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

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

Lecturer: Jacob Abernethy *Scribes: Namrata Nadagouda, Nauman Ahad*

22.1 Review

- Let $\mathcal{G}$ be a class of functions (interpreted as *the family of loss functions associated to* $\mathcal{H}$) mapping from $\mathcal{Z}$ to $[0,1]$:
$$\mathcal{G} = \{g : (x,y) \mapsto L(h(x),y) : h \in \mathcal{H}\}$$
where $\mathcal{H}$ denotes a hypothesis set,
$\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $\mathcal{X}$ is the sample space and $\mathcal{Y}$ is the set of labels,
L is a loss function $L : \mathcal{Y} \times \mathcal{Y} \mapsto [0,1]$.

- Let $S = (z_1, \ldots, z_m)$ be a fixed sample of size m with elements in $\mathcal{Z}$. We define $\hat{\mathbb{E}}_S[g]$, the empirical mean of g over S and $\mathbb{E}[g]$, the true mean as below
$$\hat{\mathbb{E}}_S[g] = \frac{1}{m}\sum_{i=1}^{m} g(z_i), \quad \mathbb{E}[g] = \mathop{\mathbb{E}}_{Z\sim D}[g(Z)]$$
where D is the distribution according to which the samples S are drawn.

- For a fixed $g \in \mathcal{G}$, we can prove that using the Hoeffding's inequality that
$$\hat{\mathbb{E}}_S[g] - \mathbb{E}[g] \leq \sqrt{\frac{\log 1/\delta}{m}} \quad \text{w.p.} \geq 1-\delta$$
under the assumption that the samples S are drawn independently and identically from the distribution D.

- Define the function Φ for any sample S by
$$\Phi(S) = \sup_{g\in\mathcal{G}} \left(\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]\right).$$
We are interested in bounding this function.

- In the previous lecture, we proved that
$$\Phi(S) \leq \mathfrak{R}_m(\mathcal{G}) + \sqrt{\frac{\log 1/\delta}{2m}}$$
where $\mathfrak{R}_m(\mathcal{G})$ is the Rademacher complexity given by
$$\mathfrak{R}_m(\mathcal{G}) = \mathop{\mathbb{E}}_{S\sim D}\mathop{\mathbb{E}}_{\sigma_{1:m}}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m} g(z_i)\sigma_i\right]$$

22.2 Massart's Lemma

Theorem 22.1 *Let $\mathcal{A} \subseteq \mathbb{R}^m$ be a finite set, with $r = \max_{\mathbf{a}\in\mathcal{A}} \|\mathbf{a}\|_2$, then*

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

where σ_i's are Rademacher random variables (which are independent and identically distributed random variables taking values {-1,1} with equal probability) and a_i are components of vector $\mathbf{a}$.

Proof: Here's a proof of the Massart's Lemma. It basically follows from Hoeffding's Lemma.

$$\begin{aligned}
\exp(\lambda.\mathop{\mathbb{E}}_{\sigma_{1:m}}[\sup_{a\in\mathcal{A}}\sum_{i=1}^{m}\sigma_i.a_i]) &\leq \mathop{\mathbb{E}}_{\sigma_{1:m}}[\exp(\lambda.\sup_{a\in\mathcal{A}}\sum_{i=1}^{m}\sigma_i.a_i] \text{ (Jensen's for } \forall\lambda > 0) \\
&= \mathop{\mathbb{E}}_{\sigma_{1:m}}[\sup_{a\in\mathcal{A}}\exp(\sum_{i=1}^{m}\lambda.\sigma_i.a_i] \\
&\leq \mathop{\mathbb{E}}_{\sigma_{1:m}}[\sum_{a\in\mathcal{A}}\exp(\sum_{i=1}^{m}\lambda.\sigma_i.a_i)] \\
&= \sum_{a\in\mathcal{A}}\mathop{\mathbb{E}}_{\sigma_{1:m}}[\Pi_{i=1}^{m}\exp(\lambda.\sigma_i.a_i)] \text{ (As } \sigma_i\text{'s are i.i.d)} \\
&= \sum_{a\in\mathcal{A}}\Pi_{i=1}^{m}\mathop{\mathbb{E}}_{\sigma_{1:m}}[\exp(\lambda.\sigma_i.a_i)] \\
&\leq \sum_{a\in\mathcal{A}}\Pi_{i=1}^{m}\exp(\frac{\lambda^2.(2a_i)^2}{8}) \text{ (Using Hoeffding's Lemma)} \\
&= \sum_{a\in\mathcal{A}}\exp(\frac{\lambda^2}{2}.\sum_{i=1}^{m}(a_i)^2) \\
&= \sum_{a\in\mathcal{A}}\exp(\frac{\lambda^2}{2}.r^2) \text{ (From definition of } r) \\
&\leq |\mathcal{A}|\exp(\frac{\lambda^2}{2}.r^2)
\end{aligned}$$

Applying the logarithm operator to the inequality and multiplying by $\frac{1}{\lambda}$

$$\begin{aligned}
\frac{1}{\lambda}\log\Big(\exp(\lambda.\mathop{\mathbb{E}}_{\sigma_{1:m}}[\sup_{a\in\mathcal{A}}\sum_{i=1}^{m}a_i.\sigma_i])\Big) &\leq \frac{1}{\lambda}\log\Big(|\mathcal{A}|\exp(\frac{\lambda^2}{2}.r^2)\Big) \\
\mathop{\mathbb{E}}_{\sigma_{1:m}}[\sup_{a\in\mathcal{A}}\sum_{i=1}^{m}a_i.\sigma_i] &\leq \frac{\log|\mathcal{A}|}{\lambda} + \frac{\lambda}{2}.r^2
\end{aligned}$$

Set value of $\lambda = \sqrt{\frac{2\log|A|}{r^2}}$ above to obtain

$$\mathop{\mathbb{E}}_{\sigma_{1:m}}[\sup_{a\in\mathcal{A}}\sum_{i=1}^{m}a_i.\sigma_i] \leq r\sqrt{2\log|A|}$$

∎

Corollary 22.2 *The Radamachar complexity of function class $\mathcal{G}$ is upper bounded by* $\sqrt{\frac{2\log \Pi_{\mathcal{G}}}{m}}$

Proof: *For a fixed sample $S = (z_1, z_2, ..z_m)$, $\mathcal{G}_{|S}$ is the set of vectors of function values $(g(z_1), g(z_2), ...g(z_m))$, where $g \in \mathcal{G}$. Massart's Lemma can be used to upper bound the Rachmader complexity in terms of the growth function $\Pi_{\mathcal{G}}(m)$ as:*

$$\begin{aligned}
\hat{\mathfrak{R}}_s(\mathcal{G}) &= \mathop{\mathbb{E}}_{\sigma_{1:m}}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m} g(z_i)\sigma_i\right] \quad \textit{(Sample Radamacher Complexity)} \\
&= \mathop{\mathbb{E}}_{\sigma_{1:m}}\left[\frac{1}{m}\sup_{g\in\mathcal{G}} \mathbf{g}.\sigma\right] \\
&\le \frac{1}{m} r\sqrt{2\log|G_{|s}|} \quad \textit{(Massart's Lemma)} \\
&\le \frac{1}{m}\sqrt{m}\sqrt{2\log|G_{|s}|} \quad \textit{(L2 norm on a binary set)} \\
&\le \sqrt{\frac{2\log \Pi_{\mathcal{G}}(m)}{m}} \quad \left(As\ \Pi_{\mathcal{G}}(m) = \max_{\substack{s\subseteq z \\ |s|=m}} |G_{|s}|\right)
\end{aligned}$$

Expressing Radamacher complexity in-terms of the sample Radamacher complexity

$$\begin{aligned}
\mathfrak{R}_m(\mathcal{G}) &= \mathop{\mathbb{E}}_{S}\left[\mathop{\mathbb{E}}_{\sigma_{1:m}}\left[\sup_{g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^{m} g(z_i)\sigma_i\right]\right] \\
&\le \mathop{\mathbb{E}}_{S}\left[\sqrt{\frac{2\log \Pi_{\mathcal{G}}(m)}{m}}\right] \\
&= \sqrt{\frac{2\log \Pi_{\mathcal{G}}(m)}{m}}
\end{aligned}$$

■

Comments: This corollary causes

$$\Phi(S) \le \mathfrak{R}_m(\mathcal{G}) + \sqrt{\frac{\log 1/\delta}{2m}}$$

to be replaced by

$$\Phi(S) \le \sqrt{\frac{2\log \Pi_{\mathcal{G}}(m)}{m}} + \sqrt{\frac{\log 1/\delta}{2m}}$$

The Massart's lemma allows us to replace the Rademacher complexity with a term that depends on the growth function. We will now see if this growth function can be expressed in terms of the VC-dimension.

22.3 Sauer's Lemma

In this section, we discuss the relation between VC-dimension and growth function. This quantity is often easier to compute compared to the growth function or the Rademacher complexity.

The growth function grows exponentially in the sample size m until it is lesser than the VC-dimension beyond which it increase polynomially. We are interested in bounding this polynomial nature of the growth function.

Theorem 22.3 *Let $\mathcal{G}$ be a binary function (hypothesis) class with VC-dim(G) = d, then for all $m \in \mathbb{N}$ the following inequality holds:*

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

where $\Pi_{\mathcal{G}}$ is the growth function of function class G over m samples.

Proof: Here's a sketch of the proof.

Let $M_{S,G}$ be the matrix whose unique rows are $(g(x_1), \dots, g(x_m))$ for all the functions $g \in \mathcal{G}$ for a given sample $S = \{x_1, \dots, x_m\}$.

Facts:

1.
$$\Pi_G(m) = \max_{\substack{S \in Z \\ S = m}} \# M_{S,G}$$

2. If the number of $1's$ in all the rows of $M_{S,G}$ was lesser than or equal to d, then
$$\# M_{S,G} \leq \sum_{i=0}^{d} \binom{m}{i}$$

Trick:

Modify $M_{S,G}$ such that the number of $1's$ in every row is lesser than or equal to d. This process does not lead to duplication of any row and also doesn't result in an increase in the VC-dimension.

Procedure:

```
for epoch = 1, 2, ...
    for col j = 1, 2, ..., m
        For row i = 1, 2, ...
            M_{S,G}(i,j) = 0 if this doesn't duplicate another row
```

An example of this shifting is given below.

$$\begin{pmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \rightarrow \begin{pmatrix} 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

This procedure can only reduce the number of shatterings but cannot add a new one. ■