CS281B/Stat241B: Advanced Topics in Learning & Decision Making

Vapnik-Chervonenkis Dimension

Lecturer: Michael I. Jordan *Scribe: Songhwai Oh*

1 An Axis-Aligned Rectangle Classifier

Let's consider a classification problem using a set of axis-aligned rectangles. This will give you concrete understanding of the terms that have been used in the last couple of lectures on statistical learning theory. We make observations $x \in \mathbb{R}^2$ and want to find a rectangle that can best fit our observations according to the labels. The rectangle in Figure 1 classifies any points inside to be 1 and any points outside to be 0. Let's examine how each term can be applied in this example.

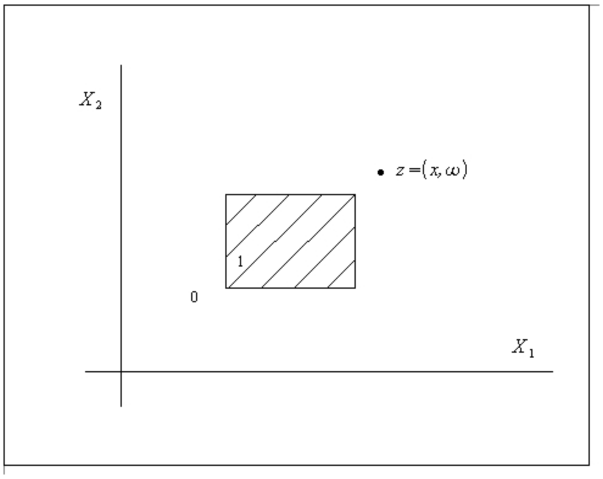

Figure 1: An Axis-Aligned Rectangle Classifier

- Λ is a set of all possible arragements of axis-aligned rectangles on this $\mathbb{R}^2$ plane. In general Λ is a set of abstract parameters (e.g., a set of functions).
- $\alpha \in \Lambda$ is a specific instance of the rectangles from the set Λ as shown in Figure 1.
- z describes a pair (x, ω) where $\omega \in \{0, 1\}$ is the label for the observation x.
- $g(x, \alpha) \in \{0, 1\}$ is our learning machine's function. It predicts the label associated with the observation. In this example, it labels every point inside the rectangle as 1 and any point outside as 0.
- $Q(z, \alpha)$ is a loss function and for this example it is

$$Q(z, \alpha) = \begin{cases} 0 & \text{if } g(x, \alpha) = \omega \\ 1 & \text{otherwise.} \end{cases}$$

- $R(\alpha)$ is a risk functional which is the expected value of the loss function $Q(z, \alpha)$. Note that the probability measure $F(z)$ is unknown and we cannot calculate this risk functional in practice.

$$R(\alpha) = EQ(z, \alpha) = \int Q(z, \alpha) dF(z), \qquad \alpha \in \Lambda \tag{1}$$

- $R_{emp}(\alpha)$ is the empirical risk functional and calculated from the training set as

$$R_{emp}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} Q(z_i, \alpha), \qquad \alpha \in \Lambda \tag{2}$$

  for a given set of i.i.d. observations $z_1, \ldots, z_l$. Observe that $E[R_{emp}(\alpha)] = R(\alpha)$.

- α_l is a particular rectangle that yields the minimum empirical risk functional. As we have seen from the previous lectures, one can obtain a bound for the risk functional in terms of the empirical risk and a "Chernoff bound"-like loss bound and the relationship can be summarized as

$$R(\alpha_l) \leq R_{emp}(\alpha_l) + O\left(\frac{1}{\sqrt{l}}\right). \tag{3}$$

- $N^\Lambda(z_1, \ldots, z_l)$ denotes the number of different separations of the given sample that can be achieved using functions from Λ. In our example it is the number of different labelings we can achieve using rectangles. The count N^Λ is smaller for the sample on the left of Figure 2 because you cannot have a rectangle that can only group the distant points while the sample on the right can be separated every possible way (Figure 3). Note that $N^\Lambda(z_1, \ldots, z_l) \leq 2^l$.

- The annealed VC entropy $H^\Lambda_{ann}(l)$ is defined as $\ln E[N^\Lambda(z_1, \ldots, z_l)]$. It is the logarithm of the expected value of the count $N^\Lambda(z_1, \ldots, z_l)$.

- However, the annealed VC entropy can't generally be computed in practice. Instead, we consider the growth function $G^\Lambda(l)$ which is the worst case entropy and $H^\Lambda_{ann}(l) \leq G^\Lambda(l)$. We show how to bound the growth function using VC dimension in the next sections.

$$G^\Lambda(l) \triangleq \ln \sup_{z_1, \ldots, z_l} N^\Lambda(z_1, \ldots, z_l)$$

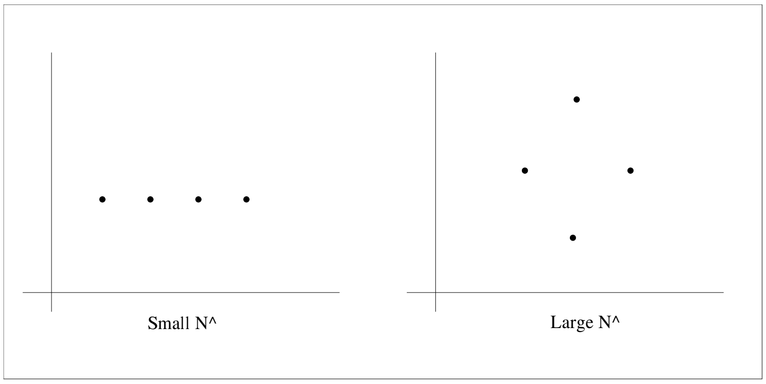

Figure 2: Different N^Λ's

2 VC Dimension

Lemma 1 (VC Lemma) *Any growth function is either linear or is bounded by a logarithmic function.*

$$G^{\Lambda}(l) = \begin{cases} = & \ln 2^l & \text{if} \quad l \leq h \\ \leq & \ln \sum_{i=0}^{h} \binom{l}{i} & \text{if} \quad l > h \end{cases}$$

Definition 1 (Shattering) *A family Λ shatters a set of points $z_1, \ldots, z_l$ if it picks out all possible 2^l labels on $z_1, \ldots, z_l$.*

The shattering of four points by a rectangle classifier is illustrated in Figure 3.

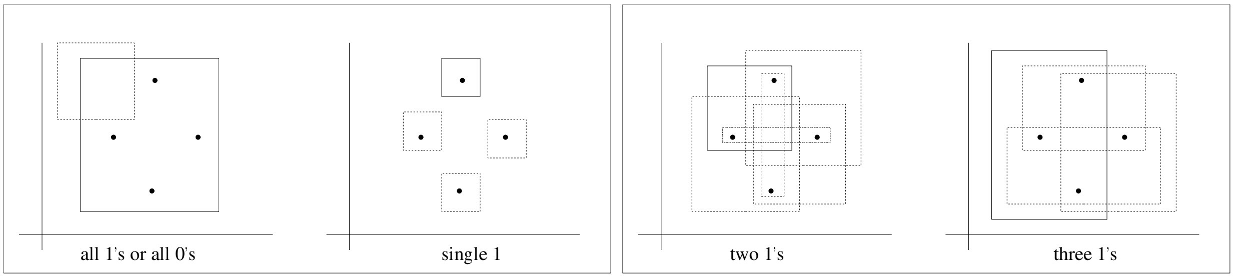

Figure 3: Shattering with Rectangles

Definition 2 (VC Dimension) *The VC dimension h is the cardinality of the maximal set that can be shattered.*

You can think of the VC dimension h as one number that characterizes the statistical properties of the family Λ. Note that if for any n there exists a set of n samples that can be shattered by the functions in Λ, then the VC dimension is equal to infinity.

3 Examples of VC Dimension

3.1 VC Dimension of Axis-Aligned Rectangles

Claim: We cannot shatter 5 or more points with rectangles.

Proof: For any set of 5 points, choose the 4 points that have the max and min in the first and the second coordinates (Figure 4). Then by the definition the fifth point must be inside a rectangle given by the max and min points. In this arrangement we cannot assign 1 on the edges and 0 inside. Thus, this proves that we cannot shatter 5 or more points with rectangles. ■

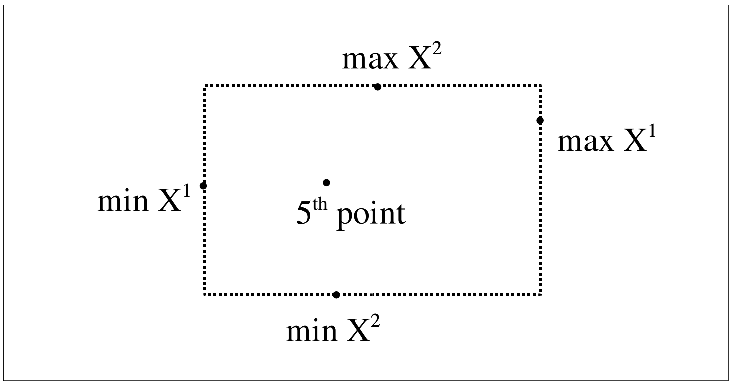

Figure 4: 5 Points in a Rectangle

3.2 VC Dimension of hyperplanes

A hyperplane that goes through the origin can be denoted by

$$\sum_{k=1}^{n} a^k u_k = 0.$$

Then the loss function $Q(\cdot)$ can be defined as

$$Q(z, \alpha) = \theta\Big(\sum_{k=1}^{n} a^k u_k\Big) = \theta\Big(\mathbf{a}^T \mathbf{u}\Big).$$

First, we claim that the VC dimension h of half-planes is greater than or equal to the dimension of the data n. Then we show that h is exactly n by showing that $n+1$ vectors cannot be shattered.

Claim: $VC_{dim} = n$.

Proof: Consider the n unit basis vectors in $\mathbb{R}^n$

$$(1, 0, \ldots, 0), (0, 1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1).$$

It is easy to see that these vectors can be shattered by hyperplanes.

Let's show that there are no $n+1$ vectors in $\mathbb{R}^n$ that can be shattered by hyperplanes passing through the origin. We prove this by contradiction. Suppose that $u_1, \ldots, u_{n+1}$ can be shattered. This implies that there exist 2^{n+1} vectors $a_i \in \mathbb{R}^n, i = 1, \ldots, 2^{n+1}$ such that the matrix of inner products denoted by $z_{i,j} = u_i^T a_j$ has columns with all possible combination of signs. So we have the matrix of inner products

$$A = \begin{pmatrix} z_{1,1} & \cdots & z_{1,2^{n+1}} \\ \vdots & \ldots & \vdots \\ z_{(n+1),1} & \cdots & z_{(n+1),2^{n+1}} \end{pmatrix}$$

and it has all 2^{n+1} possible combination of signs

$$sign(A) = \begin{pmatrix} - & - & \cdots & - & + \\ - & \cdot & \cdots & \cdot & + \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ - & + & \cdots & \cdot & + \end{pmatrix}.$$

Then the rows of A are linearly independent because there are no constants $c_1, c_2, \ldots, c_{n+1}$ such that $\sum_{i=1}^{n+1} c_i z_{i,(:)} = 0$ because for any $c_1, c_2, \ldots, c_{n+1}$, there is a column with the same signs which makes $\sum_{i=1}^{n+1} c_i z_{i,(:)}$ always non-zero. This implies that $n+1$ vectors in $\mathbb{R}^n$ are linearly independent but it is a false statement. This contradiction proves that there are no $n+1$ vectors in $\mathbb{R}^n$ that can be shattered by hyperplanes passing through the origin. Thus, the VC dimension of this class is n. ■

Note that the VC dimension of general hyperplanes that does not go through the origin is $n+1$ since one additional degree of freedom is added by the offset.

The same idea can be applied to a set of linearly independent functions. Consider

$$\theta\Big(\sum_{i=1}^{n} a^i \phi_i(z)\Big).$$

Now define $u_i \triangleq \phi_i(z)$. It follows the same argument as the hyperplane case and the resulting VC dimension of linearly independent functions is n. This shows that the VC dimension does not depend on the complexity of the function $\phi_i(\cdot)$.

3.3 VC Dimension of Closed Balls in $\mathbb{R}^d$

A set of closed balls in $\mathbb{R}^d$ is defined as

$$\Big\{ \quad \mathbf{z} = (z^1, \cdots, z^d) \quad \Big| \quad \sum_{i=1}^{d} (z^i - a_i)^2 \leq b \quad \Big\}.$$

Claim: $VC_{dim} \leq d+2$.

Proof: After expanding the quadradic term, we get

$$\sum_i (z^i - a_i)^2 - b = \sum (z^i)^2 - 2\sum z^i a_i + \sum a_i^2 - b. \tag{4}$$

If we define appropriate basis functions as

$$\phi_1(z) = \sum (z^i)^2, \quad \phi_2(z) = z^1, \quad \ldots, \quad \phi_{d+1}(z) = z^d, \quad \text{and} \quad \phi_{d+2}(z) = 1,$$

then clearly the Eq. (4) is a linear combination of the basis functions. Since $\phi_i(\cdot)$ may not be linearly independent to each other, the VC dimension is upper bounded by the dimension of basis functions $d+2$. Thus, $VC_{dim} \leq d+2$. ■

3.4 An Example with $VC_{dim} = \infty$

Consider a set of convex polygons in $\mathbb{R}^2$. If you place points on a circle, you can shatter any combination of points (Figure 5). This means that the VC dimension of a set of polygons is infinity. (Note that this does not mean that you cannot learn using polygons. I.e., the converse of the VC dimension theory is not true in general.)

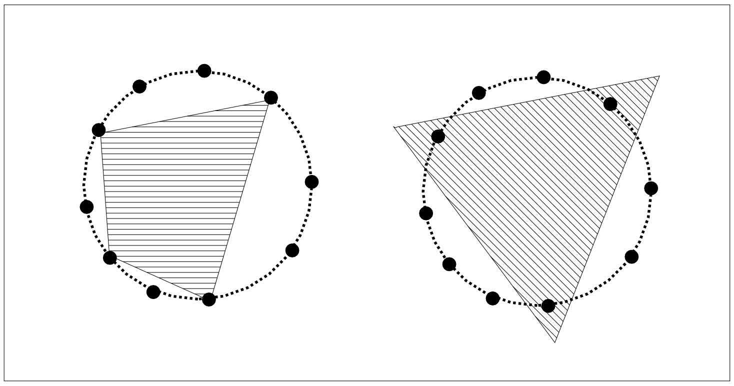

Figure 5: Convex Polygons

3.5 Finite Number of Parameters and Infinite VC Dimension

In this example, we show a set of single parameter functions whose VC dimension is infinity. Consider the one-dimensional indicator functions

$$Q(z, \alpha) = \theta(\sin \alpha z), \qquad z \in (0, 2\pi), \quad \alpha \in (0, \infty)$$

defined on the interval $(0, 2\pi)$. Let $\delta_1, \ldots, \delta_l$ be a binary sequence. Then there exist l points $z_1, \ldots, z_l$ such that the system of equation

$$\theta(\sin \alpha z_i) = \delta_i$$

has a solution in α. Let $z_i = 2\pi 10^{-i},\ i = 1, 2, \ldots, l$. This gives us a solution

$$\alpha^* = \frac{1}{2}\Big(\sum_{i=1}^{l}(1 - \delta_i)10^i + 1\Big).$$

Since it works with any l, the VC dimension of the set of functions equals to infinity.

4 Growth Function

We can now bound the growth function using the VC dimension. Then we can obtain the bound for $R(\alpha_l)$ in a form which is most widely used in literature.

Lemma 2 *For $l \geq 2h$,*

$$\sum_{i=0}^{h} \binom{l}{i} \leq \exp\Big\{lH\Big(\frac{h}{l}\Big)\Big\}.$$

Proof of Lemma 2:

Let $\alpha = \frac{h}{l} \leq \frac{1}{2}$ (since $l \geq 2h$). Then we have following inequality.

$$\begin{aligned} 1 &= (\alpha + 1 - \alpha)^l && (5)\\ &\underset{binomial\ theorem}{\geq} \sum_{i=1}^{\alpha l} \binom{l}{i} \alpha^i (1-\alpha)^{l-i} && (6)\\ &= \sum_{i=1}^{\alpha l} \binom{l}{i} \Big(\frac{\alpha}{1-\alpha}\Big)^i (1-\alpha)^l && (7)\\ &\geq \exp\Big\{-lH(\alpha)\Big\} \sum_{i=0}^{h} \binom{l}{i} && (8) \end{aligned}$$

Eq. (8) holds because $\alpha \leq \frac{1}{2}$ and consequently $\frac{\alpha}{1-\alpha} \leq 1$. To get Lemma 2, divide both sides with the exponential term. ■

Lemma 3 *For $l \geq 2h$,*

$$N^\Lambda(l) \underset{VC\ lemma}{\leq} \sum_{i=0}^{h} \binom{l}{i} \leq \Big(\frac{el}{h}\Big)^h$$

Proof of Lemma 3:

Here is the proof for the second inequality.

$$\begin{aligned} \sum_{i=0}^{h} \binom{l}{i} &\leq \exp\Big\{lH\Big(\frac{h}{l}\Big)\Big\} && (9)\\ &= \Big(1 - \frac{h}{l}\Big)^{-l} \Big(\frac{l}{h}\frac{l-h}{l}\Big)^h && (10)\\ &\leq e^h \Big(\frac{l-h}{h}\Big)^h && (11)\\ &\leq e^h \Big(\frac{l}{h}\Big)^h && (12)\\ &= \Big(\frac{el}{h}\Big)^h \quad ■ && (13) \end{aligned}$$

It is important to note that the Eq. (13) is polynomial in l. Using Lemma 3, the growth function is now bounded by

$$G^{\Lambda}(l) \quad \leq \quad \log\left(\frac{el}{h}\right)^{h} \quad = \quad h\log\left(\frac{el}{h}\right). \tag{14}$$

After plugging the last term in Eq. (14) into the $H^{\Lambda}_{ann}(l)$ part of the inequality for $R(\alpha_l)$, we finally get the desired inequality:

$$R(\alpha_l) \leq R_{emp}(\alpha_l) + \sqrt{\frac{h\left(\ln\frac{2l}{h}+1\right)-\ln\frac{\eta}{4}}{l}} + \frac{1}{l} \tag{15}$$

which holds with probability greater than $1-\eta$.