CS 598: Theoretical Machine Learning

Lecturer: Pranjal Awasthi
Scribe: Nathaniel Hobbs

Lecture #6
September 23, 2016

Recall from our previous class the bound on over-fitting based on the VC dimension of hypothesis function class H.

Theorem 1. *Fix $\epsilon, \delta > 0$. If $m \geq \frac{8}{\epsilon^2}(d + \ln(\frac{1}{\delta}))$ where $d = VCdim(H)$, then $\forall h \in H, \forall D$,*

$$\Pr_{s \sim D}\left(|err_s(h) - err(h)| \leq \epsilon\right) \geq 1 - \delta$$

Corollary 1. *Given same setup as above*

$$\Pr_{s \sim D}\left(|err_s(h) - err(h)| \leq \sqrt{\tfrac{8}{m}\big(d + \ln(1/\delta)\big)}\right) \geq 1 - \delta$$

The benefits of this over-fitting bound are that it (1) can apply to any learning scenario so long as the VC dimension of the hypothesis class H is known, and (2) it holds for **any** data distribution D.

However, if one has some knowledge of the problem and the type of data distribution, then one might require much fewer than the $\approx \sqrt{\frac{d}{m}}$ samples required above to avoid over-fitting. This will be the focus of today's lecture.

Rademacher Bounds

We initially defined the training error as $err(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}[h(x_i) \neq y_i]$. This is equivalent to defining the error as $err(h) = \frac{1}{2} - \frac{1}{2m}\sum_{i}^{m} y_i h(x_i)$. This implies that maximizing the second term is the same as minimizing the error, i.e. $err(h) = \text{argmax}_{h \in H} \frac{1}{m}\sum_{i}^{m} y_i h(x_i)$.

Given a space X and a fixed distribution $D|_X$, let sample $S = (x_1, \ldots, x_m)$ be a set of examples drawn i.i.d. from $D|_X$. Furthermore, let hypothesis class $\mathcal{H}$ be a class of functions $h : X \to \{+1, -1\}$.

Definition 1 (Empirical Rademacher Complexity)**.**

$$\widehat{R}_S(\mathcal{H}) = \mathbb{E}_\sigma\big[\sup_{h \in \mathcal{H}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i \cdot h(x_i)\big]$$

*where $\sigma_\mathbf{1}, \sigma_\mathbf{2}, \ldots, \sigma_\mathbf{m}$, called **Rademacher variables**, are independent random variables such that $\Pr(\sigma_i = +1) = \Pr(\sigma_i = -1) = 1/2$ for $i = 1, 2, \ldots, m$.*

Definition 2 (Rademacher Complexity)**.**

$$R_m(\mathcal{H}) = \mathbb{E}_S\big[\widehat{R}_S(\mathcal{H})\big]$$

Note that the Empirical Rademacher Complexity is defined with respect to a data set S that it is given. In contrast, the Rademacher Complexity is an expectation over **all** sets of size m. Intuitively, the $\frac{1}{m}\sum \sigma_i \cdot h(x_i)$ is behaving as if taking the dot product between two vectors in m-dimensional space, i.e a measure of their correlation or interdependence. Taking the expectation over σ, the empirical Rademacher complexity is measuring the ability of

functions from $\mathcal{H}$ (when applied to *fixed* set S) to fit random noise. The Rademacher complexity of $\mathcal{H}$ then is a measure of the expected noise-fitting-ability of $\mathcal{H}$ over all sets $S \in X^m$ that could be drawn according to distribution $D|_X$

Now we state the bound for over-fitting using the empirical Rademacher complexity as it applies to binary classification:

Theorem 2 (Over-fitting bound using Rademacher)**.** *With probability* $\geq 1-\delta$*, for all* $h \in \mathcal{H}$

$$|err(h) - err_S(h)| \leq R_{2m}(H) + 2\sqrt{\frac{\ln(1/\delta)}{m}}$$

Note that in most cases, the second term will be lower order because $\ln(1/\delta)$ is small, and dividing that by m makes it even smaller. This implies that the important term in this bound will be the Rademacher complexity of $\mathcal{H}$. For notational simplicity, we first define function $\Phi(S) = \sup_{h \in H} \big(err(h) - err_S(h)\big)$. Now the proof of this theorem will be done in two steps:

1. Bound $\Phi(S)$ with high probability in terms of its expectation
2. Bound the expectation of $\Phi(S)$ it terms of the Rademacher complexity of $\mathcal{H}$

The first part of our proof will rely on upper bounding the probability of how far the realization of a random variable can lie from its expected value. The bound we will use is called **McDiarmid's Inequality**, and it applies to functions that satisfy the **Lipschitz Condition**.

Definition 3 (Lipschitz Condition)**.** *Let* $f : \mathbb{R}^m \to \mathbb{R}$ *be a continuous function. For some constant* $c \in \mathbb{R}$*, a function is* ***c-Lipschitz*** *if* $\forall i, \forall x_1, \ldots, x_m, x_i' \in \mathbb{R}$

$$\Big|f(x_1, \ldots, x_m) - f(x_1, \ldots, x_i', \ldots, x_m)\Big| \leq c$$

Intuitively, a Lipschitz continuous function f is limited in how fast it can change, i.e. there is a definite real number, call it c, such that for every pair of points on the function, the absolute value of the slope connecting these points doesn't exceed c. Another way to characterize this is that f changes by at most c if any single variable changes within its domain.

Theorem 3 (McDiamrid's Inequality)**.** *Fix* $\epsilon > 0$ *and let* $x_1, \ldots, x_m$ *be independent random variables taking values in the set A. Furthermore, let* $f(x_1, \ldots, x_m) : X^m \to \mathbb{R}$ *be a function that is c-Lipschitz. Then*

$$\Pr\left(\Big|f(X) - \mathbb{E}[f(X)]\Big| \geq \epsilon\right) \leq 2e^{-\epsilon^2/mc^2}$$

Lemma 1 ($\Phi(S)$ is bounded by its expectation)**.** *For all* $\delta \in (0, 1)$*, with probability at least* $1 - \delta$

$$\Big|\mathbb{E}_S\big[\Phi(S)\big] - \Phi(S)\Big| \leq \sqrt{\frac{\ln(1/\delta)}{m}}$$

Proof. Recall $\Phi(S) = \sup_{h \in H} \big(err(h) - err_S(h)\big)$ and $err_S(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}[h(x_i) \neq y_i]$.

Because $h(x_i) \in \{0, 1\}$, changing any x_i to x_i' in the training set will change $err_S(h)$ by at most $\frac{1}{m}$. If we denote the change of a single variable x_i in S by the set S_i', we have $\big|\Phi(S) - \Phi(S')\big| \leq \frac{1}{m}$.

Applying McDiamrid's Inequality to $\Phi(S)$ with $c = \frac{1}{m}$, we have

$$\Pr\Big(\big|\Phi(S) - \mathbb{E}\big[\Phi(S)\big]\big| \geq \epsilon\Big) \leq 2e^{-\epsilon^2/m(\frac{1}{m})^2} \leq 2e^{-\epsilon^2 m}$$

If we wish to upper bound this by δ, it suffices to choose m such that:

$$m \geq \frac{2}{\epsilon^2}\ln(1/\delta)$$

□

Lemma 2 ($\mathbb{E}_S[\Phi(S)]$ is bounded by $R_m(\mathcal{H})$)**.** $\forall h \in H$

$$\mathbb{E}_S\big[err_S(h) - err(h)\big] \leq R_{2m}(\mathcal{H})$$

Proof. We will show this by symmetrization. We first introduce ghost sample S' drawn from the same distribution as S.

$$\begin{aligned}
\mathbb{E}_S[\Phi(S)] &= \mathbb{E}_S\Big[\sup_{h\in H}\Big(err(h) - err_S(h)\Big)\Big] \\
&= \mathbb{E}_S\Big[\sup_{h\in H}\Big(err(h) - \overbrace{\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}[h(x_i) \neq y_i]}^{\text{def of } err_S(h)}\Big)\Big] \\
&= \mathbb{E}_S\Big[\sup_{h\in H}\Big(\overbrace{\mathbb{E}_{S'}\Big[\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}[h(x_i') \neq y_i']\Big]}^{err(h) = \mathbb{E}_{S'}[err_{S'}(h)]} - \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}[h(x_i) \neq y_i]\Big)\Big] \qquad (1) \\
&= \mathbb{E}_S\Big[\sup_{h\in H}\Big(\overbrace{\mathbb{E}_{S'}\Big[\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}[h(x_i') \neq y_i'] - \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}[h(x_i) \neq y_i]\Big]}^{\text{can extend } \mathbb{E}_{S'} \text{ to include } err_S(h) \text{ b/c } S \text{ doesn't depend on } S' \text{ and LOF}}\Big)\Big] \\
&\leq \mathbb{E}_{S'}\mathbb{E}_S\Big[\sup_{h\in H}\Big(\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}[h(x_i') \neq y_i'] - \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}[h(x_i) \neq y_i]\Big)\Big]
\end{aligned}$$

Where the last inequality comes from moving the expectation over $\mathbb{E}_{S'}$ outside the *sup*. This is valid because the max of an expectation is always at most the expectation of a max.

This is a nice setup, because we have the difference in two expectations, one over the set S' and another over the set S, both being drawn from the same distribution.

From here we will just do a bookkeeping trick. We will use S and S' to create a new set $T = S \bigcup S'$, and we swap elements between S and S' with probability $\frac{1}{2}$ via Rademacher random variables. Note that the new set T is still i.i.d. from the same distribution as S

and S'.

$$\mathbb{E}_{S'}\mathbb{E}_S\Big[\sup_{h\in H}\Big(\underbrace{\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}[h(x_i')\neq y_i']}_{\text{if }\sigma=1\text{ then contributes here}} - \underbrace{\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}[h(x_i)\neq y_i]}_{\text{if }\sigma=-1\text{ then contributes here}}\Big)\Big]$$

$$= \mathop{\mathbb{E}}_{T=S\bigcup S'}\mathbb{E}_\sigma\Big[\sup_{h\in H}\Big(\frac{1}{m}\sum_{i=1}^{2m}\sigma_i\mathbb{I}[h(x_i')\neq y_i']\Big)\Big]$$

$$= \mathbb{E}_T\mathbb{E}_\sigma\Big[\sup_{h\in H}\Big(\frac{1}{m}\sum_{i=1}^{2m}\sigma_i\Big(\frac{1-y_ih(x_i)}{2}\Big)\Big)\Big]$$

$$= \underbrace{\mathbb{E}_T\mathbb{E}_\sigma\Big[\sup_{h\in H}\frac{1}{m}\sum_{i=1}^{2m}\frac{\sigma_i}{2}\Big]}_{0\text{ in expectation b/c }\sigma\in\{+1/-1\}} + \mathbb{E}_T\mathbb{E}_\sigma\Big[\sup_{h\in H}\frac{1}{m}\sum_{i=1}^{2m}\frac{\sigma_iy_ih(x_i)}{2}\Big] \tag{2}$$

$$= \mathbb{E}_T\mathbb{E}_\sigma\Big[\sup_{h\in H}\frac{1}{2m}\sum_{i=1}^{2m}\underbrace{\sigma_iyi}_{\text{same as randomly coosing +1/-1}}h(x_i)\Big]$$

$$= \mathbb{E}_T\mathbb{E}_\sigma\Big[\sup_{h\in H}\frac{1}{2m}\sum_{i=1}^{2m}\sigma_ih(x_i)\Big]$$

$$= \mathbb{E}_T(\widehat{R}_T(\mathcal{H}))$$

$$= R_{2m}(\mathcal{H})$$

□

Resources: http://www.cs.huji.ac.il/ shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf. See Chapter 26.