APPENDIX B

VAPNIK–CHERVONENKIS THEORY

Vapnik–Chervonenkis (VC) Theory (Devroye et al., 1996; Vapnik, 1998) introduces intuitively satisfying metrics for classification complexity, via quantities known as the shatter coefficients and the VC dimension, and uses them to uniformly bound the difference between apparent and true errors over a family of classifiers, in a distribution-free manner. The VC theorem also leads to a simple bound on the expected classifier design error for certain classification rules, which obey the empirical risk minimization principle.

The main result of VC theory is the VC theorem, which is related to the Glivenko–Cantelli theorem and the theory of empirical processes. All bounds in VC theory are worst-case, as there are no distributional assumptions, and thus they can be very loose for a particular feature–label distribution and small sample sizes. Nevertheless, the VC theorem remains a powerful tool for the analysis of the large sample behavior of both the true and apparent (i.e., resubstitution) classification errors.

B.1 SHATTER COEFFICIENTS

Intuitively, the complexity of a classification rule must have to do with its ability to "pick out" subsets of a given set of points. For a given n, consider a set of points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in R^d. Given a set $A \subseteq R^d$, then

$$A \cap \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subseteq \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \tag{B.1}$$

Error Estimation for Pattern Recognition, First Edition. Ulisses M. Braga-Neto and Edward R. Dougherty.

is the subset of $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ "picked out" by A. Now, consider a family $\mathcal{A}$ of measurable subsets of R^d, and let

$$N_{\mathcal{A}}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \left|\{A \cap \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \mid A \in \mathcal{A}\}\right|, \tag{B.2}$$

that is, the total number of subsets of $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ that can be picked out by sets in $\mathcal{A}$. The nth *shatter coefficient* of the family $\mathcal{A}$ is defined as

$$s(\mathcal{A}, n) = \max_{\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}} N_{\mathcal{A}}(\mathbf{x}_1, \ldots, \mathbf{x}_n). \tag{B.3}$$

The shatter coefficients $s(\mathcal{A}, n)$ measure the richness (the size, the complexity) of $\mathcal{A}$. Note also that $s(\mathcal{A}, n) \leq 2^n$ for all n.

B.2 THE VC DIMENSION

The Vapnik–Chervonenkis (VC) dimension is a measure of the size, that is, the complexity, of a class of classifiers $\mathcal{C}$. It agrees very naturally with our intuition of complexity as the ability of a classifier to cut up the space finely.

From the previous definition of shatter coefficient, we have that if $s(\mathcal{A}, n) = 2^n$, then there is a set of points $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ such that $N_{\mathcal{A}}(\mathbf{x}_1, \ldots, \mathbf{x}_n) = 2^n$, and we say that $\mathcal{A}$ *shatters* $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. On the other hand, if $s(\mathcal{A}, n) < 2^n$, then any set of points $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ contains at least one subset that cannot be picked out by any member of $\mathcal{A}$. In addition, we must have $s(\mathcal{A}, m) < 2^m$, for all $m > n$.

The VC dimension $V_{\mathcal{A}}$ of $\mathcal{A}$ (assuming $|\mathcal{A}| \geq 2$) is the largest integer $k \geq 1$ such that $s(\mathcal{A}, k) = 2^k$. If $s(\mathcal{A}, n) = 2^n$ for all n, then $V_{\mathcal{A}} = \infty$. The VC dimension of $\mathcal{A}$ is thus the *maximal number of points in R^d that can be shattered by $\mathcal{A}$*. Clearly, $V_{\mathcal{A}}$ measures the complexity of $\mathcal{A}$.

Some simple examples:

- Let $\mathcal{A}$ be the class of half-lines: $\mathcal{A} = \{(-\infty, a] \mid a \in R\}$, then

$$s(\mathcal{A}, n) = n + 1 \quad \text{and} \quad V_{\mathcal{A}} = 1. \tag{B.4}$$

- Let $\mathcal{A}$ be the class of intervals: $\mathcal{A} = \{[a, b] \mid a, b \in R\}$, then

$$s(\mathcal{A}, n) = \frac{n(n+1)}{2} + 1 \quad \text{and} \quad V_{\mathcal{A}} = 2. \tag{B.5}$$

- Let $\mathcal{A}_d$ be the class of "half-rectangles" in R^d: $\mathcal{A}_d = \{(-\infty, a_1] \times \cdots \times (-\infty, a_d] \mid (a_1, \ldots, a_d) \in R^d\}$, then $V_{\mathcal{A}_d} = d$.
- Let $\mathcal{A}_d$ be the class of rectangles in R^d: $\mathcal{A}_d = \{[a_1, b_1] \times \cdots \times [a_d, b_d] \mid (a_1, \ldots, a_d, b_1, \ldots, b_d]) \in R^{2d}\}$, then $V_{\mathcal{A}_d} = 2d$.

Note that in the examples above, the VC dimension is equal to the number of parameters. While this is intuitive, it is not true in general. In fact, one can find a one-parameter family $\mathcal{A}$ for which $V_{\mathcal{A}} = \infty$. So one should be careful about naively attributing complexity to the number of parameters.

Note also that the last two examples generalize the first two, by means of cartesian products. It can be shown that in such cases

$$s(\mathcal{A}_d, n) \leq s(\mathcal{A}, n)^d, \quad \text{for all } n. \tag{B.6}$$

A general bound for shatter coefficients is

$$s(\mathcal{A}, n) \leq \sum_{i=0}^{V_{\mathcal{A}}} \binom{n}{i}, \quad \text{for all } n. \tag{B.7}$$

The first two examples achieve this bound, so it is tight. From this bound it also follows, in the case that $V_{\mathcal{A}} < \infty$, that

$$s(\mathcal{A}, n) \leq (n+1)^{V_{\mathcal{A}}}. \tag{B.8}$$

B.3 VC THEORY OF CLASSIFICATION

The preceding concepts can be applied to define the complexity of a class $\mathcal{C}$ of classifiers (i.e., a classification rule). Given a classifier $\psi \in \mathcal{C}$, let us define the set $A_\psi = \{\mathbf{x} \in R^d \mid \psi(\mathbf{x}) = 1\}$, that is, the 1-decision region (this specifies the classifier completely, since the 0-decision region is simply A_ψ^c). Let $\mathcal{A}_\mathcal{C} = \{A_\psi \mid \psi \in \mathcal{C}\}$, that is, the family of all 1-decision regions produced by $\mathcal{C}$. We define the shatter coefficients $\mathcal{S}(\mathcal{C}, n)$ and VC dimension $V_\mathcal{C}$ for $\mathcal{C}$ as

$$\begin{aligned} \mathcal{S}(\mathcal{C}, n) &= s(\mathcal{A}_\mathcal{C}, n), \\ V_\mathcal{C} &= V_{\mathcal{A}_\mathcal{C}}. \end{aligned} \tag{B.9}$$

All the results discussed previously apply in the new setting; for example, following (B.8), if $V_\mathcal{C} < \infty$, then

$$\mathcal{S}(\mathcal{C}, n) \leq (n+1)^{V_\mathcal{C}}. \tag{B.10}$$

Next, results on the VC dimension and shatter coefficients for commonly used classification rules are given.

B.3.1 Linear Classification Rules

Linear classification rules are those that produce hyperplane decision-boundary classifiers. This includes NMC, LDA, Perceptrons, and Linear SVMs. Let $\mathcal{C}$ be the class

of hyperplane decision-boundary classifiers in R^d. Then, as proved in Devroye et al. (1996, Cor. 13.1),

$$S(\mathcal{C}, n) = 2 \sum_{i=0}^{d} \binom{n-1}{i},$$
$$V_{\mathcal{C}} = d + 1. \tag{B.11}$$

The fact that $V_{\mathcal{C}} = d + 1$ means that there is a set of $d + 1$ points that can be shattered by oriented hyperplanes in R^d, but no set of $d + 2$ points (in general position) can be shattered—a familiar fact in the case $d = 2$.

The VC dimension of linear classification rules thus increases linearly with the number of variables. Note however that the fact that all linear classification rules have the same VC dimension does not mean they will perform the same in small-sample cases.

B.3.2 *k*NN Classification Rule

For $k = 1$, clearly any set of points can be shattered (just use the points as training data). This is also true for any $k > 1$. Therefore, $V_{\mathcal{C}} = \infty$.

Classes $\mathcal{C}$ with finite VC dimension are called *VC classes*. Thus, the class $\mathcal{C}_k$ of *k*NN classifiers is not a VC class, for each $k > 1$. Classification rules that have infinite VC dimension are not necessarily useless. For example, there is empirical evidence that 3NN is a good rule in small-sample cases. In addition, the Cover-Hart Theorem says that the asymptotic *k*NN error rate is near the Bayes error. However, the worst-case scenario if $V_{\mathcal{C}} = \infty$ is indeed very bad, as shown in Section B.4.

B.3.3 Classification Trees

A binary tree with a depth of k levels of splitting nodes has at most $2^k - 1$ splitting nodes and at most 2^k leaves. Therefore, for a classification tree with data-independent splits (that is, fixed-partition tree classifiers), we have that

$$S(\mathcal{C}, n) = \begin{cases} 2^n, & n \le 2^k \\ 2^{2^k}, & n > 2^k \end{cases} \tag{B.12}$$

and it follows that $V_{\mathcal{C}} = 2^k$. The shatter coefficients and VC dimension thus grow very fast (exponentially) with the number of levels. The case is different for data-dependent decision trees (e.g., CART and BSP). If stopping or pruning criteria are not strict enough, one may have $V_{\mathcal{C}} = \infty$ in those cases.

B.3.4 Nonlinear SVMs

It is easy to see that the shatter coefficients and VC dimension correspond to those of linear classification in the transformed high-dimensional space. More precisely, if the *minimal* space where the kernel can be written as a dot product is m, then $V_C = m + 1$.

For example, for the polynomial kernel

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathbf{y})^p = (x_1 y_1 + \cdots + x_d y_d)^p, \tag{B.13}$$

we have $m = \binom{d+p-1}{p}$, that is, the number of distinct powers of $x_i y_i$ in the expansion of $K(\mathbf{x}, \mathbf{y})$, so $V_C = \binom{d+p-1}{p} + 1$.

For certain kernels, such as the Gaussian kernel,

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(\frac{-\|\mathbf{x} - \mathbf{y}\|^2}{\sigma^2}\right), \tag{B.14}$$

the minimal space is infinite dimensional, so $V_C = \infty$.

B.3.5 Neural Networks

A basic result for neural networks is proved in (Devroye et al., 1996, Theorem 30.5). For the class $\mathcal{C}_k$ of neural networks with k neurons in one hidden layer, and arbitrary sigmoids,

$$V_{\mathcal{C}_k} \geq 2 \left\lfloor \frac{k}{2} \right\rfloor d \tag{B.15}$$

where $\lfloor x \rfloor$ is the largest integer $\leq x$. If k is even, this simplifies to $V_C \geq kd$.

For threshold sigmoids, $V_{\mathcal{C}_k} < \infty$. In fact,

$$\mathcal{S}(\mathcal{C}_k, n) \leq (ne)^\gamma \text{ and } V_{\mathcal{C}_k} \leq 2\gamma \log_2(e\gamma), \tag{B.16}$$

where $\gamma = kd + 2k + 1$ is the number of weights. The threshold sigmoid achieves the smallest possible VC dimension among all sigmoids. In fact, there are sigmoids for which $V_{\mathcal{C}_k} = \infty$ for $k \geq 2$.

B.3.6 Histogram Rules

For a histogram rule with a finite number of partitions b, it is easy to see that the the shatter coefficients are given by

$$\mathcal{S}(\mathcal{C}, n) = \begin{cases} 2^n, & n < b, \\ 2^b, & n \geq b. \end{cases} \tag{B.17}$$

Therefore, the VC dimension is $V_C = b$.

B.4 VAPNIK–CHERVONENKIS THEOREM

The celebrated Vapnik–Chervonenkis theorem uses the shatter coefficients $S(\mathcal{C}, n)$ and the VC dimension $V_\mathcal{C}$ to bound

$$P(|\hat{\varepsilon}[\psi] - \varepsilon[\psi]| > \tau), \quad \text{for all } \tau > 0, \tag{B.18}$$

in a distribution-free manner, where $\varepsilon[\psi]$ is the true classification error of a classifier $\psi \in \mathcal{C}$ and $\hat{\varepsilon}_n[\psi]$ is the *empirical error* of ψ given data S_n:

$$\hat{\varepsilon}_n[\psi] = \frac{1}{n}\sum_{i=1}^{n} |y_i - \psi(x_i)|. \tag{B.19}$$

Note that if S_n could be assumed independent of every $\psi \in \mathcal{C}$, then $\hat{\varepsilon}_n[\psi]$ would be an independent test-set error, and we could use Hoeffding's Inequality (c.f. Section A.13) to get

$$P(|\hat{\varepsilon}[\psi] - \varepsilon[\psi]| > \tau) \leq 2e^{-2n\tau^2}, \quad \text{for all } \tau > 0. \tag{B.20}$$

However, that is not sufficient. If one wants to study $|\hat{\varepsilon}[\psi] - \varepsilon[\psi]|$ for any distribution, and any classifier $\psi \in \mathcal{C}$, in particular a designed classifier ψ_n, one cannot assume independence from S_n.

The solution is to bound $P(|\hat{\varepsilon}[\psi] - \varepsilon[\psi]| > \tau)$ uniformly for all possible $\psi \in \mathcal{C}$, that is, to find a (distribution-free) bound for the probability of the worst-case scenario

$$P\left(\sup_{\psi\in\mathcal{C}} |\hat{\varepsilon}[\psi] - \varepsilon[\psi]| > \tau\right), \quad \text{for all } \tau > 0. \tag{B.21}$$

This is the purpose of the Vapnik–Chervonenkis theorem, given next. This is a deep theorem, a proof of which can be found in Devroye et al. (1996).

Theorem B.1 *(Vapnik–Chervonenkis theorem.) Let $S(\mathcal{C}, n)$ be the nth shatter coefficient for class $\mathcal{C}$, as defined previously. Regardless of the distribution of (X, Y),*

$$P\left(\sup_{\psi\in\mathcal{C}} |\hat{\varepsilon}[\psi] - \varepsilon[\psi]| > \tau\right) \leq 8S(\mathcal{C}, n)e^{-n\tau^2/32}, \quad \textit{for all } \tau > 0. \tag{B.22}$$

If $V_\mathcal{C}$ is finite, we can use the inequality (B.10) to write the bound in terms of $V_\mathcal{C}$:

$$P\left(\sup_{\psi\in\mathcal{C}} |\hat{\varepsilon}[\psi] - \varepsilon[\psi]| > \tau\right) \leq 8(n+1)^{V_\mathcal{C}} e^{-n\tau^2/32}, \quad \text{for all } \tau > 0. \tag{B.23}$$

Therefore, if $V_\mathcal{C}$ is finite, the term $e^{-n\tau^2/32}$ dominates, and the bound decreases exponentially fast as $n \to \infty$.

If, on the other hand, $V_C = \infty$, we cannot make $n \gg V_C$, and a worst-case bound can be found that is independent of n (this means that there exists a situation where the design error cannot be reduced no matter how large n may be). More specifically, for every $\delta > 0$, and every classification rule associated with C, it can be shown (Devroye et al., 1996, Theorem 14.3) that there is a feature–label distribution for $(\mathbf{X}, Y)$ with $\varepsilon_C = 0$ but

$$E[\varepsilon_{n,C} - \varepsilon_C] = E[\varepsilon_{n,C}] > \frac{1}{2e} - \delta, \quad \text{for all } n > 1. \tag{B.24}$$