Rademacher Complexity

Ashish Rastogi

Introduction

- In a learning task, there is a relationship between
  - complexity of the class of functions
  - learning algorithm's generalizability
- Measures of complexity of a class of functions:
  - VC dimension, VC entropy
  - Covering numbers, Fat-shattering dimensions
  - Rademacher complexity
- Typical form of a generalization bound:

$$R(h) \leq \widehat{R}(h) + f(\text{complexity of class of functions}, m)$$

(Risk) (Training error) (a function that approaches 0 as m approaches infinity)

This Lecture

- Definition of Rademacher complexity.
- Technical tool: McDiarmid's inequality.
- Generalization bounds.

3

Rademacher Complexity

- Empirical Rademacher complexity: Given a training sample $S = \{x_1, \ldots, x_m\}$, and a hypotheses set H, the "empirical Rademacher complexity" of H, is defined as:

$$\overline{R}_m(H) = \mathbb{E}_\sigma \left[\max_{h \in H} \frac{2}{m} \sum_{i=1}^{m} \sigma_i h(x_i) \right]$$

where $\sigma = (\sigma_1, \ldots, \sigma_m), \sigma_i \in \{-1, +1\}$. $h : X \mapsto [0, 1]$

- Notes:
  - sample dependent complexity measure.
  - can be computed.
  - measures how well correlated the most-correlated hypothesis is to a random labeling of points in S.

Rademacher Complexity

- Rademacher complexity (of H):

$$\overline{R}(H) = \mathbb{E}_S \left[\overline{R}_m(H) \right] .$$

5

McDiarmid's Inequality

- Theorem: Let $X_1, \ldots, X_m$ be independent random variables all taking values in the set $\mathcal{X}$. Further, let $f : \mathcal{X}^m \mapsto \mathbb{R}$ be a function of $X_1, \ldots, X_m$ that satisfies $\forall i, \forall x_1, \ldots, x_m, x_i' \in \mathcal{X}$,

$$|f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x_i', \ldots, x_m)| \leq c_i.$$

Then for all $\epsilon > 0$,

$$\Pr\left[f - \mathbb{E}[f] \geq \epsilon\right] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^m c_i^2}\right).$$

McDiarmid'89.

- Corollary: For $X_i \in [a_i, b_i]$, $f = \frac{1}{m}\sum_{i=1}^m X_i$, $c_i = \frac{b_i - a_i}{m}$.

$$\Pr\left[f - \mathbb{E}[f] \geq \epsilon\right] \leq \exp\left(\frac{-2\epsilon^2 m^2}{\sum_{i=1}^m (b_i - a_i)^2}\right).$$

Hoeffding's Inequality

Generalization Bound

- Consider the random variable:

$$\phi(S) = \sup_{h \in H} \{R(h) - \widehat{R}_S(h)\}.$$

- Let S, S' be two training samples that differ in one point.
- To apply McDiarmid's inequality, need to bound:
  - $|\phi(S) - \phi(S')|$:
    easy to show that: $|\phi(S) - \phi(S')| \leq \frac{1}{m}$.
  - $\mathbb{E}_S[\phi(S)]$: we use a "symmetrization" type step to show (next slide):
  - $$\mathbb{E}[\phi(S)] \leq \overline{R}(H).$$

7

Bound on the Expectation

$$
\begin{aligned}
\mathbb{E}_S[\phi(S)] &= \mathbb{E}_S[\sup_{h\in H}\{R(h) - \widehat{R}_S(h)\}] && \text{(by definition)}\\
&= \mathbb{E}_S[\sup_{h\in H}\{\mathbb{E}_{S'}[\widehat{R}_{S'}(h)] - \widehat{R}_S(h)\}] && \text{(writing } R(h) \text{ as an expectation)}\\
&= \mathbb{E}_S[\sup_{h\in H}\{\mathbb{E}_{S'}[\widehat{R}_{S'}(h) - \widehat{R}_S(h)]\}]\\
&\leq \mathbb{E}_{S,S'}[\sup_{h\in H}\{\widehat{R}_{S'}(h) - \widehat{R}_S(h)\}] && \text{(concavity of sup)}\\
&= \mathbb{E}_{S,S',\sigma}\left[\sup_{h\in H}\left\{\frac{1}{m}\sum_{i=1}^{m}\sigma_i(h(x'_i) - h(x_i))\right\}\right] && \text{(introducing } \sigma_i \text{)}\\
&\leq 2\mathbb{E}_{\sigma,S}\left[\frac{1}{m}\sup_{h\in H}\sum_{i=1}^{m}\sigma_i h(x_i)\right]\\
&= \overline{R}(H).
\end{aligned}
$$

8

Generalization Bound

- For all $\delta > 0$, with probability at least $1 - \delta, \forall h \in H$,

$$R(h) \leq \widehat{R}_S(h) + \overline{R}(H) + \sqrt{\frac{1}{2m} \log(1/\delta)}.$$

- Proof: From McDiarmid's inequality, we know:

$$\Pr[\phi(S) - \overline{R}(H) \geq \epsilon] \leq \exp\left(-2\epsilon^2 m\right)$$

  Setting $\delta = \exp(-2\epsilon^2 m)$, and solving, we get the result.

- Question: How to bound $\overline{R}(H)$?
  - Use McDiarmid's inequality again!
  - For a sample $S = \{x_1, \ldots, x_m\}$, define

$$\psi(S) = \overline{R}(H) - \overline{R}_m(H).$$

9

Generalization Bound

- To use McDiarmid's inequality, once again, use:
  - $\mathbb{E}[\psi(S)] = 0.$ (by definition)
  - $|\psi(S) - \psi(S')|$
- Let $S' = S \setminus \{x_j\} \cup \{x'_j\}.$

$$\begin{aligned} \psi(S) - \psi(S') &\leq \mathbb{E}_\sigma \left[\frac{2}{m} \max_{h \in H} \{\sigma_j (h(x_j) - h'(x_j))\} \right] \\ &\leq \frac{2}{m}. \end{aligned}$$

- Applying McDiarmid's inequality with these bounds yields $\forall \delta > 0,$ with probability at least $1 - \delta,$

$$\overline{R}(H) \leq \overline{R}_m(H) + \sqrt{\frac{2}{m} \ln(1/\delta)}$$

Generalization Bound

- Theorem: For all $\delta > 0$, with probability at least $1 - \delta, \forall h \in H$,

$$R(h) \leq \widehat{R}_S(h) + \overline{R}_m(H) + C\sqrt{\frac{1}{m}\ln(2/\delta)},$$

where $C = \sqrt{2} + 1/\sqrt{2}$.

Handling Classification

- Let $\mathcal{G}$ be a class of functions s.t. $\forall g \in \mathcal{G}, g : X \mapsto \{-1, +1\}$.
- Consider $l_g(x) = \dfrac{1 - y_x g(x)}{2}$. Clear that $\forall x \in X, l_g(x) \in [0, 1]$.
- Observe $l_g(x) = 1 \ (\Rightarrow)$ error and $l_g(x) = 0 \ (\Rightarrow)$ no error.
- Let $l_{\mathcal{G}} = \{l_g \mid g \in \mathcal{G}\}$.

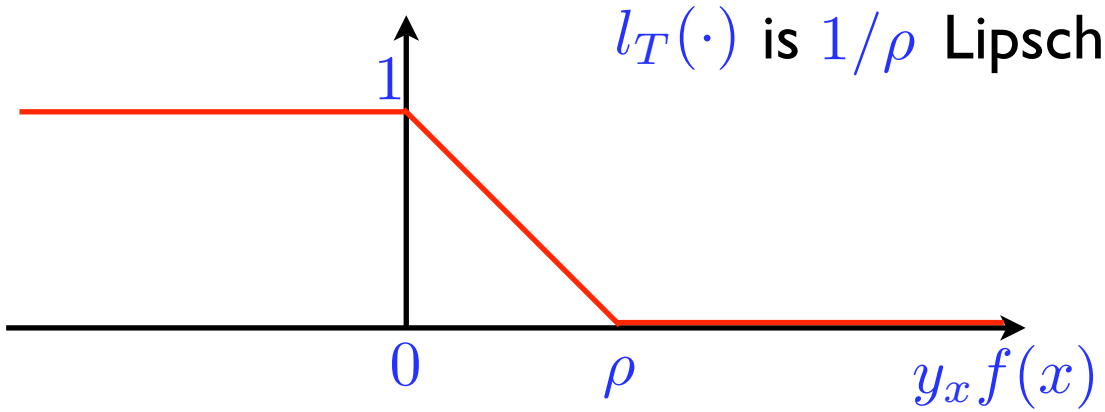
$$\begin{aligned}
\overline{R}(l_{\mathcal{G}}) &= \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \frac{2}{m} \sum_{i=1}^{m} \sigma_i \frac{1 - y_i g(x_i)}{2} \right] \\
&= \mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \frac{2}{m} \sum_{i=1}^{m} \frac{\sigma_i g(x_i)}{2} \right] \\
&= \frac{\overline{R}(\mathcal{G})}{2}
\end{aligned}$$

- Plug into previous bound.

Margin-based Bound

- Recall that the generalization bound for SVMs depends on margin (which has a geometric interpretation for SVMs).
- Another, non-geometric notion of margin:
  - Say the learning algorithm produces $f : X \mapsto \mathbb{R}$.
  - Consider the hypothesis $g \triangleq \text{sign}(f)$.
  - Define "margin" $\rho \triangleq \min_{x \in S} y_x f(x)$.
  - Truncated Hinge Loss: $l_T : X \mapsto \mathbb{R}$

$l_T(\cdot)$ is $1/\rho$ Lipschitz.



13

Margin-based Bound

- Let $\mathcal{F}$ be a family of functions, $\mathcal{F} = \{y_x f(x)\}$.
- Define training and test errors:

$$R(l_T) = \mathbb{E}_{x\sim D}\left[l_T(y_x f(x))\right], \widehat{R}(l_T) = \frac{1}{m}\sum_{i=1}^{m} l_T(y_i f(x_i)).$$

- We showed that:

$$R(l_T) \leq \widehat{R}(l_T) + \overline{R}_m(l_T(\mathcal{F})) + \sqrt{\frac{2}{m}\ln(2/\delta)}.$$

- Claim: $\overline{R}_m(l_T(\mathcal{F})) \leq \frac{2\overline{R}_m(\mathcal{F})}{\rho}$.
- Proof:
  - Talagrand's Contraction Lemma: Let $\phi : \mathbb{R} \mapsto \mathbb{R}$, such that $\phi(0) = 0$ and ϕ is λ-Lipschitz. Then,

$$\left|\mathbb{E}_\sigma\left[\frac{2}{m}\sum_{i=1}^{m}\sigma_i(\phi\circ h)(x_i)\right]\right| \leq 2\lambda\overline{R}_m(\mathcal{F}).$$

14

Rademacher Bound for Kernels

- Let $\mathcal{H}$ be a set of linear hypothesis in an RHKS.

$$\forall h \in \mathcal{H}, h : x \mapsto w \cdot \phi(x), \|w\| \leq \Omega, \phi : X \mapsto \mathbb{R}^N.$$

- How do we bound $\overline{R}_m(\mathcal{H})$?

$$\begin{aligned}
\overline{R}(\mathcal{H}) &= \mathbb{E}_\sigma \left[\sup_{\|w\| \leq \Omega} \frac{2}{m} \sum_{i=1}^{m} \sigma_i w \cdot \phi(x_i) \right] \\
&\leq \mathbb{E}_\sigma \left[\frac{2\Omega}{m} \sum_{i=1}^{m} \|\sigma_i \cdot \phi(x_i)\| \right] \\
&= \frac{2\Omega}{m} \mathbb{E}_\sigma \left[\left(\sum_{i=1}^{m} \sum_{j=1}^{m} \sigma_i \sigma_j K(x_i, x_j) \right)^{1/2} \right] \\
&= \frac{2\Omega}{m} \left(\sum_{i=1}^{m} K(x_i, x_i) \right)^{1/2}
\end{aligned}$$

- Thus, $\overline{R}_m(\mathcal{H}) \leq \dfrac{2\Omega (Tr[K])^{1/2}}{m}$

15

Rademacher Bound for Kernels

- If we assume $\|\phi(x)\| \leq R$, then $\mathrm{Tr}[K] \leq mR^2$.
- Thus, $\overline{R}_m(\mathcal{H}) \leq \dfrac{2\Omega R}{\sqrt{m}}$.

16

Rademacher Bound with VC-dim.

- For simplicity, consider the finite hypothesis set case.
- Let $f : X \mapsto [-M, M], \mathcal{F} = \{f\}, |\mathcal{F}| < \infty.$

$$\overline{R}_m(\mathcal{F}) = \mathbb{E}_\sigma \left[\max_{f \in \mathcal{F}} \frac{2}{m} \sum_{i=1}^{m} \sigma_i f(x_i) \right].$$

- Let $A = \mathbb{E}_\sigma \left[\max_{f \in \mathcal{F}} \sum_{i=1}^{m} \sigma_i f(x_i) \right].$

- Claim: $\exp(tA) \leq |\mathcal{F}| \exp(t^2 M^2 m/2)$. (proof on next slide)
- Then, $m\overline{R}_m(\mathcal{F}) \leq \frac{2 \log |\mathcal{F}|}{t} + tM^2 m.$

- Choose best t to get $\overline{R}_m(\mathcal{F}) \leq 2M \sqrt{\frac{2 \log |\mathcal{F}|}{m}}.$

Proof of Claim

$$A = \mathbb{E}_\sigma \left[\max_{f \in \mathcal{F}} \sum_{i=1}^{m} \sigma_i f(x_i) \right].$$

- Claim: $\exp(tA) \leq |\mathcal{F}| \exp(t^2 M^2 m/2)$.
- Proof:

$$\begin{aligned}
\exp(tA) &= \exp\left(t\mathbb{E}_\sigma \left[\max_{f \in \mathcal{F}} \sum_i \sigma_i f(x_i) \right] \right) \\
&\leq \mathbb{E}_\sigma \left[\exp\left(t \max_{f \in \mathcal{F}} \sum_i \sigma_i f(x_i) \right) \right] \\
&= \mathbb{E}_\sigma \left[\max_{f \in \mathcal{F}} \exp\left(t \sum_i \sigma_i f(x_i) \right) \right] \\
&\leq \mathbb{E}_\sigma \left[\sum_{f \in \mathcal{F}} \exp\left(t \sum_i \sigma_i f(x_i) \right) \right] \\
&\leq \sum_{f \in \mathcal{F}} \prod_i \mathbb{E}_{\sigma_i} \left[\exp\left(t\sigma_i f(x_i) \right) \right]
\end{aligned}$$

18

Proof of Claim

$$\begin{aligned}
\exp(tA) &\leq \sum_{f\in\mathcal{F}}\prod_{i}\mathbb{E}_{\sigma_i}\left[\exp\left(t\sigma_i f(x_i)\right)\right] \\
&\leq \sum_{f\in\mathcal{F}}\prod_{i}\exp(t^2M^2/2) \\
&= \sum_{f\in\mathcal{F}}\exp(t^2M^2m/2) \\
&\leq |\mathcal{F}|\exp(t^2M^2m/2)
\end{aligned}$$

- Thus, $A \leq \dfrac{\log|\mathcal{F}|}{t} + \dfrac{tM^2m}{2}$.
- Plugging in the best value of t, we obtain:

$$\overline{R}_m(\mathcal{F}) \leq 2M\sqrt{\frac{2\log|\mathcal{F}|}{m}}.$$

19

Rademacher Bound with VC-dim.

- If $\mathcal{F}$ has VC-dimension d, then distinct functions at most $\left(\frac{2em}{d}\right)^d$.
- Thus, $\log|\mathcal{F}| \leq d \log\left(\frac{2em}{d}\right)$.
- And so, we obtain the following bound:

$$\overline{R}_m(\mathcal{F}) \leq 2M\sqrt{\frac{2d}{m}\log\left(\frac{2em}{d}\right)}.$$

20

References

- Colin McDiarmid. *On the method of bounded differences*. In Surveys in Combinatorics, pages 148–188. Cambridge University Press, Cambridge, 1989.

- Michel Ledoux and Michel Talagrand.*Probability in Banach Spaces, Isoperimetry and Processes*. Series of Modern Surveys in Mathematics ,Vol. 23, 1991.