CSE522, Winter 2011, Learning Theory Lecture 11 - 02/03/2011

Growth Function and VC-dimension

Lecturer: Ofer Dekel *Scribe: Xu Miao*

1 Review of VC theory

Our primary interest so far is deriving the generalization bound for the binary classifiers. We have studied the Rademacher complexity techniques, and shown that the VC bound is an upper bound of the Rademacher complexity bound. For a binary classification with 0-1 loss l, we have

$$R_m(l \circ H) \leq \sqrt{\frac{2 \log g_H(m)}{m}} \tag{1}$$

where $H = \{h : \mathcal{X} \to \{+1, -1\}\}$ is a hypothesis space. The growth function $g_H(m)$ is defined to be the number of ways the hypothesis space H assign an arbitrary m point sample set (Definition 1).

Definition 1.

$$g_H(m) = \max_{S \in \mathcal{X}^m} \left| \{(h(x_1), h(x_2), \cdots, h(x_m))\}_{h \in H} \right|$$

The growth function can be bounded by applying Sauer's Lemma, i.e., $g_H(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{em}{d}\right)^d = O(m^d)$, where d is the VC-dimension of H (Definition 2)

Definition 2.

$$VCdim(H) = \max\{|S| : H \text{ shatters } S\} = \max\{m : g_H(m) = 2^m\}$$

In summary, the VC bound is stated in Theorem 3.

Theorem 3. *$\forall \delta > 0$, w.p. $\geq 1 - \delta$, over a random sampling of $S \sim \mathcal{D}^m$, $\forall h \in H$, we have*

$$l(h; \mathcal{D}) \leq l(h; S) + \sqrt{\frac{2d \log(em/d)}{m}} + \sqrt{\frac{\log(1/\delta)}{2m}}$$

If $d < \infty$, this universal convergence is attained as the number of samples goes to infinity. Generally speaking, VC bound is looser than Rademacher complexity bound because Rademacher complexity is distribution dependent, while the growth function is not (see Definition 1).

2 Calculating Growth Function and VC dimension

2.1 Interval Classifiers

$\mathcal{X} \in \mathbb{R}$, $H_I = \{h_{a,b} : a \leq b \in \mathbb{R}\}$, where $h_{a,b}(x) = \begin{cases} +1, & a \leq x \leq b \\ -1, & otherwise \end{cases}$

In last class, we have proven that $VCdim(H_I) = 2$, and $g_{H_I}(m) = \sum_{i=0}^{2} \binom{m}{i}$

2.2 Axis Parallel Rectangle Classifiers

$\mathcal{X} = \mathbb{R}^2$, $H_R = \{h_{l,r,t,b} : l \leq r, b \leq t \in \mathbb{R}\}$, where

$$h_{l,r,t,b}(x) = \begin{cases} +1, & \text{if } x \text{ is inside of the rectangle } (l,r,t,b) \\ -1, & \text{otherwise} \end{cases}$$

In last class, we have proven that $VCdim(H_R) = 4$.

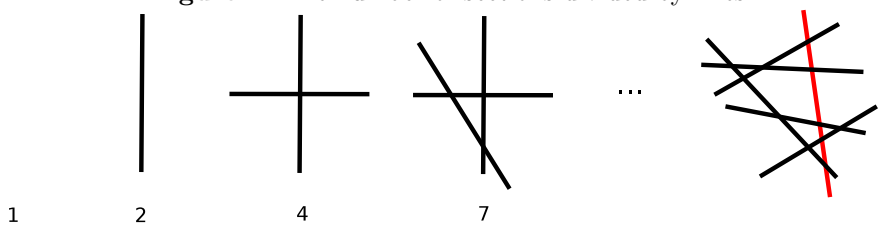
2.3 Linear Classifiers

$\mathcal{X} = \mathbb{R}^d$, $H_L = \left\{h_w \equiv \text{sign}\left(< w, \cdot >\right) : w \in \mathbb{R}^d\right\}$ is the hypothesis space for linear classifiers.

We first consider the following question, how many different sections of the space $\mathbb{R}^d$ m random hyperplanes can divide at most. Let $\Phi_d(m)$ denote this number.

For $d = 2$ dimentional space, $\Phi_d(m)$ is depicted by Figure 1.

Figure 1: The number of sections divided by lines



$$\begin{aligned} \Phi_2(0) &= 1 \\ \Phi_2(1) &= 2 \\ &\cdots \\ \Phi_2(m) &\leq \Phi_2(m-1) + m \end{aligned}$$

The last inequality holds because the m-th new line intersects at most $(m-1)$ previous lines that produces at most m new sections. Therefore, $\Phi_2(m) =_{a.s.} 1 + \sum_{i=1}^{m} i = \binom{m}{0} + \binom{m}{1} + \binom{m}{2}$. The equality holds because the probability of overlaps of random line intersections goes to 0.

Theorem 4. $\Phi_d(m) \leq \sum_{i=0}^{d} \binom{m}{i}$.

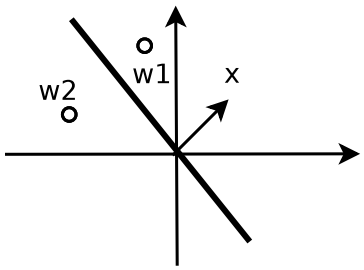
Proof. $m-1$ hyperplanes divide $\mathbb{R}^d$ into at most $\Phi_d(m-1)$ sections. The m-th plane will intersect at most $m-1$ hyperplanes. These $m-1$ hyperplanes are projected onto the m-th hyperplane ($\mathbb{R}^{d-1}$) and divide this space into at most $\Phi_{d-1}(m-1)$ sections, which is also the maximum number of sections in $\mathbb{R}^d$ intersecting with the m-th hyperplane and divided into 2 for each. Therefore,

$$\begin{aligned} \Phi_d(m) &= \Phi_d(m-1) + \Phi_{d-1}(m-1) \\ &\leq \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \\ &= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d} \binom{m-1}{i-1} \\ &= \sum_{i=0}^{d} \binom{m}{i} \end{aligned}$$

□

For linear classifiers, i.e., $h_w(x) = \text{sign}(< w, x >)$, x is a hyperplane that divides the parameter space $\mathbb{R}^d$ into two sections (Figure 2). Any w_1 from the smae side of the plane as x will assign x positive, i.e.,$h_{w_1}(x) = +1$. Any w_2 from the opposite side will assign x negative, i.e., $h_{w_2}(x) = -1$. If there are m

Figure 2: x divides w space into two sections each assigning x in one way.



samples that divide the space into multiple sections, any w from the same section will assign S the same way. According to Theorem 4, there are at most $\Phi_d(m)$ sections. Therefore, $g_{H_L}(m) \leq \Phi_d(m)$. Now, we prove that they are actually equal.

Theorem 5. *For $\mathcal{X} = \mathbb{R}^d$, $VCdim(H_L) = d$.*

Proof. I). we prove $VCdim(H_L) \geq d$ by an example. For $S = \{e_i = (\mathbb{I}_{\{1=i\}}, \mathbb{I}_{\{2=i\}}, \cdots, \mathbb{I}_{\{d=i\}}) : 1 \leq i \leq d\}$, to assign $y = \{+1, -1, \ldots, +1\}$, we can set $< w, e_i >= w_i = y_i$, $\forall i$. Therefore, H_L does shatter d samples.

II). we prove $VCdim(H_L) < d + 1$ by proving that $\forall d + 1$ points, $\exists y \in \{+1, -1\}^{d+1}$, s.t., $\nexists h \in H_L, \forall i, h(x_i) = y_i$. Let $x_1, x_2, \cdots, x_{d+1}$ be arbitrary points in $\mathbb{R}^d$, $\exists i$. s.t. x_i is a linear combination of the rest. Without loss of generality, let $x_{d+1} = \sum_{i=1}^{d} \alpha_i x_i$. Therefore, $< w, x_{d+1} >=< w, \sum_{i=1}^{d} \alpha_i x_i >= \sum_{i=1}^{d} \alpha_i < w, x_i >$. Consider $y = (\text{sign}(\alpha_1), \cdots, \text{sign}(\alpha_d), -1)$, assume that $\exists w$, $\forall i$, $\text{sign}(< w, x_i >) = y_i$, then $\text{sign}(< w, x_{d+1} >) = \text{sign}\left(\sum_{i=1}^{d} \alpha_i < w, x_i >\right) = +1$ that conflicts with y_{d+1}.

Overall, $VCdim(H_L) = d$. □

2.4 Affine Classifiers

$\mathcal{X} = \mathbb{R}^d$, $H_A = \{h_{w,b} \equiv \text{sign}(< w, \cdot > +b) : w \in \mathbb{R}^d, b \in \mathbb{R}\}$.

Lemma 6. *(Radon's Lemma): Any $d + 2$ points in $\mathbb{R}^d$ can be partitioned into disjoint sets N and P, s.t., $convex(N) \bigcap convex(P) \neq \emptyset$.*

Proof. $\forall x_1, x_2, \cdots, x_{d+2}$, $\exists_{nontrivial} \alpha_1, \alpha_2, \cdots, \alpha_{d+2} \in \mathbb{R}$, s.t.,

$$\sum_{i=0}^{d+2} \alpha_i x_i = \overrightarrow{0} \qquad \text{d equalities}$$

$$\sum_{i=0}^{d+2} \alpha_i = 0 \qquad \text{1 equality}$$

We split the points into two disjoint sets, $N = \{i : \alpha_i < 0\}$ and $P = \{i : \alpha_i \geq 0\}$.

$$\begin{aligned} \sum_{i \in N} (-\alpha_i) &= \sum_{i \in P} \alpha_i \equiv \beta > 0 \\ \sum_{i \in N} (-\alpha_i) x_i &= \sum_{i \in P} \alpha_i x_i \end{aligned}$$

By combining these eqaulities, we obtain $\sum_{i \in N} (-\frac{\alpha_i}{\beta}) x_i = \sum_{i \in P} (\frac{\alpha_i}{\beta}) x_i$. The LHS is in $convex(N)$, and the RHS is in $convex(P)$, i.e., $convex(N) \bigcap convex(P) \neq \emptyset$. □

Theorem 7. *For $\mathcal{X} = \mathbb{R}^d$, $VCdim(H_A) = d + 1$*

Proof. I). H_A shatter $S = \left\{e_1, e_2, \cdots, e_d, \overrightarrow{0}\right\}$, where $\overrightarrow{0}$ is the origin. The argument is similar to part I) in the proof of Theorem 5.

II). We prove that $\forall S = \{x_1, x_2, \cdots, x_{d+2}\}$, $\exists y \in \{+1, -1\}^{d+2}$, s.t., $\nexists h \in H_A$, $\forall i$, $h(x_i) = y_i$.

According to Radon's Lemma, $\exists_{nontrivial} \alpha_1, \cdots, \alpha_{d+2} \in \mathbb{R}$, that $\sum_{i\in N}(-\frac{\alpha_i}{\beta})x_i = \sum_{i\in P}(\frac{\alpha_i}{\beta})x_i$. Therefore,

$$\forall w \in \mathbb{R}^d, b \in \mathbb{R}, \quad \left\langle w, \sum_{i\in N}(-\frac{\alpha_i}{\beta})x_i \right\rangle + b = \left\langle w, \sum_{i\in P}(\frac{\alpha_i}{\beta})x_i \right\rangle + b \tag{2}$$

For y that $y_{i\in N} = -1$, and $y_{i\in P} = +1$, we prove that $\nexists h \in H_A$ $\forall i$ $h(x_i) = y_i$ by contradiction.

We Assume that there is a w and a b such that $\forall i, \text{sign}\,(< w, x_i > +b) = y_i$, i.e. $< w, x_{i\in N} > +b < 0$ and $< w, x_{i\in P} > +b > 0$. Since $\alpha_{i\in N} < 0$, $\alpha_{i\in P} > 0$ and $\beta > 0$, we obtain

$$\begin{aligned}
\left\langle w, \sum_{i\in N}(-\frac{\alpha_i}{\beta})x_i \right\rangle + b &= \sum_{i\in N}(-\frac{\alpha_i}{\beta})(\langle w, x_i\rangle + b) < 0 \\
\left\langle w, \sum_{i\in P}(\frac{\alpha_i}{\beta})x_i \right\rangle + b &= \sum_{i\in P}(\frac{\alpha_i}{\beta})(\langle w, x_i\rangle + b) > 0 \\
\Rightarrow \quad \left\langle w, \sum_{i\in N}(-\frac{\alpha_i}{\beta})x_i \right\rangle + b &\neq \left\langle w, \sum_{i\in P}(\frac{\alpha_i}{\beta})x_i \right\rangle + b
\end{aligned}$$

This contradicts with Equation 2. Therefore, $\nexists h \in H_A$ $\forall i$ $h(x_i) = y_i$.

Overall, $VCdim(H_A) = d + 1$. □

2.5 Bit Classifiers

So far, we have seen that the VC dimension coincides with the geometric dimension of the hypothesis space for H_I,H_R,H_L and H_A. However, it is not true in general. Here is a negative example.

$\mathcal{X} = \mathbb{N} = \{1, 2, 3, \cdots\}$, $H_B = \{h_\alpha : \alpha \in \mathbb{R}\}$, where

$$h_\alpha(x) = \begin{cases} +1, & \text{if the x-th bit in the binary representation of } \alpha \text{ is 1} \\ -1, & \text{if the x-th bit in the binary representation of } \alpha \text{ is 0} \end{cases}$$

Obviously, $\forall m, S \in \mathcal{X}^m$ can be shattered by H_B, therefore $VCdim(H_B) = \infty$.

In general, one can use space filling curves, e.g., Peano curves, to encode arbitrary dimensional real space into one dimensional real line. Hence, the geometric dimension does not necessarily imply the VC dimension.

2.6 Union Classifiers

$H_U^K = \left\{\tilde{h} \equiv \bigcup_{i=1}^K h_i : \forall i h_i \in H\right\}$, where

$$\tilde{h}(x) = \bigcup_{i=1}^{K} h_i(x) = \begin{cases} +1, & \text{if } \exists i, \text{ s.t. } h_i(x) = +1 \\ -1, & \text{if } \forall i, \text{ s.t. } h_i(x) = -1 \end{cases}$$

Lemma 8. *(Blumer, Ehrentrecht,Haussler,Warmuth 89') let $VCdim(H) = d$, $K \geq 1$*

$$VCdim(H_U^K) \leq 2Kd\log(5K)$$

Proof.

$$\begin{aligned} g_{H_U^K}(m) &\leq (g_H(m))^K \leq \left(\frac{em}{d}\right)^{dK} \\ VCdim(H_U^K) &= \max\left\{m : g_{H_U^K}(m) = 2^m\right\} \\ &\leq \min\left\{m : \left(\frac{em}{d}\right)^{dK} \leq 2^m\right\} \\ &= \min V \end{aligned}$$

We can verify that $m = 2Kd\log(5K) \in V$, for $K \geq 1$. Therefore, $VCdim(H_U^K) \leq 2Kd\log(5K)$ □