CS281B/Stat241B (Spring 2003) Statistical Learning Theory	**Lecture: 20**

Covering numbers of VC-classes

Lecturer: Peter L. Bartlett *Scribe: Ambuj Tewari*

1 Recapitulation

In the previous lecture we stated a theorem due to Haussler which gives a bound on the packing numbers of a class with finite VC dimension. We saw that bounds on packing numbers imply bounds on covering numbers and vice versa. The proof of Haussler's result involves 1-inclusion graphs and we started to look at it in the previous lecture.

In this lecture, we will first state and prove a two-part lemma and use it to complete the proof. Later, we will introduce another combinatorial dimension, the *pseudodimension*.

2 Proof of Haussler's result

If $V \subseteq \{0,1\}^n$ has $d_{VC}(V) = d$, the 1-inclusion graph (V, E) has nodes V, and (undirected) edges $E = \{\{v, w\} : v \text{ and } w \text{ differ in exactly one bit}\}$. We now have the following lemma.

Lemma 2.1. The 1-inclusion graph has the following properties.

(i) For all subgraphs (W, F) of the 1-inclusion graph, $|F|/|W| \leq d$.

(ii) The edges of (V, E) can be directed so that the maximum outdegree of any node is no more than d.

Proof. Let us first prove part (i) of the lemma. Recall the shifting proof of Sauer's Lemma, where we represented elements of V as rows in a table. We repeatedly changed 1's to 0's if doing that did not lead to a row that was already in the table (see lecture 15 for details). Let (V^*, E^*) be the new 1-inclusion graph for the table obtained after doing the shifting operations.

We can extend the proof to show that,

$$\frac{|E|}{|V|} \leq \frac{|E^*|}{|V^*|} \leq d$$

This is because,

- Our way of doing shifting operations and the fact that the VC-dimension of V is d implies that no row of V^* has more than d 1's. Therefore, each $v \in V$ is 1 bit away from no more than d elements of V having fewer 1's. Hence, $|E^*| \leq d|V^*|$.
- When we shift a column, $|V|$ is unchanged and $|E|$ does not decrease. Therefore, $|E|/|V| \leq |E^*|/|V^*|$.

To prove part (ii), we shall make use of Hall's Theorem which states that a set system $S_1, \ldots, S_k$ has a set of k distinct representatives $x_1 \in S_1, \ldots, x_k \in S_k$ if and only if for every $l \leq k$, the union of any l of the S_i contains at least l elements.

Define a set system consisting of sets $S_e = \{u^{(1)}, \ldots, u^{(d)}, v^{(1)}, \ldots, v^{(d)}\}$ for each edge $e = \{u, v\} \in E$.

Then, any $l \leq |E|$ of these correspond to a subgraph (V', E') with $|E'| = l$ and the union U of these sets has $d|V'|$ distinct elements. But part (i) implies that,

$$|U| = d|V'| \geq |E'| = l$$

Hence there exist distinct representatives $v_e \in S_e$ for all $e \in E$. We can use these distinct representatives to define directions of edges as follows. Direct the edge $\{u, v\}$ as (u, v) if S_e's representative v_e belongs to $\{u^{(1)}, \ldots, u^{(d)}\}$ and as (v, u) otherwise. We make the following two observations.

- Every edge is assigned a direction because each S_e has a representative.
- Outdegree of a vertex $\leq$ number of copies of each vertex $\leq d$.

$\square$

We can now proceed to prove the main result at last.

Theorem 2.2. For $V \subseteq \{0, 1\}^n$ with $d_{VC}(V) = d$, and metric d_1 defined as:

$$d_1(a, b) \triangleq \frac{1}{n} \sum_{i=1} n\, |a_i - b_i|$$

we have,

$$\mathcal{M}(\epsilon, V, d_1) \leq 2 \left(\frac{4e}{\epsilon} \right)^d$$

Proof. We shall prove the theorem by first proving two lemmas. The result will follow simply by combining the two lemmas.

For $m \geq 1$, suppose that we chose a subset $I = \{i_1, \ldots, i_m\}$ uniformly among all subsets of $\{1, \ldots, n\}$ of size M. Let $\Pi_m : \{0, 1\}^n \to \{0, 1\}^m$ denote the random projection

$$\Pi_m(a) = (a_{i_1}, a_{i_2}, \ldots, a_{i_m})$$

Let $W = \Pi_m(V)$. Suppose that $Y, Y' \in V$ are chosen uniformly and independently. Let $X = \Pi_m(Y)$ and $X' = \Pi_m(Y')$.

Lemma 2.3.

$$\frac{1}{m} \sum_{i=1}^{m} P(X_i \neq X'_i \mid X_j = X'_j \text{ for } j \neq i) \leq \frac{2d}{m}$$

We saw last time that if (W, F) is the 1-inclusion graph of W and we define,

$$F_i = \{\{u, v\} \in F \ : \ u_j = v_j \text{ for } j \neq i\}$$

then,

$$\begin{aligned}
& P(X_i \neq X'_i \mid X_j = X'_j \text{ for } j \neq i) \\
\leq \quad & 2 \sum_{\{u,w\} \in F_i} \min\{P(u), P(w)\} \\
\leq \quad & 2 \sum_{(u,w) \in \vec{F}_i} P(v) \\
\leq \quad & 2 \max\nolimits_{v \in W} \text{outdegree}(v) \\
\leq \quad & 2d
\end{aligned}$$

where $\vec{F}_i$ is the version of F_i when edges are directed.

Lemma 2.4. If V is ϵ-separated, for any Π_m,

$$P(X_m \neq X'_m \mid X_j = X'_j \text{ for } j \neq m) \geq \left(1 - \frac{(em/d)^d}{|V|}\right)\epsilon$$

To prove this lemma, define a partition $C_1, \ldots, C_M$ of V defined by

$$\{C_1, \ldots, C_M\} = \{\{v \in V \ : \ \Pi_{m-1}(v) = w\} \ : \ w \in W = \Pi_{m-1}(v)\}$$

We then have,

$$\begin{aligned}
& P(X_m \neq X'_m \mid X_j = X'_j, j \neq m) \\
= & \sum_{j=1}^{M} P(C_j) P(Y_{i_m} \neq Y'_{i_m} \mid Y, Y' \in C_j) \\
= & \sum_{i=1}^{M} \frac{|C_j|}{|V|} P(Y \neq Y' \mid Y, Y' \in C_j) P(Y_{i_m} \neq Y'_{i_m} \mid Y \neq Y', Y, Y' \in C_j) \\
= & \sum_{i=1}^{M} \frac{|C_j|}{|V|} \left(1 - \frac{1}{|C_j|}\right) P(Y_{i_m} \neq Y'_{i_m} \mid Y \neq Y', Y, Y' \in C_j) \\
\geq & \sum_{i=1}^{M} \frac{|C_j|}{|V|} \left(1 - \frac{1}{|C_j|}\right) \epsilon \\
= & \left(1 - \frac{M}{|V|}\right) \epsilon \\
\geq & \left(1 - \frac{(e(m-1)/d)^d}{|V|}\right) \epsilon
\end{aligned}$$

where the first inequality follows because of ϵ-separation and the second is a consequence of Sauer's lemma.

Combining the two lemmas,

$$\left(1 - \frac{(em/d)^d}{\mathcal{M}(\epsilon, V, d_1)}\right) \epsilon \leq \frac{2d}{m}$$

$$\Rightarrow \mathcal{M}(\epsilon, V, d_1) \leq \frac{(em/d)^d}{1 - \frac{2d}{\epsilon m}} \leq 2 \left(\frac{4e}{\epsilon}\right)^d$$

□

Define the metric d_2 on $\{0,1\}^n$ as,

$$d_2(a, b) \triangleq \sqrt{\frac{1}{n} \sum_{i=1}^{n} (a_i - b_i)^2}$$

For two elements a, b of $\{0,1\}^n$, $d_1(a, b) = (d_2(a, b))^2$, therefore,

$$\mathcal{M}(\epsilon, V, d_1) = \mathcal{M}(\sqrt{\epsilon}, V, d_2)$$

Last time we saw inequalities relating packing numbers and covering numbers. Using them we have,

$$\mathcal{N}(\epsilon, V, d_2) \leq \mathcal{M}(\epsilon, V, d_2) = \mathcal{M}(\epsilon^2, V, d_1) \leq 2 \left(\frac{4e}{\epsilon^2}\right)^d$$

It follows then that, if F is a class of binary values functions, the inequality given below holds for some constant c.

$$\sup_P \mathcal{N}(\epsilon, F, L_2(P)) \leq \left(\frac{c}{\epsilon}\right)^2 d$$

$$\Rightarrow R_n(F) \leq c'\sqrt{\frac{d}{n}}$$

3 Another combinatorial dimension: Pseudodimension

For $F \subseteq \mathbb{R}^{\mathcal{X}}$ (real valued functions), the pseudodimension of F is

$$d_P(F) = d_{VC}(\{(x,y) \mapsto \text{sign}(f(x) - y) \ : \ f \in F\})$$

E.g., if $F \subseteq$ linear space, $d_P(F) \leq \dim(F)$.

We have seen that $\mathcal{N}(\epsilon, F, L_2(P_n)) \leq \left(\frac{c}{\epsilon}\right)^{\dim(F)}$, for $F \subseteq [-1,1]^{\mathcal{X}}$. (c.f. Sauer's Lemma: $\Pi_F(n) \leq \left(\frac{en}{d}\right)^d$)

Example F = non-decreasing functions from $\mathbb{R}$ to $[0,1]$. Then $d_P(F) = \infty$, but $\mathcal{N}(\epsilon, F, L_2(P_n)) \leq n^{(1/\epsilon)}$.

We have seen that,

$$\Pi_F(n) \text{ polynomial in n } \Leftrightarrow d_{VC}(F) < \infty$$

$$\sup_{P_n} \mathcal{N}(\epsilon, F, L_2(P_n)) \text{ polynomial in n } \not\Rightarrow d_P(F) < \infty$$

$$\sup_{P_n} \mathcal{N}(\epsilon, F, L_2(P_n)) \text{ polynomial in n } \Leftarrow d_P(F) < \infty$$

Theorem 3.1. For $F \subseteq [0,1]^{\mathcal{X}}$, $\forall P$,

$$\mathcal{M}(\epsilon, F, L_2(P)) \leq 2\left(\frac{4e}{\epsilon^2}\right)^{d_P(F)}$$

Proof. Define U be a random variable distributed uniformly over $[0,1]$. That is, $P(U \leq a) = a$. For $f, g \ : \ \mathcal{X} \to [0,1]$,

$$\begin{aligned}
\|f-g\|^2 &= \mathbb{E}(f(X) - g(X))^2 \\
&= \mathbb{E}\left(P(f(X) > U \mid X) - P(g(X) > U \mid X)\right)^2 \\
&= \mathbb{E}\left(\mathbb{E}(1[f(X) > U] - 1[g(X) > U] \mid X)\right)^2 \\
&= \mathbb{E}(1[f(X) > U] - 1[g(X) > U])^2 \\
&= \mathbb{E}\left|\text{sign}(f(X) - U) - \text{sign}(g(X) - U)\right|
\end{aligned}$$

So,

$$\mathcal{M}(\epsilon, F, L_2(P)) \leq \mathcal{M}(\epsilon^2, G, L_1(P \times U)) \leq 2\left(\frac{4e}{\epsilon^2}\right)^{d_P(F)}$$

with $G = \{(x,y) \mapsto (\text{sign}(f(x) - y) \ : \ f \in F\}$. □

Estimates for Pseudodimension
If we can compute $F = \{x \mapsto f(a,x) \ : \ a \in \mathbb{R}^d\}$ in $\leq t$ arithmetic operations on real numbers, then we can compute $G = \{(x,y) \mapsto \text{sign}(f(a,x) - y) \ : \ a \in \mathbb{R}^d\}$ in $O(t)$ steps.

Hence $d_P(F) = O(dt) \approx 4d(t+4)$.