CS 7545: Machine Learning Theory **Fall 2019**

Lecture 23: Sauer's Lemma + Complete Upper and Lower Bound Sketch

Lecturer: Jacob Abernethy *Scribes: Youngho Yoo and Mohammadreza Zandehshahvar*

23.1 Sauer's lemma

We continue and finish the proof of Sauer's lemma from the previous lecture.

Lemma 23.1 (Sauer) *Let $\mathcal{H}$ be a binary function class with VC-dimension d. Then*

$$\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq m^d$$

where

$$\Pi_{\mathcal{H}}(m) := \sup_{\substack{S \subseteq X \\ S=\{x_1,\ldots,x_m\}}} |\{(h(x_1),\ldots,h(x_m)) : h \in \mathcal{H}\}|$$

Proof: Let $\mathcal{M}_{S,\mathcal{H}}$ be a matrix with m columns and (unique) rows

$$(h(x_1),\ldots,h(x_m)) : h \in \mathcal{H}$$

We need to show that the number of rows of $\mathcal{M}_{S,\mathcal{H}}$ is at most $\sum_{i=0}^{d} \binom{m}{i}$. In the last lecture we showed the following claim:

Claim 23.2 *If every row of a matrix has at most d 1's in each row, then it has at most $\sum_{i=0}^{d} \binom{m}{i}$ rows.*

We cannot expect the rows of $\mathcal{M}_{S,\mathcal{H}}$ to satisfy such a condition, but we will show that we can modify $\mathcal{M}_{S,\mathcal{H}}$ without increasing its VC-dimension so that the condition in the claim is satisfied. We modify $\mathcal{M}_{S,\mathcal{H}}$ as in the following procedure:

For epoch $e = 1, 2, \cdots$:
  For column $j = 1, \ldots, m$:
    For row $i = 1, \ldots, \#\text{rows}$:
      Set $\mathcal{M}_{S,\mathcal{H}}(i,j) = 0$ if no duplicate row is created } `shift`(j)

Example: If we start with the left matrix below, we get the matrix on the right after epoch 1. Then there are no more 1's that can be turned into a 0 without creating a duplicate row, so we are done.

$$\begin{matrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{matrix} \quad \Longrightarrow \quad \begin{matrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{matrix}$$

We say that a subset U of the columns of a 01-matrix is *shattered* if every binary vector $\vec{b} \in \{0,1\}^U$ is present in some row restricted to the columns in U. Notice that the last two columns of the resulting matrix in the example above is shattered, since all of $(0,0), (0,1), (1,0), (1,1)$ are present in some row restricted to the last two columns.

Claim 23.3 *After the procedure, if for some subset $U \subseteq [m]$ of columns there exists a row with all 1's on U, then U is shattered.*

Proof: Since there is a row with all 1's and each 1 could not be changed to 0 without creating a duplicate row, it follows that every vector with $|U| - 1$ 1's and one 0 on U is present in some row. Similarly, every vector with $|U| - 2$ 1's and two 0's is present in some row, and continuing in this manner it follows that every binary vector on U is present in some row. Therefore U is shattered. ■

It follows that the VC-dimension of the resulting matrix is equal to the maximum number of 1's present in any one row. If we can show that the procedure does not increase the VC-dimension from the starting matrix $\mathcal{M}_{S,\mathcal{H}}$, then this would show that the number of rows of $\mathcal{M}_{S,\mathcal{H}}$ is at most $\sum_{i=0}^{d} \binom{m}{i}$ and we would be done.

For a column $j \in [m]$, let $\texttt{shift}(j)$ denote the operation of setting the (i,j)-th entry to 0 if no duplicate row is created for each $i = 1, \ldots, \#\text{rows}$ (i.e. the last two lines of the procedure).

Claim 23.4 *For any column j, the $\texttt{shift}(j)$ operation does not increase the VC-dimension of the matrix.*

Proof: Suppose to the contrary that, for some column j, applying $\texttt{shift}(j)$ increases the VC-dimension. Let M and M' denote the matrix before and after this $\texttt{shift}(j)$ operation respectively. Then there is some $U \subseteq [m]$ that is not shattered in M but becomes shattered in M'. This implies that a new dichotomy was created, which implies that there exist rows i and k such that

1. $M(i,U) = M(k,U)$ before $\texttt{shift}(j)$,
2. $M'(i,U) \neq M'(k,U)$ after $\texttt{shift}(j)$, and moreover:
3. the row vector obtained from $M(i,U)$ by shifting $M(i,j)$ to 0 is not already present in another row.

We may assume without loss of generality that $M'(i,j) = 1$ and $M'(k,j) = 0$ (and $M(i,j) = M(k,j) = 1$). Since $M(i,j)$ was not shifted to 0 in $\texttt{shift}(j)$, there exists a row $\ell \notin \{i,k\}$ that would have been a duplicate if $M(i,j)$ was shifted. Note that we cannot have $M(\ell,j) = 1$, since this would be a duplicate of $M(i,U)$. Hence $M(\ell,j) = 0$. But this contradicts the assumption (3.) that a new dichotomy was created, a contradiction. ■

So the resulting matrix $\mathcal{M}'_{S,\mathcal{H}}$ after the procedure does not have a larger VC-dimension than the starting matrix $\mathcal{M}_{S,\mathcal{H}}$, which has VC-dimension d. By Claim 23.3, all rows of $\mathcal{M}'_{S,\mathcal{H}}$ have at most d 1's, and it follows that the number of rows of $\mathcal{M}_{S,\mathcal{H}}$ is at most $\sum_{i=0}^{d} \binom{m}{i}$, completing the proof. ■

23.2 Growth Function Generalization Bound

Given hypothesis class $\mathcal{H} : \mathcal{X} \to \hat{\mathcal{Y}}$ and loss function $l : \hat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$, the **loss class** is the set of functions

$$G_{\mathcal{H}} := \{g_h(x,y) = l(h(x),y) : h \in \mathcal{H}\}$$

Recall that we want to bound the following:

$$\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_i l(h(x_i), y_i) - \mathbb{E}_{(x,y)\sim D}[l(h(x),y)]$$

which is equal to

$$\sup_{g_h \in G_{\mathcal{H}}} \hat{\mathbb{E}}_{\mathcal{S}} g_h - \mathbb{E} g_h$$

where $S = \{(x_1,y_1), \ldots, (x_m,y_m)\}$.

Claim 23.5 *Let $\mathcal{H}$ be a binary function class and let $G_{\mathcal{H}}$ be the loss class for the $\{0,1\}$-loss. Then*

$$\hat{\mathfrak{R}}_{\mathcal{S}}(G_{\mathcal{H}}) = \hat{\mathfrak{R}}_{S|_x} \mathcal{H}$$

where $S|_x = \{x_1, x_2, \ldots, x_m\}$

Proof:

$$\begin{aligned}
\hat{\Re}_{\mathcal{S}}(G_{\mathcal{H}}) &= \mathbb{E}_{\sigma_{1:m}}\left[\frac{1}{m}\sup_{g_h \in G_{\mathcal{H}}}\sum_{i=1}^{m}\sigma_i g_h\left(x_i, y_i\right)\right] \\
&= \mathbb{E}_{\sigma_{1:m}}\left[\frac{1}{m}\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i l\left(h(x_i), y_i\right)\right] \\
&= \mathbb{E}_{\sigma_{1:m}}\left[\frac{1}{m}\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i \mathbb{1}\left[h(x_i) \neq y_i\right]\right] \\
&= \mathbb{E}_{\sigma_{1:m}}\left[\frac{1}{2m}\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i (2\cdot\mathbb{1}\left[h(x_i) \neq y_i\right] - 1)\right] \\
&= \mathbb{E}_{\sigma_{1:m}}\left[\frac{1}{2m}\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i (2\cdot\mathbb{1}\left[h(x_i) \neq 0\right] - 1)\right] \quad (*) \\
&= \mathbb{E}_{\sigma_{1:m}}\left[\frac{1}{2m}\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i (2h(x_i) - 1)\right] \\
&= \mathbb{E}_{\sigma_{1:m}}\left[\frac{1}{m}\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i h(x_i)\right] \\
&= \Re_{\mathcal{S}|_x}(\mathcal{H})
\end{aligned}$$

where the equality in line $(*)$ is due to the fact that $y_i \in \{0, 1\}$ and $(2 \cdot \mathbb{1}[h(x_i) \neq y_i] - 1) \in \{+1, -1\}$, so $\mathbb{E}_{\sigma_i}\sigma_i(2 \cdot \mathbb{1}[h(x_i) \neq y_i] - 1) = \mathbb{E}_{\sigma_i}\sigma_i(2 \cdot \mathbb{1}[h(x_i) \neq 0] - 1)$ since σ_i is a Rademacher variable.

■