Machine Learning Theory (CS 6783)

Lecture 5 : Symmetrization and Infinite classes

1 Recap

Last class we showed that

$$\mathbb{E}_S\left[L_D(\hat{\mathbf{y}}_{\text{ERM}})\right] - \inf_{f\in\mathcal{F}} L_D(f) \leq \sup_D \mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left\{\mathbb{E}\left[\ell(f(x),y)\right] - \frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right\}\right]$$

This was using the Empirical Risk Minimizer (ERM)

1. When $|\mathcal{F}| < \infty$, using the above we showed that

$$\mathbb{E}_S\left[L_D(\hat{\mathbf{y}}_{\text{ERM}})\right] - \inf_{f\in\mathcal{F}} L_D(f) \leq \sqrt{\frac{\log|\mathcal{F}|}{n}}$$

2. For countably infinite class we showed MDL bound and the algorithm based on this bound.
3. However the learning rate was not uniform over $\mathcal{F}$
4. If $N(\delta)$ is the smallest number of functions that can approximate $\mathcal{F}$ to within additive factor δ uniformly over the instance space, then

$$\mathbb{E}_S\left[L_D(\hat{\mathbf{y}}_{\text{ERM}})\right] - \inf_{f\in\mathcal{F}} L_D(f) \leq \inf_{\delta>0}\left\{4\delta + \sqrt{\frac{\log N(\delta)}{n}}\right\}$$

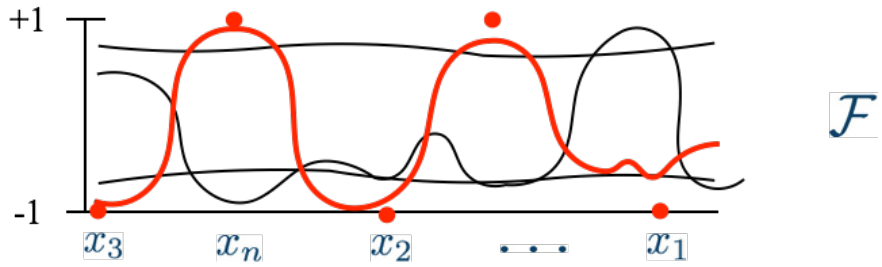
5. But $N(\delta) = \infty$ for even simple examples like threshold.

2 Symmetrization and Rademacher Complexity

$$
\begin{aligned}
&\mathbb{E}_S\left[L_D(\hat{\mathbf{y}}_{\mathrm{ERM}})\right] - \inf_{f\in\mathcal{F}} L_D(f) \\
&\qquad \le \mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left\{\mathbb{E}\left[\ell(f(x),y)\right] - \frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right\}\right] \\
&\qquad \le \mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}\ell(f(x'_t),y'_t) - \frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right\}\right] \\
&\qquad = \mathbb{E}_{S,S'}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}\epsilon_t(\ell(f(x'_t),y'_t) - \ell(f(x_t),y_t))\right\}\right] \\
&\qquad \le 2\ \mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(f(x_t),y_t)\right\}\right] \\
&\qquad =: \mathcal{R}_n(\ell\circ\mathcal{F})
\end{aligned}
$$

Where in the above each ϵ_t is a Rademacher random variable that is $+1$ with probability $1/2$ and -1 with probability $1/2$. The above is called Rademacher complexity of the loss class $\ell\circ\mathcal{F}$. In general Rademacher complexity of a function class measures how well the function class correlates with random signs. The more it can correlate with random signs the more complex the class is.

$$\text{Example}: \ \mathcal{X} = [0,1],\ \mathcal{Y} = [-1,1]$$



3 Why Does Symmetrization Help?

The main idea is that once we have introduced the Rademacher variables $\epsilon_1,\ldots,\epsilon_n$, we can look at the Rademacher complexity conditioned on sample S. Specifically, given a sample S, define

$$\mathcal{F}_{|x_1,\ldots,x_n} = \{(f(x_1),\ldots,f(x_n)) : f\in\mathcal{F}\}$$

and define

$$\mathcal{G}_{|(x_1,y_1),\ldots,(x_n,y_n)} = \{(\ell(f(x_1),y_1),\ldots,\ell(f(x_n),y_n)) : f\in\mathcal{F}\}$$

Now note that:

$$\mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(f(x_t),y_t)\right\}\right] = \mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{\mathbf{f}\in\mathcal{F}_{|x_1,\ldots,x_n}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(\mathbf{f}[t],y_t)\right]$$
$$= \mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{\mathbf{g}\in\mathcal{G}_{|(x_1,y_1),\ldots,(x_n,y_n)}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\mathbf{g}[t]\right]$$

Thus we see that we need to bound $\mathbb{E}_\epsilon\left[\sup_{\mathbf{f}\in\mathcal{F}_{|x_1,\ldots,x_n}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(\mathbf{f}[t],y_t)\right]$ or equivaently $\mathbb{E}_\epsilon\left[\sup_{\mathbf{g}\in\mathcal{G}_{|(x_1,y_1),\ldots,(x_n,y_n)}}\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\mathbf{g}[t]\right]$. Since the term instead the supremum is still a zero mean average, it is clear that only the cardinality of set $\mathcal{F}_{|x_1,\ldots,x_n}$ matters and not cardinality of all of $\mathcal{F}$. Why does this help?

Think about the threshold example, given n examples, the cardinality restricted to these samples is at most $n+1$. Why?
Well sort any given n points in ascending order, using thresholds, we can get at most n + 1 possible labeling on the n points. Hence for any $x_1,\ldots,x_n$, $|\mathcal{F}_{|x_1,\ldots,x_n}| \leq n+1$

If we use the intuition that max over a finite set (of cardinality say M) of average over n zero mean bounded variables is at most $O(\sqrt{\log M/n})$, then we see that this implies a rate of $O(\sqrt{\log(n+1)/n}$ for learning thresholds. To make the argument concrete, we below reprove the lemma for expected supremum over Rademacher averages over finite sets precisely. From the lemma we prove next we immediately conclude that more generally:

$$\mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}\epsilon_t\ell(f(x_t),y_t)\right\}\right] \leq O\left(\mathbb{E}_S\left[\sqrt{\frac{\log|\mathcal{F}_{|x_1,\ldots,x_n}|}{n}}\right]\right)$$

4 Massart's Finite Lemma

Lemma 1. *For any set $V \subset \mathbb{R}^n$:*

$$\frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{\mathbf{v}\in V}\sum_{t=1}^{n}\epsilon_t\mathbf{v}[t]\right] \leq \frac{1}{n}\sqrt{2\left(\sup_{\mathbf{v}\in V}\sum_{t=1}^{n}\mathbf{v}^2[t]\right)\log|V|}$$

Proof.

$$\sup_{\mathbf{v}\in V}\sum_{t=1}^{n}\epsilon_t\mathbf{v}[t] = \frac{1}{\lambda}\log\left(\sup_{\mathbf{v}\in V}\exp\left(\lambda\sum_{t=1}^{n}\epsilon_t\mathbf{v}[t]\right)\right)$$
$$\leq \frac{1}{\lambda}\log\left(\sum_{\mathbf{v}\in V}\exp\left(\lambda\sum_{t=1}^{n}\epsilon_t\mathbf{v}[t]\right)\right)$$
$$= \frac{1}{\lambda}\log\left(\sum_{\mathbf{v}\in V}\prod_{t=1}^{n}\exp\left(\lambda\epsilon_t\mathbf{v}[t]\right)\right)$$

Taking expectation w.r.t. Rademacher random variables,

$$\mathbb{E}_\epsilon\left[\sup_{\mathbf{v}\in V}\sum_{t=1}^{n}\epsilon_t\mathbf{v}[t]\right] \leq \frac{1}{\lambda}\mathbb{E}_\epsilon\left[\log\left(\sum_{\mathbf{v}\in V}\prod_{t=1}^{n}\exp\left(\lambda\epsilon_t\mathbf{v}[t]\right)\right)\right]$$

Since log is a concave function, by Jensen's inequality, Expected log is upper bounded by log of expectation and so:

$$\begin{aligned}
&\leq \frac{1}{\lambda}\log\left(\mathbb{E}_\epsilon\left[\sum_{\mathbf{v}\in V}\prod_{t=1}^{n}\exp\left(\lambda\epsilon_t\mathbf{v}[t]\right)\right]\right)\\
&= \frac{1}{\lambda}\log\left(\sum_{\mathbf{v}\in V}\prod_{t=1}^{n}\mathbb{E}_{\epsilon_t}\left[\exp\left(\lambda\epsilon_t\mathbf{v}[t]\right)\right]\right)\\
&= \frac{1}{\lambda}\log\left(\sum_{\mathbf{v}\in V}\prod_{t=1}^{n}\frac{e^{\lambda\mathbf{v}[t]}+e^{-\lambda\mathbf{v}[t]}}{2}\right)
\end{aligned}$$

For any x, $\frac{e^x+e^{-x}}{2} \leq e^{x^2/2}$

$$\begin{aligned}
&\leq \frac{1}{\lambda}\log\left(\sum_{\mathbf{v}\in V}e^{\lambda^2\sum_{t=1}^{n}\mathbf{v}^2[t]/2}\right)\\
&\leq \frac{1}{\lambda}\log\left(|V|e^{\lambda^2\sup_{\mathbf{v}\in V}\left(\sum_{t=1}^{n}\mathbf{v}^2[t]\right)/2}\right)\\
&= \frac{\log|V|}{\lambda} + \frac{\lambda\sup_{\mathbf{v}\in V}\left(\sum_{t=1}^{n}\mathbf{v}^2[t]\right)}{2}
\end{aligned}$$

Choosing $\lambda = \sqrt{\frac{2\log|V|}{\sup_{\mathbf{v}\in V}\left(\sum_{t=1}^{n}\mathbf{v}^2[t]\right)}}$ completes the proof. □