, \mathrm{d}P_X + \alpha \int_{f\geq 0} (1-\eta)\, \mathrm{d}P_X$$

where $\eta(x) := P(y = 1|x)$, $x \in X$. From this we easily conclude that f is a Bayes decision function if and only if $(\eta(x) - \alpha)\,\mathrm{sgn} f(x) \geq 0$ for P_X-almost all $x \in X$. Finally, the Bayes $L_{\alpha-class}$-risk is

$$R^*_{L_{\alpha-class},P} = \int_X \min\{(1-\alpha)\eta, \alpha(1-\eta)\}\, \mathrm{d}P_X$$

In the two examples above the goal was to predict labels y from the set $\{-1, 1\}$. In the next example, we wish to predict general real-valued labels.

Example 3. (Least squares regression). The informal goal in regression is to predict predict the label $y \in Y = \mathbb{R}$ of a pair (x,y) drawn from an unknown probability measure P on $X \times Y$ if only x is observed. The most common way to formalize this goal is based on the *least square loss* $L_{LS} : Y \times \mathbb{R} \to [0,\infty)$ defined by

$$L_{LS}(y,t) := (y-t)^2, \quad y \in Y,\ t \in \mathbb{R}. \tag{12}$$

In other words, the least squares loss penalizes the discrepancy between y and t *quadratically*. Obviously, for a measurable function $f : X \to R$, the L_{LS}-risk is

$$R_{L_{LS},P}(f) = \int_X \int_Y (y - f(x))^2 \, \mathrm{d}P(y|x) \, \mathrm{d}P_X(x).$$

By minimizing the inner integral with respect to $f(x)$, we then see that f is a Bayes decision function if and only if $f(x)$ almost surely equals the expected Y-value in x, i.e., if and only if

$$f(x) = \mathbb{E}_P(Y|x) := \int_Y y \, \mathrm{d}P(y|x) \tag{13}$$

for P_X-almost all $x \in X$. Moreover, plugging $x \mapsto \mathbb{E}_P(Y|x)$ into $R_{L_{LS},P}(\cdot)$ shows that the Bayes L_{LS}-risk is the average conditional Y-variance, i.e.,

$$R^*_{L_{LS},P} = \int_X \mathbb{E}_P(Y^2|x) - (\mathbb{E}_P(Y|x))^2 \, \mathrm{d}P_X(x).$$

In all examples above we assumed that $L(x,y,f(x)) = L(y,f(x))$, with no dependence on X directly. This setting is part of a more general situation.

Definition 4. A function $L : Y \times \mathbb{R} \to [0,\infty)$ is called a *supervised loss function* if it is measurable. A function $L : X \times \mathbb{R} \to [0,\infty)$ is called an *unsupervised loss function* if it is measurable.

In case L is an unsupervised loss function, the risk has the form

$$R_{L,P}(f) := \int_X L(x, f(x)) \, \mathrm{d}P_X(x)$$

and it is independent of the supervisor $P(\cdot|x)$ that generates the labels.

Example 4. (Density distribution)

Let μ be a known probability measure on X, $g : X \to [0,\infty)$ be an unknown density w.r.t. μ. The goal is to estimate g. In this case, a possible choice is the unsupervised loss $L_q : X \times \mathbb{R} \to [0,\infty), q > 0$ given by

$$L_q(X,t) = |g(x) - t|^q.$$

Let $P_X = \mu$, then

$$R_{L_q,p}(f) = \int_X |g(x) - f(x)|^q d\mu(x)$$

where f is any measurable function on X. Clearly, $R^*_{L_q,p} = 0$ if $f^* = g$ modulo sets of μ-measure zero.

2.2 Properties of loss functions and their risks

Definition 5. A loss $L : X \times Y \times \mathbb{R} \to [0,\infty)$ is strictly convex if $L(x,y,\cdot) : \mathbb{R} \to [0,\infty)$ is strictly convex for all $x \in X, y \in Y$. L is continuous if $L(x,y,\cdot) : \mathbb{R} \to [0,\infty)$ is continuous for all $x \in X, y \in Y$.

If L is a strictly convex loss then it is easy to see that also the L-risk is strictly convex. However, the continuity of the loss does not imply the continuity of the corresponding risk.

Proposition 1 (semi-continuous of Risk). *If $L : X \times Y \times \mathbb{R} \to [0,\infty)$ is a strictly convex loss, P is a distribution on $X \times Y$ and $(f_n)_{n\geq 1}$ is a sequence of measurable function on X converging to a measurable function f in probability w.r.t. the marginal distribution P_X, then*

$$R_{L,p}(f) \leq \liminf_{n\to\infty} R_{L,p}(f_n)$$

.

Before proving this proposition, we recall some definitions. Given a probability space (X,σ,P), a sequence $(f_n)_{n\geq 1}$ of measurable functions $f_n : X \to \mathbb{R}$ converge to a measurable function $f : X \to \mathbb{R}$ in probability if for any $\varepsilon > 0, \delta > 0$, there is N s.t. for any $n \geq N$,

$$P(\{x \in X : |f_n(x) - f(x)| \geq \varepsilon\}) < \delta.$$

We say that (f_n) converges to f P-almost surely if

$$f_n(x) \to f(x) \text{ for } P\text{-almost all } x \in X.$$

We have the following properties:

- P-almost sure convergence $\implies$ converge in probability.
- Convergence in probability $\implies$ there is $(f_{n_k})_{k\geq 1} \subset (f_n)_{n\geq 1}$ that converges $\mathbb{P}$-almost surely.

We can now prove the proposition.

Proof. Since $(f_n)_{n\geq 1}$ converges in probability to f, then there exists a subsequence $(f_{n_k})_{k\geq 1} \subset (f_n)_{n\geq 1}$ that converges to f almost P_X-almost surely. By the continuity of L, we have that

$$\lim_{k\to\infty} L(x,y,f_{n_k}(x)) = L(x,y,f(x)) \text{ for } P_X\text{-almost all } (x,y) \in X \times Y.$$

By Fatou's Lemma,

$$R_{L,P}(f) = \int_{X\times Y} \lim_{k\to\infty} L(x,y,f_{n_k}(x)) dP(x,y)$$

$$\leq \liminf_{k\to\infty} L(x,y,f_{n_k}(x)) dP(x,y) = \liminf_{n\to\infty} R_{L,p}(f_n).$$

Remark 1. It is clear that if there is an integrable majoranof t $L(\cdot,\cdot,f_n(\cdot))$, then we can use dominated convergence theorem to show that $R_{L,P}(f_n) \to R_{L,P}(f)$.

Definition 6. A loss $L : X \times Y \times \mathbb{R} \to (0,\infty)$ is a Nemitski loss if there is a measurable $b : X \times Y \to [0,\infty)$ and an increasing $h : [0,\infty) \to [0,\infty)$ s.t. $L(x,y,t) \leq b(x,y) + h(|t|)$ for all $(x,y,t) \in X \times Y \times \mathbb{R}$.

Definition 7. L is a Nemitski loss of order $p \in (0,\infty)$ if there is $c > 0$ s.t.

$$L(x,y,t) \leq b(x,y) + c|t|^p \text{ for all } (x,y,t) \in X \times Y \times \mathbb{P}.$$

If, in addition, P is a distribution on $X \times Y$ and b is P integrable, then L is a P-integrable Nemitski loss.

Remark 2. If L is a P-integrable Nemitski loss and $f \in L^\infty(P_X)$, then $R_{L,P}(F) < \infty$ and $R^*_{L,P} < \infty$ by continuity of the risk.

Proposition 2. *Let P be a distribution on $X \times Y$ and L a continuous P-integrable Nemitski loss.*

1. *Let $(f_n)_{n\geq 1}$ be uniformly bounded measurable functions from $X \to \mathbb{R}$ such that $||f_n||_{L^\infty} \leq B$, where $B > 0$ is independent of n. If $f_n \to f$ P_X-almost surely, then*

$$\lim_{n\to\infty} R_{L,P}(f_n) = R_{L,P}(f).$$

2. *The map $R_{L,P} : L^\infty(P_X) \to [0,\infty)$ is well-defined and continuous.*
3. *If L is of order $p \in [1,\infty)$, then $R_{L,P} : L^p(P_X) \to [0,\infty)$ is well-defined and continuous.*

Proof. 1. It is clear that $||f||_{L^\infty} \leq B$. By the continuity of L,

$$\lim_{n\to\infty} L(x,y,f_n(x)) = L(x,y,f(x)) \text{ } P\text{-almost surely for all} (x,y) \in X \times Y.$$

Also,

$$|L(x,y,f_n(x)) - L(x,y,f(x))|$$
$$\leq 2b(x,y) + h(f_n(x)) + h(|f(x)|) \leq 2b(x,y) + 2h(B).$$

Since the RHS is P-integrable, by dominated convergence theorem,

$$|R_{L,P}(f_n) - R_{L,p}(f)| \leq \int |L(x,y,f_n(x)) - L(x,y,f(x))| dP(x,y)$$

which implies

$$R_{L,P}(f_n) \to R_{L,P}(f).$$

2. The Nemitski loss assumption and the integrability of b imply that $R_{L,P}(f)$ is bounded for any $f \in L^\infty(P_X)$. The continuity follows from part 1.
3. The hypothesis on the loss function directly gives that

$$R_{L,P}(F) < \infty, \quad \text{if } f \in L^p(P_X).$$

For the continuity, let $(f_n)_{n\geq 1} \subset L^p(P_X)$ with $f_n \to f$ in L^p. Since L^p convergence implies convergence in probability, so by Proposition 1 we have that

$$R_{L,P}(f) \leq \liminf(R_{L,P}(f_n)).$$

Set $\widetilde{L}(x,y,t) := b(x,y) + c|t|^p - L(x,y,t)$. This is also a continuous loss. Thus

$$||b||_{L^1} + c||f||_{L^p}^p - R_{L,P}(f) = R_{\widetilde{L},P}(f) \leq \liminf_{n\to\infty} R_{\widetilde{L},P}(f_n)$$

$$= \liminf_{n\to\infty}(-R_{L,P}(f_n) + ||b||_{L^1} + ||f_n||_{L^p}^p).$$

Using the continuity of L^p-norm, we conclude

$$\limsup_{n\to\infty} R_{L,P}(f_n) \leq R_{L,P}(f) \text{ when } f_n \to f \text{ in } L^p\text{norm}.$$

References

1. Ingo Steinwart; Andreas Christmann. *Support Vector Machines,* Springer, 2008.