Binary classification, Part 1

Maxim Raginsky

September 22, 2015

The problem of binary classification can be stated as follows. We have a random couple $Z = (X, Y)$, where $X \in \mathbb{R}^d$ is called the *feature vector* and $Y \in \{-1, 1\}$ is called the *label*[1]. In the spirit of the model-free framework, we assume that the relationship between the features and the labels is stochastic and described by an unknown probability distribution $P \in \mathscr{P}(\mathsf{Z})$, where $\mathsf{Z} = \mathbb{R}^d \times \{-1, 1\}$. In these lectures on binary classification, I will be following mainly two excellent sources: the book by Devroye, Györfi, and Lugosi [DGL96] and the comprehensive survey article by Bousquet, Boucheron, and Lugosi [BBL05].

As usual, we consider the case when we are given an i.i.d. sample of length n from P. The goal is to learn a *classifier*, i.e., a mapping $g : \mathbb{R}^d \to \{-1, 1\}$, such that the probability of classification error, $\mathbb{P}(g(X) \neq Y)$, is small. As we have seen before, the optimal choice is the *Bayes classifier*

$$g^*(x) \triangleq \begin{cases} 1, & \text{if } \eta(x) > 1/2 \\ -1, & \text{otherwise} \end{cases} \tag{1}$$

where $\eta(x) \triangleq \mathbb{P}[Y = 1 | X = x]$ is the *regression function*. However, since we make no assumptions on P, in general we cannot hope to learn the Bayes classifier g^*. Instead, we focus on a more realistic goal: We fix a collection $\mathscr{G}$ of classifiers and then use the training data to come up with a hypothesis $\widehat{g}_n \in \mathscr{G}$, such that

$$\mathbb{P}(\widehat{g}_n(X) \neq Y) \approx \inf_{g \in \mathscr{G}} \mathbb{P}(g(X) \neq Y)$$

with high probability.

By way of notation, let us write $L(g)$ for the classification error of g, i.e., $L(g) \triangleq \mathbb{P}(g(X) \neq Y)$, and let $L^*(\mathscr{G})$ denote the smallest classification error attainable over $\mathscr{G}$:

$$L^*(\mathscr{G}) \triangleq \inf_{g \in \mathscr{G}} L(g).$$

We will assume that a minimizing $g^* \in \mathscr{G}$ exists. For future reference, we note that

$$L(g) = \mathbb{P}(g(X) \neq Y) = \mathbb{P}(Y g(X) < 0). \tag{2}$$

Warning: In what follows, we will use C or c to denote various absolute constants; their values may change from line to line.

[1] The reason why we chose $\{-1, 1\}$, rather than $\{0, 1\}$, for the label space is merely convenience.

1 Learning linear discriminant rules

One of the simplest classification rules (and one of the first to be studied) is a *linear discriminant rule*: given a nonzero vector $w = (w^{(1)}, \ldots, w^{(d)}) \in \mathbb{R}^d$ and a scalar $b \in \mathbb{R}$, let

$$g(x) \equiv g_{w,b}(x) \triangleq \begin{cases} 1, & \text{if } \langle w, x\rangle + b > 0 \\ -1, & \text{otherwise} \end{cases} \tag{3}$$

Let $\mathscr{G}$ be the class of all such linear discriminant rules as w ranges over all nonzero vectors in $\mathbb{R}^d$ and b ranges over all reals: $\mathscr{G} = \{g_{w,b} : w \in \mathbb{R}^d \setminus \{0\}, b \in \mathbb{R}\}$.

Given the training sample Z^n, let $\widehat{g}_n \in \mathscr{G}$ be the output of the ERM algorithm, i.e.,

$$\widehat{g}_n \triangleq \arg\min_{g \in \mathscr{G}} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{g(X_i) \neq Y_i\}}.$$

In other words, $\widehat{g}_n$ is any classifier of the form (3) that minimizes the number of misclassifications on the training sample. Then we have the following:

Theorem 1. *There exists an absolute constant $C > 0$, such that for any $n \in \mathbb{N}$ and any $\delta \in (0,1)$, the bound*

$$L(\widehat{g}_n) \leq L^*(\mathscr{G}) + C\sqrt{\frac{d+1}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}} \tag{4}$$

holds with probability at least $1 - \delta$.

Proof. A standard argument leads to the bound

$$L(\widehat{g}_n) \leq L^*(\mathscr{G}) + 2\Delta_n(Z^n), \tag{5}$$

where

$$\Delta_n(Z^n) \triangleq \sup_{g \in \mathscr{G}} \left| L(g) - L_n(g) \right|$$

is the uniform deviation and $L_n(g)$ denotes the *empirical classification error* of g on Z^n:

$$L_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{g(X_i) \neq Y_i\}},$$

which is the fraction of incorrectly labeled points in the training sample Z^n. Consider a classifier $g \in \mathscr{G}$ and define the set

$$C_g \triangleq \left\{ (x, y) \in \mathbb{R}^d \times \{-1, 1\} : y \cdot (\langle w, x\rangle + b) \leq 0 \right\}.$$

Then it is easy to see that

$$L(g) = P(C_g) \qquad \text{and} \qquad L_n(g) = P_n(C_g),$$

where, as before,

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{Z_i} = \frac{1}{n} \sum_{i=1}^{n} \delta_{(X_i, Y_i)}$$

is the empirical distribution of the sample Z^n. Let $\mathscr{C}$ denote the collection of all sets of the form $C = C_g$ for some $g \in \mathscr{G}$. Then

$$\Delta_n(Z^n) = \sup_{C \in \mathscr{C}} |P_n(C) - P(C)|.$$

Let $\mathscr{F} = \mathscr{F}_{\mathscr{C}}$ denote the class of indicator functions of the sets in $\mathscr{C}$: $\mathscr{F}_{\mathscr{C}} = \{\mathbf{1}_{\{\cdot \in C\}} : C \in \mathscr{C}\}$. Then we know that, with probability at least $1-\delta$,

$$\Delta_n(Z^n) \le 2\mathbb{E}R_n(\mathscr{F}(Z^n)) + \sqrt{\frac{\log(1/\delta)}{2n}}, \tag{6}$$

where $R_n(\mathscr{F}(Z^n))$ is the Rademacher average of the projection of $\mathscr{F}$ onto the sample Z^n. Now,

$$\begin{aligned}\mathscr{F}(Z^n) &= \big\{(f(Z_1),\ldots,f(Z_n)) : f \in \mathscr{F}\big\} \\ &= \big\{(\mathbf{1}_{\{Z_1\in C\}},\ldots,\mathbf{1}_{\{Z_n\in C\}}) : C \in \mathscr{C}\big\}.\end{aligned}$$

Therefore, if we prove that $\mathscr{C}$ is a VC class, then

$$R_n(\mathscr{F}(Z^n)) \le C\sqrt{\frac{V(\mathscr{C})}{n}}.$$

But this follows from the fact that any $C \in \mathscr{C}$ has the form

$$C = \left\{(x,y) \in \mathbb{R}^d \times \{-1,1\} : \sum_{j=1}^{d} w^{(j)} y x^{(j)} + by \le 0\right\}$$

for some $w \in \mathbb{R}^d \setminus \{0\}$ and some $b \in \mathbb{R}$, and the functions $(x,y) \mapsto y$ and $(x,y) \mapsto yx^{(j)}$, $1 \le j \le d$, span a linear space of dimension no greater than $d+1$. Hence, $V(\mathscr{C}) \le d+1$, so that

$$R_n(\mathscr{F}(Z^n)) \le C\sqrt{\frac{V(\mathscr{C})}{n}} \le C\sqrt{\frac{d+1}{n}}.$$

Combining this with (5) and (6), we see that (4) holds with probability at least $1-\delta$. □

1.1 Generalized linear discriminant rules

In the same vein, we may consider classification rules of the form

$$g(x) = \begin{cases} 1, & \text{if } \sum_{j=1}^{k} w^{(j)}\psi_j(x) + b > 0 \\ -1, & \text{otherwise} \end{cases} \tag{7}$$

where k is some positive integer (not necessarily equal to d), $w = (w^{(1)},\ldots,w^{(k)}) \in \mathbb{R}^k$ is a nonzero vector, $b \in \mathbb{R}$ is an arbitrary scalar, and $\Psi = \{\psi_j : \mathbb{R}^d \to \mathbb{R}\}_{j=1}^k$ is some fixed "dictionary" of real-valued functions on $\mathbb{R}^d$. For a fixed Ψ, let $\mathscr{G}$ denote the collection of all classifiers of the form (7) as w ranges over all nonzero vectors in $\mathbb{R}^k$ and b ranges over all reals. Then the ERM rule is, again, given by

$$\widehat{g}_n \triangleq \inf_{g\in\mathscr{G}} L_n(g) \equiv \inf_{g\in\mathscr{G}} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{g(X_i)\neq Y_i\}}.$$

The following result can be proved pretty much along the same lines as Theorem 1:

Theorem 2. *There exists an absolute constant $C > 0$, such that for any $n \in \mathbb{N}$ and any $\delta \in (0,1)$, the bound*

$$L(\widehat{g}_n) \le L^*(\mathscr{G}) + C\sqrt{\frac{k+1}{n}} + \sqrt{\frac{2\log(1/\delta)}{n}} \tag{8}$$

holds with probability at least $1-\delta$.

1.2 Two fundamental issues

As Theorems 1 and 2 show, the ERM algorithm applied to the collection of all (generalized) linear discriminant rules is guaranteed to work well in the sense that the classification error of the output hypothesis will, with high probability, be close to the optimum achievable by any discriminant rule with the given structure. The same argument extends to any collection of classifiers $\mathscr{G}$, for which the "error sets" $\{(x,y) : y \cdot g(x) \le 0\}$, $g \in \mathscr{G}$, form a VC class of dimension much smaller than the sample size n. In other words, with high probability the difference

$$L(\widehat{g}_n) - L^*(\mathscr{G}) = L(\widehat{g}_n) - \inf_{g \in \mathscr{G}} L(g)$$

will be small. However, precisely because the VC dimension of $\mathscr{G}$ cannot be too large, the approximation properties of $\mathscr{G}$ will be limited. Another problem is computational. For instance, the problem of finding an empirically optimal linear discriminant rule is NP-hard. In other words, unless P is equal to NP, there is no hope of coming up with an efficient ERM algorithm for linear discriminant rules that would work for all feature space dimensions d. If d is fixed, then it is possible to enumerate all projections of a given sample Z^n onto the class of indicators of all halfspaces in $O(n^{d-1}\log n)$ time, which allows for an exhaustive search for an ERM solution, but the usefulness of this naive approach is limited to $d < 5$.

2 Risk bounds for combined classifiers via surrogate loss functions

One way to sidestep the above approximation-theoretic and computational issues is to replace the 0–1 Hamming loss function that gives rise to the probability of error criterion with some other loss function. What we gain is the ability to bound the performance of various complicated classifiers built up by combining simpler *base classifiers* in terms of the complexity (e.g, the VC dimension) of the collection of the base classifiers, as well as considerable computational advantages, especially if the problem of minimizing the empirical surrogate loss turns out to be a convex programming problem. What we lose, though, is that, in general, we will not be able to compare the generalization error of the learned classifier to the minimum classification risk. Instead, we will have to be content with the fact that the generalization error will be close to the smallest *surrogate loss*.

We will consider classifiers of the form

$$g_f(x) = \operatorname{sgn} f(x) \equiv \begin{cases} 1, & \text{if } f(x) \ge 0 \\ -1, & \text{otherwise} \end{cases} \tag{9}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is some function. From (2) we have

$$L(g_f) = \mathbb{P}(g_f(X) \ne Y) \le \mathbb{P}(Y g_f(X) < 0) = \mathbb{P}(Y f(X) < 0).$$

From now on, when dealing with classifiers of the form (9), we will write $L(f)$ instead of $L(g_f)$ to keep the notation simple. Now we introduce the notion of a surrogate loss function.

Definition 1. *A* surrogate loss function *is any function* $\varphi : \mathbb{R} \to \mathbb{R}_+$, *such that*

$$\varphi(x) \ge \mathbf{1}_{\{x>0\}}. \tag{10}$$

Some examples of commonly used surrogate losses:

1. Exponential, $\varphi(x) = e^x$
2. Logit, $\varphi(x) = \log_2(1 + e^x)$
3. Hinge loss, $\varphi(x) = (x+1)_+ \equiv \max\{x+1, 0\}$

Let φ be a surrogate loss. Then for any $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$ and any $f : \mathbb{R}^d \to \mathbb{R}$ we have

$$yf(x) < 0 \qquad \Rightarrow \qquad \varphi(-yf(x)) \geq \mathbf{1}_{\{-yf(x)>0\}} = \mathbf{1}_{\{yf(x)<0\}}. \tag{11}$$

Therefore, defining the *φ-risk* of f by

$$A_\varphi(f) \triangleq \mathbb{E}[\varphi(-Yf(X))]$$

and its empirical version

$$A_{\varphi,n}(f) \triangleq \frac{1}{n}\sum_{i=1}^{n} \varphi(-Y_i f(X_i)),$$

we see from (11) that

$$L(f) \leq A_\varphi(f) \qquad \text{and} \qquad L_n(f) \leq A_{\varphi,n}(f). \tag{12}$$

Now that these preliminaries are out of the way, we can state and prove the basic surrogate loss bound:

Theorem 3. *Consider any learning algorithm $\mathscr{A} = \{\mathscr{A}_n\}_{n=1}^\infty$, where, for each n, the mapping $\mathscr{A}_n$ receives the training sample $Z^n = (Z_1, \ldots, Z_n)$ as input and produces a function $\widehat{f}_n : \mathbb{R}^d \to \mathbb{R}$ from some class $\mathscr{F}$. Suppose that $\mathscr{F}$ and the surrogate loss φ are chosen so that the following conditions are satisfied:*

1. *There exists some constant $B > 0$ such that*

$$\sup_{(x,y)\in\mathbb{R}^d\times\{-1,1\}} \sup_{f\in\mathscr{F}} \varphi(-yf(x)) \leq B$$

2. *There exists some constant $M_\varphi > 0$ such that φ is M_φ-Lipschitz, i.e.,*

$$|\varphi(u) - \varphi(v)| \leq M_\varphi |u - v|, \qquad \forall u, v \in \mathbb{R}.$$

Then for any n and any $\delta \in (0, 1)$ the following bound holds with probability at least $1 - \delta$:

$$L(\widehat{f}_n) \leq A_{\varphi,n}(\widehat{f}_n) + 4M_\varphi \mathbb{E}R_n(\mathscr{F}(X^n)) + B\sqrt{\frac{\log(1/\delta)}{2n}}. \tag{13}$$

Proof. Using (12), we can write

$$\begin{aligned} L(\widehat{f}_n) &\leq A_\varphi(\widehat{f}_n) \\ &= A_{\varphi,n}(\widehat{f}_n) + A_\varphi(\widehat{f}_n) - A_{\varphi,n}(\widehat{f}_n) \\ &\leq A_{\varphi,n}(\widehat{f}_n) + \sup_{f\in\mathscr{F}} \left| A_\varphi(f) - A_{\varphi,n}(f) \right|. \end{aligned}$$

Now let $\mathscr{H}$ denote the class of functions $h : \mathbb{R}^d \times \{-1,1\} \to \mathbb{R}$ of the form $h(x,y) = -yf(x)$, $f \in \mathscr{F}$. Then

$$\begin{aligned}\sup_{f\in\mathscr{F}} \left|A_\varphi(f) - A_{\varphi,n}(f)\right| &= \sup_{f\in\mathscr{F}} \left|\mathbb{E}[\varphi(-Yf(X))] - \frac{1}{n}\sum_{i=1}^n \varphi(-Y_i f(X_i))\right| \\ &= \sup_{h\in\mathscr{H}} \left|P(\varphi\circ h) - P_n(\varphi\circ h)\right|,\end{aligned}$$

where $\varphi\circ h(z) \triangleq \varphi(h(z))$ for every $z = (x,y) \in \mathbb{R}^d \times \{-1,1\}$. Let

$$\begin{aligned}\Delta_n(Z^n) &\triangleq \sup_{h\in\mathscr{H}} \left|P(\varphi\circ h) - P_n(\varphi\circ h)\right| \\ &= \sup_{h\in\mathscr{H}} \left|P(\varphi\circ h - \varphi(0)) - P_n(\varphi\circ h - \varphi(0))\right|,\end{aligned}$$

where the second line follows from the fact that adding the same constant to each $\varphi\circ h$ does not change the value of $P_n(\varphi\circ h) - P(\varphi\circ h)$. Using the familiar symmetrization argument, we can write

$$\mathbb{E}\Delta_n(Z^n) \le 2\mathbb{E}R_n\left(\mathscr{H}_\varphi(Z^n)\right), \tag{14}$$

where $\mathscr{H}_\varphi$ denotes the class of all functions of the form $(x,y) \mapsto \varphi(h(x,y)) - \varphi(0)$, $h \in \mathscr{H}$. We now use a very powerful result about the Rademacher averages called the *contraction principle*, which states the following [LT91]: If $\mathscr{A} \subset \mathbb{R}^n$ is a bounded set and $F : \mathbb{R} \to \mathbb{R}$ is an M-Lipschitz function satisfying $F(0) = 0$, then

$$R_n(F\circ\mathscr{A}) \le 2MR_n(\mathscr{A}), \tag{15}$$

where $F\circ\mathscr{A} \triangleq \{(F(a_1),\ldots,F(a_n)) : a = (a_1,\ldots,a_n) \in \mathscr{A}\}$. (The proof of the contraction principle is somewhat involved, and we do not give it here.) Consider the function $F(u) = \varphi(u) - \varphi(0)$. This function clearly satisfies $F(0) = 0$, and it is M_φ-Lipschitz, by our assumptions on φ. Moreover, from our definition of $\mathscr{H}_\varphi$, we immediately see that

$$\begin{aligned}\mathscr{H}_\varphi(Z^n) &= \left\{\left(\varphi(h(Z_1) - \varphi(0),\ldots,\varphi(h(Z_n) - \varphi(0)\right) : h \in \mathscr{H}\right\} \\ &= \{(F(h(Z_1)),\ldots,F(h(Z_n))) : h \in \mathscr{H}\} \\ &= F\circ\mathscr{H}(Z^n).\end{aligned}$$

Therefore, applying (15) to $\mathscr{A} = \mathscr{H}(Z^n)$ and then using the resulting bound in (14), we obtain

$$\mathbb{E}\Delta_n(Z^n) \le 4M_\varphi\mathbb{E}R_n\left(\mathscr{H}(Z^n)\right).$$

Furthermore, letting σ^n be an i.i.d. Rademacher tuple independent of Z^n, we have

$$\begin{aligned}R_n(\mathscr{H}(Z^n)) &= \frac{1}{n}\mathbb{E}_{\sigma^n}\left[\sup_{h\in\mathscr{H}}\left|\sum_{i=1}^n \sigma_i h(Z_i)\right|\right] \\ &= \frac{1}{n}\mathbb{E}_{\sigma^n}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i Y_i f(X_i)\right|\right] \\ &= \frac{1}{n}\mathbb{E}_{\sigma^n}\left[\sup_{f\in\mathscr{F}}\left|\sum_{i=1}^n \sigma_i f(X_i)\right|\right] \\ &\equiv R_n(\mathscr{F}(X^n)),\end{aligned}$$

which leads to

$$\mathbb{E}\Delta_n(Z^n) \le 4M_\varphi \mathbb{E}R_n\left(\mathscr{F}(X^n)\right). \tag{16}$$

Now, since every function $\varphi \circ h$ is bounded between 0 and B, the function $\Delta_n(Z^n)$ has bounded differences with $c_1 = \ldots = c_n = B/n$. Therefore, from (16) and from McDiarmid's inequality, we have for every $t > 0$ that

$$\mathbb{P}\Big(\Delta_n(Z^n) \ge 4M_\varphi \mathbb{E}R_n(\mathscr{F}(X^n)) + t\Big) \le \mathbb{P}\Big(\Delta_n(Z^n) \ge \mathbb{E}\Delta_n(Z^n) + t\Big) \le e^{-2nt^2/B^2}.$$

Choosing $t = B\sqrt{(2n)^{-1}\log(1/\delta)}$, we see that

$$\Delta_n(Z^n) \le 4M_\varphi \mathbb{E}R_n(\mathscr{F}(X^n)) + B\sqrt{\frac{\log(1/\delta)}{2n}}$$

with probability at least $1-\delta$. Therefore, since

$$L(\widehat{f}_n) \le A_{\varphi,n}(\widehat{f}_n) + \Delta_n(Z^n),$$

we see that (13) holds with probability at least $1-\delta$. □

What the above theorem tells us is that the performance of the learned classifier $\widehat{f}_n$ is controlled by the Rademacher average of the class $\mathscr{F}$, and we can always arrange it to be relatively small. We will now look at several specific examples.

3 Weighted linear combination of classifiers

Let $\mathscr{G} = \{g : \mathbb{R}^d \to \{-1,1\}\}$ be a class of *base classifiers* (not to be confused with *Bayes* classifiers!), and consider the class

$$\mathscr{F}_\lambda \triangleq \left\{ f = \sum_{j=1}^N c_j g_j : N \in \mathbb{N}, \sum_{j=1}^N |c_j| \le \lambda; g_1, \ldots, g_N \in \mathscr{G} \right\},$$

where $\lambda > 0$ is a tunable parameter. Then for each $f = \sum_{j=1}^N c_j g_j \in \mathscr{F}_\lambda$ the corresponding classifier g_f of the form (9) is given by

$$g_f(x) = \operatorname{sgn}\left(\sum_{j=1}^N c_j g_j(x) \right).$$

A useful way of thinking about g_f is that, upon receiving a feature $x \in \mathbb{R}^d$, it computes the outputs $g_1(x), \ldots, g_N(x)$ of the N base classifiers from $\mathscr{G}$ and then takes a weighted "majority vote" – indeed, if we had $c_1 = \ldots = c_N = \lambda/N$, then $\operatorname{sgn}(g_f(x))$ would precisely correspond to taking the majority vote among the N base classifiers. Note, by the way, that the number of base classifiers is not fixed, and can be learned from the data.

Now, Theorem 3 tells us that the performance of any learning algorithm that accepts a training sample Z^n and produces a function $\widehat{f}_n \in \mathscr{F}_\lambda$ is controlled by the Rademacher average $R_n(\mathscr{F}_\lambda(X^n))$. It turns out, moreover, that we can relate it to the Rademacher average of the base class $\mathscr{G}$. To start, note that

$$\mathscr{F}_\lambda = \lambda \cdot \operatorname{absconv}\mathscr{G},$$

where

$$\text{absconv}\mathscr{G} = \left\{ \sum_{j=1}^{N} c_j g_j : N \in \mathbb{N}; \sum_{j=1}^{N} c = |c_j| \leq 1; g_1, \ldots, g_N \in \mathscr{G} \right\}$$

is the absolute convex hull of $\mathscr{G}$. Therefore

$$R_n(\mathscr{F}_\lambda(X^n)) = \lambda \cdot R_n(\mathscr{G}(X^n)).$$

Now note that the functions in $\mathscr{G}$ are binary-valued. Therefore, assuming that the base class $\mathscr{G}$ is a VC class, we will have

$$R_n(\mathscr{G}(X^n)) \leq C\sqrt{\frac{V(\mathscr{G})}{n}}.$$

Combining these bounds with the bound of Theorem 3, we conclude that for any $\widehat{f}_n$ selected from $\mathscr{F}_\lambda$ based on the training sample Z^n, the bound

$$L(\widehat{f}_n) \leq A_{\varphi,n}(\widehat{f}_n) + C\lambda M_\varphi \sqrt{\frac{V(\mathscr{G})}{n}} + B\sqrt{\frac{\log(1/\delta)}{2n}}$$

will hold with probability at least $1-\delta$, where B is the uniform upper bound on $\varphi(-yf(x))$, $f \in \mathscr{F}_\Lambda$, $(x,y) \in \mathbb{R}^d \times \{-1,1\}$ and M_φ is the Lipschitz constant of the surrogate loss φ.

Note that the above bound involves only the VC dimension of the *base class*, which is typically small. On the other hand, the class $\mathscr{F}_\lambda$ obtained by forming weighted combinations of classifiers from $\mathscr{G}$ is extremely rich, and will generally have infinite VC dimension! But there is a price we pay: The first term is the empirical surrogate loss $A_{\varphi,n}(\widehat{f}_n)$, rather than the empirical classification error $L_n(\widehat{f}_n)$. However, it is possible to choose the surrogate φ in such a way that $A_{\varphi,n}(\cdot)$ can be bounded in terms of a quantity *related* to the number of misclassified training examples. Here is an example.

Fix a positive parameter $\gamma > 0$ and consider

$$\varphi(x) = \begin{cases} 0, & \text{if } x \leq -\gamma \\ 1, & \text{if } x \geq 0 \\ 1 + x/\gamma, & \text{otherwise} \end{cases}$$

This is a valid surrogate loss with $B = 1$ and $M_\varphi = 1/\gamma$, but in addition we have $\varphi(x) \leq \mathbf{1}_{\{x > -\gamma\}}$, which implies that $\varphi(-yf(x)) \leq \mathbf{1}_{\{yf(x) < \gamma\}}$. Therefore, for any f we have

$$A_{\varphi,n}(f) = \frac{1}{n}\sum_{i=1}^{n} \varphi(-Y_i f(X_i)) \leq \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{Y_i f(X_i) < \gamma\}}. \tag{17}$$

The quantity

$$L_n^\gamma(f) \triangleq \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{Y_i f(X_i) < \gamma\}} \tag{18}$$

is called the *margin error* of f. Notice that:

- For any $\gamma > 0$, $L_n^\gamma(f) \geq L_n(f)$
- The function $\gamma \mapsto L_n^\gamma(f)$ is increasing.

Notice also that we can write

$$L_n^\gamma(f) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{Y_i f(X_i)<0\}} + \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{0\le Y_i f(X_i)<\gamma\}},$$

where the first term is just $L_n(f)$, while the second term is the number of training examples that were classified correctly, but only with small "margin" (the quantity $Yf(X)$ is often called the *margin* of the classifier f).

Theorem 4 (Margin-based risk bound for weighted linear combinations)**.** *For any $\gamma > 0$, the bound*

$$L(\widehat{f}_n) \le L_n^\gamma(\widehat{f}_n) + \frac{C\lambda}{\gamma}\sqrt{\frac{V(\mathscr{G})}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \tag{19}$$

holds with probability at least $1-\delta$.

Remark 1. Note that the first term on the right-hand side of (19) increases with γ, while the second term decreases with γ. Hence, if the learned classifier $\widehat{f}_n$ has a small margin error for a large γ, i.e., it classifies the training samples well and with high "confidence," then its generalization error will be small.

References

[BBL05] O. Bousquet, S. Boucheron, and G. Lugosi. Theory of classification: a survey of recent advances. *ESAIM Probability and Statistics*, 9:323–375, 2005.

[DGL96] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[LT91] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 1991.