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

Lecture 20: VC-Dimension and Rademacher Complexity

Lecturer: Jacob Abernethy *Scribes: Jun Qi, Wanrong Zhang*

20.1 Recap Empirical Risk Minimization

Basic concept:

- Some Input space $\mathcal{X}$
- Label space $\mathcal{Y}$, prediction space $\mathcal{Y}'$
- Hypothesis class $\mathcal{H}$, $h \in \mathcal{H} : \mathcal{X} \to \mathcal{Y}'$
- Loss function $\ell : \mathcal{Y}' \times \mathcal{Y} \to \mathcal{R}$.
- Distribution $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$
- Risk of $h \in \mathcal{H}$: $R(h) = \mathbb{E}_{(x,y)\sim D}[l(h(x), y)]$
- Empirical risk for samples $\{(x_1, y_1), \ldots, (x_m, y_m)\} \subseteq \mathcal{X} \times \mathcal{Y}$: $\hat{R}_m(h) = \frac{1}{m}\sum_{i=1}^m \ell(h(x_i), y_i)$
- Empirical Risk Minimization (ERM): select $\hat{h} = \arg\min_{h\in\mathcal{H}} \hat{R}_m(h)$.

Last time, we have the following simple bound:

$$R(\hat{h}) - \min_{h^*\in\mathcal{H}} R(h^*) \leq 2 \sup |\hat{R}_m(h) - R(h)|$$

Proof:

$$\begin{aligned} R(\hat{h}) - R(h^*) &= R(\hat{h}) - \hat{R}_m(h^*) = R(\hat{h}) - \hat{R}_m(\hat{h}) + \hat{R}_m(\hat{h}) - \hat{R}_m(h^*) + \hat{R}_m(h^*) - R(h^*) \\ &\leq |R(\hat{h}) - \hat{R}_m(\hat{h})| + |\hat{R}_m(h^*) - R(h^*)| \end{aligned}$$

■

If we can show that $\sup_{h\in\mathcal{H}} |\hat{R}_m(h) - R(h)|$ has an uniform bound, then we have a bound on the estimation error. For finite $\mathcal{H}$, recall that for fixed $h \in \mathcal{H}$, we have

$$\Pr(|\mathbb{E}[\ell(h(x), y)] - \frac{1}{m}\sum_{i=1}^m \ell(h(x_i), y_i)| > t) \leq 2\exp(-2mt^2).$$

Then, by union bound, we have

$$\Pr(\exists h \in \mathcal{H} : |\mathbb{E}[\ell(h(x), y)] - \frac{1}{m}\sum_{i=1}^m \ell(h(x_i), y_i)| > t) \leq 2|\mathcal{H}|\exp(-2mt^2).$$

Thus we bound the empirical risk as following w.p. at least $1 - \delta$, $\forall h \in \mathcal{H}$

$$|R(h) - \hat{R}_m(h)| \leq \sqrt{\frac{2\log 2/\delta + \log|\mathcal{H}|}{2m}}.$$

Note that the infinite $\mathcal{H}$ case requires more careful analysis.

20.2 VC-Dimension

The VC-dimension, named after Vapnik and Chervonenkis, is a parameter that captures learnability for finite and infinite hypothesis classes.

Definition 20.1 *The VC-dimension of a hypothesis class H is given by*

$$\textit{VC-dim(H)} = \max\{d : \exists S \subseteq X, |S| = d, H \textit{ shatters } S\} \tag{20.1}$$

Claim 20.2 *Let H be a class of binary function, and $S = \{x_1, x_2, ..., x_m\}$. We say H shatters S, if $|\{(h(x_1), ..., h(x_m)) : h \in H\}| = 2^m$.*

Claim 20.3 *The VC-dimension of the linear-threshold. functions in R^d is $d + 1$.*

20.3 Rademacher Complexity

The Rademacher Complexity of a class of functions measures how rich the class is. It does so by measuring how well the class can fit random noise. In particular, it uses Rademacher random variables.

Definition 20.4 *A Rademacher Random Variable takes on values ± 1 and is defined by the Rademacher distribution*

$$\sigma_i = \begin{cases} 1, & w.p.\ \frac{1}{2} \\ -1, & w.p.\ \frac{1}{2} \end{cases} \tag{20.2}$$

Definition 20.5 *The Empirical Rademacher Complexity of a class G of functions $g : X \to R$ with respect to a sample $S = (x_1, x_2, ..., x_m)$ is*

$$\hat{R}_S := E_{\epsilon_1,...,\epsilon_m}[\sup_{h \in H} \frac{1}{m} \sum_{i=1}^{m} \epsilon_i h(x_i)] \tag{20.3}$$

where $\epsilon_1, \epsilon_2, ..., \epsilon_m$ are independent Rademacher random variables.

Definition 20.6 *Given a distribution $D \in \Delta(x)$, the Rademacher Complexity of H w.r.t. D is $E_{S \sim D}[\hat{R}_S(H)]$.*

The following theorem connects the Rademacher Complexity with the uniform deviation bounds.

Theorem 20.7 *Let G be a class of functions $g : X \to [0, 1]$ and $D \in \Delta(X)$ be a distribution. Then with probability at least $1 - \delta$, we have that*

$$\sup_{g \in G}(E[g] - \hat{E}_g[g]) \leq 2R_m(G) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}. \tag{20.4}$$

where $S \sim D^m$, $E[g] = E_{x \sim D}[g(x)]$, and $\hat{E}_S[g] = \frac{1}{m} \sum_{i=1}^{m} g(x_i)$.