CS 7545: Machine Learning Theory	**Fall 2018**

Lecture 23: Overview of VC-Dimension Upper & Lower Bounds

Lecturer: Jacob Abernethy — *Scribes: Jinsol Lee, Gukyeong Kwon*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

23.1 Review: Sauer's Lemma

Let $\mathcal{H}$ be a class of binary functions with a VC dimension of d.

$$\prod_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq \underbrace{\left(\frac{me}{d}\right)^d \leq m^d}_{\text{if } d \geq 3}$$

$$\prod_{\mathcal{H}}(m) = \max_{\substack{\mathcal{S} \subseteq \mathcal{X} \\ |\mathcal{S}|=m \\ \{x_1,\cdots,x_m\}=\mathcal{S}}} \left|\{(h(x_1),\cdots,h(x_m)) : h \in \mathcal{H}\}\right|$$

This is the largest number of dichotomies that can be produced in a space $\mathcal{X}$.

Proof: Let M be a matrix whose rows are unique vectors from $\{(h(x_1),\cdots,h(x_m)) : h \in \mathcal{H}\}$ for a fixed $\mathcal{S} = \{x_1,\cdots,x_m\} \subseteq \mathcal{X}$

- Goal: Show that the number of rows of $M \leq \sum_{i=0}^{d} \binom{m}{i}$
- Trick: Modify M to be sparse.
  - Shift column j of M such that for each row i, $M_{ij} = 1$.
  - Set $M_{ij} = 0$ if it does not create duplicates.
- Procedure: Continue shifting columns one by one until it's not possible to shift further.

$$\text{ex) } M = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \quad \Rightarrow \quad M' = \begin{bmatrix} 0 & 1 & 0 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 0 & 0 & 1 & 1 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & 0 & 1 \\ 0 & \mathbf{0} & \mathbf{0} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

- Facts:
  1. M' has no duplicated rows
  2. Given $Q \subset [m]$, if $\exists$ row i such that $M'_{ij} = 1 \;\; \forall j \in Q$, then $\underline{M' \text{ shatters } Q}$.
     - M' shatters Q: M' restricted to columns Q has all possible $2^{|Q|}$ rows.
  3. VC-dim$(M') \leq$ VC-dim$(M) \leq$ VC-dim$(\mathcal{H})$
     - Specifically, if a column j is part of a shattered set Q after shifting, then Q was shattered before as well.

     Proof: Use proof by contradiction. Assume that col j is part of a shattered set Q after shifting, but Q was not shattered before the procedure. Now rearrange the columns so it the columns not in Q are on the left side of column j and other columns in Q are on the right side of column j. The procedure must have created an combination in set Q that did not exist before by changing 1 digit from 1 to 0, assume it happened on row i column j. There must be another row i' that share same value as row i in set Q (otherwise changing M_{ij} from 1 to 0 will leave its original combination in Q missing, then set Q is still not shattered), also the 1 on row i' cannot be changed to 0 during the procedure. Because of no duplicate rows rule, and row i and i' share the same value in set Q, they must have difference(s) in some column(s) outside set Q. Also the combination in Q that row i will achieve by changing M_{ij} from 1 to 0 cannot exist initially(Otherwise the procedure will not create new combination in Q). Now no rows will prevent both row i and i' changing their col j to 0, which contradicts the assumption.

$$M = \begin{array}{c} \begin{matrix} \overbrace{\quad\quad}^{other} & \overbrace{\quad}^{j} & \overbrace{\quad}^{Q} \end{matrix} \\ \begin{bmatrix} \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & 1 & q_i \\ \cdots & 1 & q'_i \\ \vdots & \vdots & \vdots \end{bmatrix} \end{array}$$

Row $\begin{bmatrix} \cdots & 0 & q_i \end{bmatrix}$ cannot exist. ■

The above facts combined imply,

$$\begin{aligned} \text{Number of rows in } M &= \text{Number of rows in } M' \\ &\leq \text{Number of subsets of } [m] \text{ with } \leq d \text{ elements} \\ &= \sum_{i=0}^{d} \binom{m}{i} \end{aligned}$$

- Subclaim: VC-dim(M') = Largest number of 1's in a row.

■

23.2 Growth Function Generalization Bound

Recap $\mathcal{H}$: binary class, $\ell(\cdot,\cdot)$: $0-1$ loss

$$R(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\ell(h(x),y)\big]$$
$$\hat{R}_s(h) = \frac{1}{|\mathcal{S}|}\sum_{(x_i,y_i)\in\mathcal{S}} \ell(h(x_i),y_i)$$
$$\text{ERM: } \hat{h} \leftarrow \arg\min_{h\in\mathcal{H}} \hat{R}_s(h)$$

We showed,

$$R(\hat{h}) - \min_{h^*\in\mathcal{H}} R(h^*) \leq 2\sup_{h\in\mathcal{H}} \big|R(h) - \hat{R}_s(h)\big| \tag{23.1}$$

Definition 23.1 (Loss class) *The loss class of hypothesis set $\mathcal{H}$ is defined as follow:*

$$G := \ell \circ \mathcal{H} := \{g_h(z) := \ell(h(x),y) : h \in \mathcal{H}\}$$

With the definition above,

$$\begin{aligned}
\sup_{h\in\mathcal{H}} \left|R(h) - \hat{R}_s(h)\right| &= \sup_{g\in\ell\circ\mathcal{H}} \Big| \mathop{\mathbb{E}}_{z\sim D}[g(z)] - \frac{1}{m}\sum_{z_i\in S} g(z_i)\Big| \\
&\leq 2 \sup_{g\in\ell\circ\mathcal{H}} \mathbb{E}g - \hat{\mathbb{E}}g \\
&\leq 4R_m(\ell\circ\mathcal{H}) + \sqrt{\frac{\log 2/\delta}{2m}} \quad \text{(symmetrization)} \\
&= 4\left(\frac{1}{2}R_n(\mathcal{H})\right) + \sqrt{\frac{\log 2/\delta}{2m}} \\
&\leq \sqrt{\frac{2\log|A|}{m}} \quad \text{(Massart)} \\
&\leq \sqrt{\frac{2\log \Pi_{\mathcal{H}(m)}}{m}} \\
&\leq \sqrt{\frac{2\log m^d}{m}} \quad \text{(Sauer)} \\
&= O\left(\sqrt{\frac{d\log m}{m}}\right)
\end{aligned}$$

where $R_m(\mathcal{H})$ is defined as below:

$$R_m(\mathcal{H}) := \mathop{\mathbb{E}}_{s\sim D}\mathop{\mathbb{E}}_{\sigma_1\cdots\sigma_m}\left[\sup_{h\in\mathcal{H}}\frac{1}{m}\sum_i \sigma_i h(x_i)\right] = \mathbb{E}\mathbb{E}\left[\sup_{a\in\{(h(x_1),h(x_2),\cdots,h(x_m)):h\in\mathcal{H}\}}\frac{1}{m}\sum_i \sigma_i a_i\right]$$

- Fact: For any $\mathcal{H}$ of VC-dim = d, $\exists D\in\Delta(x,y)$ such that $R(\hat{h}) - R(h^*) \geq \sqrt{\frac{d}{m}}$ with probability $\geq c$.

  Proof: First, we sample $\sigma_1,\cdots\sigma_d \sim \{-1,1\}$.

  $$x\sim Unif(shattered_{\mathcal{H}}(\mathcal{X})) \qquad y_i = \begin{cases} 1 & \text{w.p} \quad 1/2 + \sigma_1\sqrt{d/m} \\ 0 & \text{w.p} \quad 1/2 - \sigma_1\sqrt{d/m}\end{cases}$$

  To obtain error $\leq \sqrt{\frac{d}{m}}$, we need $\frac{1}{(\sqrt{d/m})^2}\cdot d$ samples. ■