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

Lecture 22: Massart's Lemma and Sauer's Lemma

Lecturer: Jacob Abernethy *Scribes: Daniel Barg, Hongzhao Guan*

We are continuing towards proving a bound for the testing error of classes of binary functions, and today we proved two important lemmas towards that goal.

22.1 Review and Notation

- Input Space $\mathcal{X}$
- Label Space $\mathcal{Y} : \{-1, 1\}$.
- Class of Function $\mathcal{H} : \mathcal{X} \to \mathcal{Y}$
- Distribution $\mathcal{D} \in \triangle(\mathcal{X} \times \mathcal{Y})$
- Prediction Space $\mathcal{Y}'$
- Loss Function $\ell : \mathcal{Y}' \times \mathcal{Y} \to R$
- Risk $R(h) = \underset{(x,y)\sim\mathcal{D}}{\mathbb{E}}[\ell(h(x), y)]$
- Empirical Risk: $\hat{R}_m(h) = \frac{1}{m} \cdot \sum_{i=1}^{m} \ell(h(x_i), y_i)$
- Given a dataset $\{(x_1, y_1), ..., (x_m, y_m)\}$
- Empirical Risk Minimization (ERM) $\hat{h} = \arg\min_{h\in\mathcal{H}} \hat{R}_m(h)$
- Fact $R(\hat{h}) - \min_{h^*\in\mathcal{H}} R(h^*) \leq 2 \cdot \sup_{h\in\mathcal{H}} |R(h) - \hat{R}_m(h)|$
- Let $\mathcal{G}$ be a class of binary function, give a distribution p on $\mathcal{X}$
  $\mathbb{E}[g] = \underset{x\sim p}{\mathbb{E}}[g(x)]$, $\hat{\mathbb{E}}_s[g] = \frac{1}{|s|} \sum_{X_i\in\mathcal{S}} g(x_i)$ where $\mathcal{S} \sim p^m$ and $|\mathcal{S}| = m$
- Fact $\sup_{g\in\mathcal{G}} E[g] - \hat{\mathbb{E}}_s[g] \leq 2 \cdot \underset{x_1,...,x_m\sim p}{\mathbb{E}} \left[\sup_{g\in\mathcal{G}} \frac{1}{m} \sum_i \sigma_i g(x_i) \right]$

Definition 22.1 (Growth Function) *The growth function $\Pi_{\mathcal{G}}(m)$ of class $\mathcal{G}$ is the following:*

$$\max_{\mathcal{S}\subseteq\mathcal{X}, |S|=m} |\{g(x_1), ..., g(x_m) : g \in \mathcal{G}\}|$$

Definition 22.2 (VC-dimension)

$$VCD(\mathcal{G}) = \max\{m | \Pi_{\mathcal{G}}(m) = 2^m\}$$

22.2 Massart's Lemma

Theorem 22.3 (Massart's Lemma) *Let* $\mathcal{A} \subseteq \mathbb{R}^m$, $\max_{a \in \mathcal{A}} ||a||_2 \leq r$, *then:*

$$\mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_m}\left[\sup_{a\in\mathcal{A}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i a_i\right] \leq \frac{r\sqrt{2\log|\mathcal{A}|}}{m}$$

Note:when $\mathcal{A} \leq 0, 1^m \Rightarrow r \leftarrow \sqrt{m}$, *right hand side becomes* $\sqrt{\dfrac{2\log|A|}{m}}$

Proof:
(Hint: this is essentially Hoeffding)

$$\begin{aligned}
& \exp\left(\lambda \cdot \mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_m}\left[\sup_{a\in\mathcal{A}}\sum_i \sigma_i a_i\right]\right) \forall\lambda > 0 \\
\leq\ & \mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_m}\left[\exp\left(\lambda \sup_{a\in\mathcal{A}}\sum_{i=1}^{m}\sigma_i a_i\right)\right] \text{ By Jensen's} \\
=\ & \mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_m}\left[\sup_{a\in\mathcal{A}}\exp\left(\lambda\sum_{i=1}^{m}\sigma_i a_i\right)\right] \\
\leq\ & \mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_m}\left[\sum_{a\in\mathcal{A}}\exp\left(\lambda\sum_{i=1}^{m}\sigma_i a_i\right)\right] \\
=\ & \sum_{a\in\mathcal{A}}\mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_m}\left[\prod_{i=1}^{m}\exp(\lambda\cdot\sigma_i\cdot a_i)\right] \text{ Since the } \sigma_i\text{'s are IID} \\
=\ & \sum_{a\in\mathcal{A}}\prod_{i=1}^{m}\mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_m}\left[\exp(\lambda\cdot\sigma_i\cdot a_i)\right] \\
\leq\ & \sum_{a\in\mathcal{A}}\prod_{i=1}^{m}\mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_m}\exp\left(\frac{\lambda^2\cdot(2a_i)^2}{8}\right) \text{ Because of Hoeffding's Lemma} \\
=\ & \sum_{a\in\mathcal{A}}\exp\left(\frac{\lambda^2}{2}\cdot\sum_{i=1}^{m}{a_i}^2\right) \\
\leq\ & |\mathcal{A}|\cdot\exp\left(\frac{\lambda^2}{2}\cdot r^2\right)
\end{aligned}$$

Now, take log on both sides and divide by λ, above inequality becomes:

$$\mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_m}\left[\sup_{a\in\mathcal{A}}\sum_{i=1}^{m}a_i\cdot\sigma_i\right] \leq \frac{\log|\mathcal{A}|}{\lambda} + \frac{\lambda^2}{2}\cdot r^2$$

Set:

$$\lambda = \sqrt{\frac{2\log|\mathcal{A}|}{r^2}}$$

Now, the bound follows. ∎

22.3 Sauer's Lemma

Theorem 22.4 (Sauer's Lemma)

$$\Pi_{\mathcal{G}}(m) \leq \sum_{i=0}^{VCD(\mathcal{G})} \binom{m}{i} \leq c \cdot m^{VCD(\mathcal{G})}$$

Proof: [Proof of the second inequality]
Since,

$$\binom{n}{k} \leq \binom{ne}{k}^k$$

Then

$$\sum_{i=1}^{d} \binom{m}{i} \leq \sum_{i=0}^{d} \binom{me}{i} \leq c \cdot m^d$$

■

Proof: [Proof of the first inequality]
Given sample $\mathcal{S} = x_1, ..., x_m$, let M be a matrix whose rows are unique elements of $\{(g(x_1), ..., g(x_m)) : g \in \mathcal{G}\}$, we want to bound number of rows of M, since the upper bound is $\Pi_{\mathcal{G}}(m)$. The problem is that it is hard to analyze this matrix M. To aid in the analysis, we modify M to a matrix M', which we define as follow:

```
For j = 1,...,m :
    For row i = 1,..., number of rows of M :
        if M_ij = 0: do nothing
        if M_ij = 1: M_ij ← 0, only if it does not duplicate another row
```

Call the matrix you get at the end of these operations M'. Here is an example of the shifting process :

$$M = \begin{pmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \Longrightarrow M' = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Now, we make the following claims :

(1) M' has unique rows.

Why? By the definition of our shifting process, we do not create duplicates.

(2) If, in row i, there are k columns $j_1, j_2, \ldots, j_k \in [m]$ such that for $M'_{ij_1} = \ldots = M'_{ij_k} = 1$, then M' shatters these columns. Shattering columns means that all the dichotomies of length k are generated in the rows of the chosen columns. For example, the last 2 columns of M' are shattered.

Why? If, in our chosen subset of columns, there is a row of all 1's, it means that we were unable to shift these 1's down. In other words, any dichotomy with $(k-1)$ ones and 1 zero already exists within the columns, and continuing this logic, every dichotomy of length k exists within these columns.

(3) $\text{VCD}(M') \leq \text{VCD}(M) \leq \text{VCD}(\mathcal{G})$, where VCD of a matrix is the maximum number of shattered columns.

Why? Assigned as an exercise.

Together, (1)+(2)+(3) imply that that:

num rows of M = num rows of $M' \leq$ num subsets of $[m]$ of size less than or equal to $\text{VCD}(\mathcal{G}) = \sum_{i=0}^{VCD(\mathcal{G})} \binom{m}{i}$.

This is true since we just need to count how many ways we can have k ones in a row of M', which has m columns. We cannot have more ones (shatter more columns) than $VCD(\mathcal{G})$ by definition of our matrices, and counting the number of ones is equivalent to counting subsets. Since this bound is independent of the matrix M, we have established a bound for $\Pi_{\mathcal{G}}(m)$. ■