6 Consistency

Apart from the uniform convergence of the empirical error rates of decision rules to the true risks, another fundamental question in classification is whether it is possible to construct a sequence of rules $\{f_n\}$ such that their theoretical risk, $R(f_n)$, gets arbitrarily close to the lowest possible risk, R^*, with a large probability. This brings the classical notion of consistency for classification just as in statistical estimation problems where an analogous question is posed for estimation of model parameters. This section reviews various concepts of consistency related to classification and examines some consistent classification procedures.

Definition 5 (Consistency)**.** A classification rule f_n is said to be consistent for a certain distribution $\mathcal{P}$ of (X, Y) if

$$E_{\mathcal{D}_n} R(f_n) \to R^* \quad \text{as } n \to \infty,$$

and *strongly* consistent if

$$R(f_n) \to R^* \quad \text{almost surely.}$$

Note that the definition of consistency of f_n depends on the distribution $\mathcal{P}$ of (X, Y). In general, a decision rule can be consistent for a certain family of distributions of (X, Y) but may not be consistent for others. Since the distribution of (X, Y) is unknown in practice, it is desirable to have a rule that is consistent for a large family of distributions. This consideration leads to much stronger notion of consistency as defined below.

Definition 6 (Universal Consistency)**.** A sequence of classification rules is called *universally* consistent if it is consistent for any distribution of (X, Y).

6.1 Fisher Consistency

Different from the definitions of consistency mentioned above, there is another useful idea of consistency relevant to the procedures defined through empirical risk minimization. It originates from a classical parameter estimation setting where a model parameter, say, θ is estimated with a random sample of size n. Suppose that an estimator of θ is defined as a functional of the empirical distribution F_n, $T(F_n)$. The estimator (or the estimation procedure) is said to be *Fisher consistent* if its population analog, $T(F)$, is the same as the parameter θ.

Adapting the idea to the procedures of empirical risk minimization for classification, consider a procedure of finding $f \in \mathcal{F}$ that minimizes an empirical risk with respect to L, $(1/n)\sum_{i=1}^{n} L(f(x_i), y_i)$. We say that a loss function L is Fisher consistent (or classification-calibrated) if the population minimizer of the risk $EL(f(X), Y)$ for all measurable functions leads to the Bayes optimal decision rule. Fisher consistency is regarded as a very minimal condition for the loss function L to be proper for approximation of the optimal decision rule. Recall that the Bayes decision rule is given by $f^*(x) = I(\eta(x) \geq 1/2)$ with labels 0 and 1, and equivalently $f^*(x) = sign(\eta(x) - 1/2)$ with labels 1 and -1. Direct minimization of the empirical error rate is computationally difficult due to the non-convexity of the 0-1 loss function. Instead, many classification procedures attempt to minimize convex surrogate loss functions for computational ease or due to other motivations. Under the symmetric labeling of 1 and -1, various procedures can be compared via the corresponding loss functions based

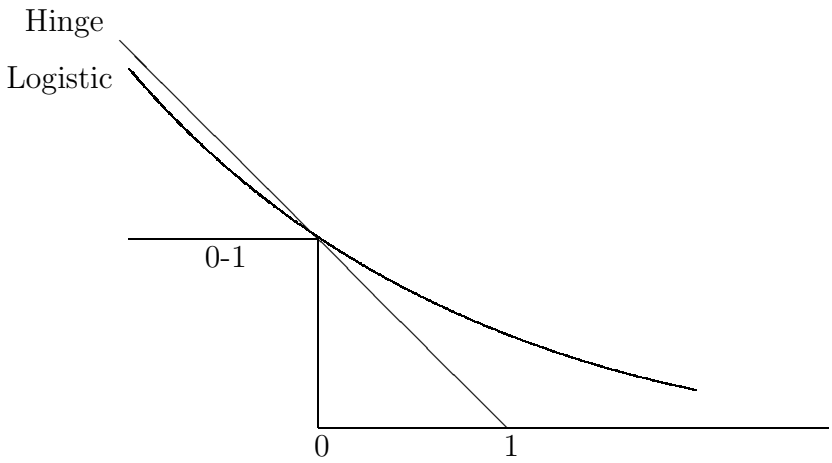

Figure 7: Margin-based loss functions

on the so-called *functional margin*, $yf(x)$. Examples of such loss functions for margin-based classifiers are the hinge loss for the SVM and the negative log-likelihood for logistic regression.

For a real-valued function f, which induces a decision rule via $sign(f)$,

(a) *Misclassification (0-1) loss*: $L(f(x), y) = I(yf(x) \leq 0)$

(b) *Hinge loss* for the SVM: $L(f(x), y) = (1 - yf(x))_+$

(c) *Negative log-likelihood* for logistic regression: $L(f(x), y) = \log_2\{1 + \exp(-yf(x))\}$.
Let $\eta(x) = P(Y = 1|X = x)$ and $f(x) = \log \eta(x)/(1 - \eta(x))$, the logit function. Then

$$\eta(x) = \frac{e^{f(x)}}{1 + e^{f(x)}} \text{ and } 1 - \eta(x) = \frac{1}{1 + e^{f(x)}}.$$

The likelihood of f given (x, y) is

$$l(f(x), y) = \begin{cases} \eta(x) & \text{for } y = 1 \\ 1 - \eta(x) & \text{for } y = -1. \end{cases}$$

Taking the negative log-likelihood as a new loss function, we get

$$\begin{aligned} L(f(x), y) &= -\log_2 l(f(x), y) = \log_2 \frac{1}{l(f(x), y)} \\ &= \begin{cases} \log_2(1 + e^{-f(x)}) & \text{for } y = 1 \\ \log_2(1 + e^{f(x)}) & \text{for } y = -1 \end{cases} \\ &= \log_2\{1 + \exp(-yf(x))\}. \end{aligned}$$

Figure 7 compares the margin-based loss functions with the 0-1 loss. They are convex upper bounds of the misclassification loss. To verify whether each loss function is Fisher consistent, we find the population minimizer over all measurable functions.

(a) *SVM*:

The population minimizer is given by

$$f_{\text{SVM}}(x) := \underset{\text{f:measurable}}{\arg\min}\ E(1 - Yf(X))_+.$$

For each x, we seek the minimizer $f(x)$ of

$$E\Big((1 - Yf(X))_+ | X = x\Big) = (1 - f(x))_+\eta(x) + (1 + f(x))_+(1 - \eta(x)).$$

Then, the value $f(x)$ should be in $[-1, 1]$ because otherwise truncation of f at -1 or 1 gives a lower loss. Thus, when $-1 \le f(x) \le 1$,

$$\begin{aligned} E\Big((1 - Yf(X))_+ | X = x\Big) &= (1 - f(x))\eta(x) + (1 + f(x))(1 - \eta(x)) \\ &= 1 + (1 - 2\eta(x))f(x). \end{aligned}$$

Therefore,

$$f_{\text{SVM}}(x) = \begin{cases} 1 & \text{if } \eta(x) \ge 1/2 \\ -1 & \text{if } \eta(x) < 1/2 \end{cases} = \text{sign}(\eta(x) - 1/2) = f^*(x).$$

The population minimizer of the hinge loss itself coincides with the Bayes rule. Obviously, $sign(f_{\text{SVM}}(x)) = f^*(x)$. So, the hinge loss is Fisher consistent. In addition, the population minimizer shows that the SVM aims at the Bayes decision rule directly without estimation of $\eta(x)$.

(b) *Logistic regression*:

Similarly, it can be shown that the population minimizer of the negative log-likelihood is the true logit function itself, $f_{LR}(x) = \log \frac{\eta(x)}{1-\eta(x)}$. Hence,

$$\text{sign}(f_{LR}(x)) = \text{sign}(\eta(x) - 1/2) = f^*(x),$$

and the negative log-likelihood is classification-calibrated.

Casting the problem in a general setting, consider the procedures that find a classification rule by convex risk minimization. That is, the classification rule is determined by minimizing the empirical risk $(1/n)\sum_{i=1}^{n} L(y_i f(x_i))$ with respect to a margin-based convex loss function L. What conditions does L need for Fisher consistency? To state such conditions for L, first define $R_L(f) := EL(Yf(X))$ and note that

$$E\big(L(Yf(X))|X = x\big) = \eta(x)L(f(x)) + (1 - \eta(x))L(-f(x)).$$

For specification of the optimal value $f(x)$ at each x, let $C_\eta(\alpha) := \eta L(\alpha) + (1-\eta)L(-\alpha)$ for $\alpha \in \mathbb{R}$ and α_η^* be the minimizer of $C_\eta(\alpha)$. The following theorem provides sufficient conditions for Fisher consistency of convex risk minimization with L.

Theorem 16. *For a convex L, if L is differentiable at 0 and $L'(0) < 0$, then L is classification-calibrated.*

Proof. For brevity, denote $\eta(x)$ at x by η. Then we need to show that

$$sign(\alpha_\eta^*) = sign(\eta - 1/2).$$

From $C'_\eta(\alpha) = \eta L'(\alpha) - (1-\eta)L'(-\alpha)$, we have $C'_\eta(0) = (2\eta - 1)L'(0)$. If $\eta > 1/2$, then $C'_\eta(0) < 0$. Then there exists, say, $\alpha_0 > 0$ such that

$$C_\eta(\alpha_0) \le C_\eta(0) + C'_\eta(0)\frac{\alpha_0}{2}.$$

By the convexity of L and hence of $C_\eta(\alpha)$, for all α

$$C_\eta(\alpha) \ge C_\eta(0) + C'_\eta(0)\alpha.$$

In particular, for $\alpha < \alpha_0/4$

$$C_\eta(\alpha) \ge C_\eta(0) + C'_\eta(0)\frac{\alpha_0}{4} > C_\eta(0) + C'_\eta(0)\frac{\alpha_0}{2} \ge C_\eta(\alpha_0).$$

Therefore, the minimizer $\alpha_\eta^* > 0$. Similarly, for $\eta < 1/2$, $\alpha_\eta^* < 0$. Hence L is classification-calibrated. □

As shown in Figure 7, the hinge loss and the negative log likelihood apparently satisfy the conditions in the theorem. Furthermore, it implies that a variety of convex loss functions can be used for construction of classification rules as a surrogate of the misclassification loss. Other examples include

i) $L(yf(x)) = (1 - yf(x))^2 = (y - f(x))^2$ (quadratic loss),

ii) $L(yf(x)) = (1 - yf(x))_+^2$ (truncated quadratic loss or squared hinge loss),

iii) $L(yf(x)) = \exp(-yf(x))$ (exponential loss for boosting).

For more discussions about convex risk minimization, see *'Convexity, classification, and risk bounds'*, JASA (2006) by Bartlett, Jordan, and McAuliffe.