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

Lecture 20: VC Dimension + Rademacher complexity

Lecturer: Jacob Abernethy *Scribes: Mihir Mavalankar*

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

20.1 Statistical Learning Theory

From Previous Lecture Recall that in a statistical learning theory setting discussed in the last lecture,

- X is the input space
- Y is the label space and Y' is the prediction space
- H is the hypothesis class where $h \in H : X \mapsto Y$
- We also have a unknown distribution $D \in \triangle(X \times Y)$
- We define risk for $h \in H$ as $R(h) = \mathrm{E}_{x,y \sim D}[l(h(x), y)]$ where $l : Y \times Y' \mapsto R$
- Given data of size n, $(x_1, y_1), ..., (x_n, y_n) \subseteq X \times Y$, we define empirical risk as $\hat{R}_m(h) = \dfrac{1}{n} \sum\limits_{i=1}^{n} l(h(x_i), y_i)$
- The empirical risk minimization algorithm (ERM) produces a hypothesis, $\hat{h} = \underset{h \in H}{\operatorname{argmin}} \hat{R}_m(h)$
- Also in the last lecture we showed a bound that we will use in today to get a bound on the estimation error, $R(\hat{h}) - R(h^*) \leq 2 \sup\limits_{h \in H} |R(\hat{h}) - \hat{R}_m(h)|$ where $h^* \in H$ and minimizes $R(.)$

20.2 Bound on Estimation Error

Claim 20.1 *If the hypothesis space H is of finite size then the estimation error is,*
$\mathrm{E}_{data}[R(\hat{h}) - \hat{R}_m(h)] = O(\sqrt{\dfrac{log(1/\delta) + log|H|}{m}})$ *with probability* $\geq 1 - \delta$ *and* $l(.) \in [0, 1]$

Proof: First we use the bound from last lecture, $R(\hat{h}) - R(h^*) \leq 2 \sup\limits_{h \in H} |R(h) - \hat{R}_m(h)| = U$

Now we try to bound the probability of U,

$$\begin{aligned} Pr(U > t) &= Pr(\exists h \in H : |R(h) - \hat{R}_m(h)| > t) \\ &\leq \sum_{h \in H} Pr(|R(h) - \hat{R}_m(h)| > t) \end{aligned} \tag{20.1}$$

Now let, $Z_i^h = l(h(x_i), y_i)$ and $\mu^h = E_{x,y \sim D}[l(h(x)), y]$ do,

$$\begin{aligned} Pr(U > t) &\leq \sum_{h \in H} Pr(|\mu^h - 1/m \sum_{i=1}^{m} Z_i^h| > t) \\ &\leq \sum_{h \in H} 2exp(-2mt^2) \qquad (By\ Hoeffding's\ inequality) \\ &= 2|H|exp(-2mt^2) = \delta \end{aligned} \tag{20.2}$$

Now we get t in terms of δ, and we get $t = \sqrt{\dfrac{log(2/\delta) + log|H|}{2m}}$
Plugging this is the last equation we get,
$\mathrm{E}_{data}[R(\hat{h}) - \hat{R}_m(h)] = O(\sqrt{\dfrac{log(1/\delta) + log|H|}{m}})$ with probability $\geq 1 - \delta$

■

20.3 Vapnik–Chervonenkis (VC) dimension

First we need take H, which is a class of binary functions

Definition 20.2 *The growth function of H is defined as,*
$\Pi_H(m) = \sup\limits_{S=x_1,..,x_m \subseteq X} |\{((h(x_1),..,h(x_m)) : h \in H\}|$
This is essentially a vector of labels

Claim 20.3 $\Pi_H(m) \leq 2^m$

Proof: (Exercise) ■

Definition 20.4 *The VC dimension of H is the largest value of d such that* $\Pi_H(d) = 2^d$
(OR Equivalently)
VC dimension is the size of the largest set of X's that can be shattered.

Claim 20.5 *For* $\mathbb{R}^d$ *the class H of binary threshold functions has VC dimension equal to d+1*

Proof: (Exercise) ■

Goal: Show that V.C. dimension characterizes the "learnability" of a function class H, which means

1. $\forall D \in (X \times Y) : \sup\limits_{h \in H} || = O(\sqrt{\dfrac{log(1/\delta) + VC(H)}{m}})$, where $VC(H)$ is the VC dimension of H
2. The equation above is tight upto log factors. This means that, $\exists D \in \triangle\ (X, Y)$ such that no algorithm guarantees that,
   $|R_m(\hat{h}) - R(h^*)| \leq \sqrt{\dfrac{VC(H)}{m}}$

Definition 20.6 *We say that a random variable X such that X=1 with probability 1/2 and X=-1 with probability 1/2 is called a Rademacher random variable.*

Definition 20.7 *Given samples* $S = X_1, ..., X_m \in X$ *the empirical Rademacher complexity of class H is,*
$\hat{\Re}_s(H) = E_{\sigma_1,..,\sigma_m}[\dfrac{1}{m} \sup\limits_{h \in H} \sum\limits_{i=1}^{m} \sigma_i h(x_i)]$

In the next few lectures we will prove that, $\hat{\Re}_s(H) \leq \sqrt{\dfrac{VC(H) log(m)}{m}}$