Machine Learning Theory (CSC 482A/581A) - Lecture 11

Nishant Mehta

1 Uniform convergence in the realizable case

We already have seen that agnostically learning is possible when $\mathcal{F}$ is a VC class, and the excess risk obtainable via ERM converges (ignoring logarithmic factors) at the rate

$$O\left(\sqrt{\frac{\mathrm{VCdim}(\mathcal{F}) + \log\frac{1}{\delta}}{n}}\right).$$

However, in the PAC learning (i.e. realizable) setting, at least for finite classes we were able to obtain a better convergence rate in that the rate did not have a square root. The same is true for VC classes, as we will now see.

Theorem 1 (Vapnik and Chervonenkis (1971))**.** *Let $\mathcal{F} \subset \{0,1\}^{\mathcal{X}}$ be a VC class with $\mathrm{VCdim}(\mathcal{F}) = V$, and let $\hat{f}$ be an ERM classifier (which, given a training sample, outputs a hypothesis in $\mathcal{F}$ that minimizes the empirical risk), and let P be an arbitrary probability distribution P over $\mathcal{X} \times Y$ that satisfies $Y = c(X)$ for some $c \in \mathcal{F}$.*

Then for any $n \geq V$, and any $\varepsilon > 0$.

$$\Pr\left(R(\hat{f}) > \varepsilon\right) \leq 2\left(\frac{2en}{V}\right)^V e^{-n\varepsilon/2}.$$

Equivalently, for any $n \geq V$, with probability at least $1-\delta$

$$R(\hat{f}) \leq \frac{2\left(V\log\frac{2en}{V} + \log\frac{2}{\delta}\right)}{n}.$$

Before proving this result, observe that we can reframe our goal in terms of the convergence of the empirical risk $\hat{R}(\hat{f})$ of ERM to its true risk. Since $R(\hat{f}) = R(\hat{f}) - \hat{R}(\hat{f})$ for ERM, it follows that a high probability bound on $\left|R(\hat{f}) - \hat{R}(\hat{f})\right|$ is exactly equivalent to a high probability bound on the risk $R(\hat{f})$.

The proof of Theorem 1 relies on the above observation and two lemmas. As before, we use the notation that $Z = (X, Y)$ (likewise for $Z_j = (X_j, Y_j)$ and $Z'_j = (X'_j, Y'_j)$). We again introduce a ghost sample $Z'_1, \ldots, Z'_n$; recall that each $Z_j = (X_j, Y_j)$ is a labeled sample drawn from probability distribution P.

Lemma 1. *If $n\varepsilon > 2$, then*

$$\Pr\left(\sup_{f\in\mathcal{F}:\ \hat{R}(f)=0}\left|\frac{1}{n}\sum_{j=1}^n \ell_f(Z_j) - \mathsf{E}_{Z\sim P}[\ell_f(Z)]\right| > \varepsilon\right)$$
$$\leq 2\Pr\left(\sup_{f\in\mathcal{F}:\ \hat{R}(f)=0}\left|\frac{1}{n}\sum_{j=1}^n \ell_f(Z_j) - \frac{1}{n}\sum_{j=1}^n \ell_f(Z'_j)\right| > \varepsilon/2\right).$$

We won't cover the proof of this result. However, the high-level argument is similar the one we used for the general (agnostic) case.

Lemma 2. *It holds that*

$$\Pr\left(\sup_{f\in\mathcal{F}\colon \hat{R}(f)=0}\left|\frac{1}{n}\sum_{j=1}^{n}\ell_f(Z_j)-\frac{1}{n}\sum_{j=1}^{n}\ell_f(Z'_j)\right|>\varepsilon/2\right)\leq \Pi_{\mathcal{F}}(2n)2^{-n\varepsilon/2}.$$

Proof. Let $\pi(Z_1),\ldots,\pi(Z_n),\pi(Z'_1),\ldots,\pi(Z'_n)$ be an arbitrary permutation of $Z_1,\ldots,Z_n,Z'_1,\ldots,Z'_n$. Observe that from the i.i.d. property of the double sample, the distribution of the random variable

$$\sup_{f\in\mathcal{F}\colon \sum_{j=1}^{n}\ell_f(Z_j)=0}\left|\frac{1}{n}\sum_{j=1}^{n}\ell_f(Z_j)-\frac{1}{n}\sum_{j=1}^{n}\ell_f(Z'_j)\right|$$

is equal to the distribution of the random variable

$$\sup_{f\in\mathcal{F}\colon \sum_{j=1}^{n}\ell_f(\pi(Z_j))=0}\left|\frac{1}{n}\sum_{j=1}^{n}\ell_f(\pi(Z_j))-\frac{1}{n}\sum_{j=1}^{n}\ell_f(\pi(Z'_j))\right|$$

Let $U(S_{2n})$ be the uniform distribution over the symmetric group S_{2n}, the set of all permutations over $2n$ items. It therefore holds that

$$\begin{aligned}
&\Pr\left(\sup_{f\in\mathcal{F}\colon \hat{R}(f)=0}\left|\frac{1}{n}\sum_{j=1}^{n}\ell_f(Z_j)-\frac{1}{n}\sum_{j=1}^{n}\ell_f(Z'_j)\right|>\varepsilon/2\right)\\
&=\mathsf{E}\left[\sup_{f\in\mathcal{F}\colon \hat{R}(f)=0}\mathbf{1}\left[\left|\frac{1}{n}\sum_{j=1}^{n}\ell_f(Z_j)-\frac{1}{n}\sum_{j=1}^{n}\ell_f(Z'_j)\right|>\varepsilon/2\right]\right]\\
&=\mathsf{E}\left[\mathsf{E}_{\pi\sim U(S_{2n})}\left[\sup_{f\in\mathcal{F}\colon \sum_{j=1}^{n}\ell_f(\pi(Z_j))=0}\mathbf{1}\left[\left|\frac{1}{n}\sum_{j=1}^{n}\ell_f(\pi(Z_j))-\frac{1}{n}\sum_{j=1}^{n}\ell_f(\pi(Z'_j))\right|>\varepsilon/2\right]\right]\right].
\end{aligned}$$

We will get a small upper bound just for the internal expectation over π. For a fixed double sample, let $\mathcal{F}_{2n}\subset\mathcal{F}$ be a class which, for each labeling of $X_1,\ldots,X_n,X'_1,\ldots,X'_n$ attainable by a hypothesis in $\mathcal{F}$, contains precisely one representative from f that obtains this labeling. Then the conditional expectation above (conditional on the double sample) is equal to

$$\mathsf{E}_{\pi\sim U(S_{2n})}\left[\sup_{f\in\mathcal{F}_{2n}\colon \sum_{j=1}^{n}\ell_f(\pi(Z_j))=0}\mathbf{1}\left[\left|\frac{1}{n}\sum_{j=1}^{n}\ell_f(\pi(Z_j))-\frac{1}{n}\sum_{j=1}^{n}\ell_f(\pi(Z'_j))\right|>\varepsilon/2\right]\right],$$

which is at most

$$\begin{aligned}
&\mathsf{E}_{\pi\sim U(S_{2n})}\left[\sum_{f\in\mathcal{F}_{2n}\colon \sum_{j=1}^{n}\ell_f(\pi(Z_j))=0}\mathbf{1}\left[\left|\frac{1}{n}\sum_{j=1}^{n}\ell_f(\pi(Z_j))-\frac{1}{n}\sum_{j=1}^{n}\ell_f(\pi(Z'_j))\right|>\varepsilon/2\right]\right]\\
&=\mathsf{E}_{\pi\sim U(S_{2n})}\left[\sum_{f\in\mathcal{F}_{2n}}\mathbf{1}\left[\sum_{j=1}^{n}\ell_f(\pi(Z_j))=0\right]\mathbf{1}\left[\left|\frac{1}{n}\sum_{j=1}^{n}\ell_f(\pi(Z_j))-\frac{1}{n}\sum_{j=1}^{n}\ell_f(\pi(Z'_j))\right|>\varepsilon/2\right]\right]\\
&=\sum_{f\in\mathcal{F}_{2n}}\Pr_{\pi\sim U(S_{2n})}\left(\sum_{j=1}^{n}\ell_f(\pi(Z_j))=0\ \bigwedge\ \left|\frac{1}{n}\sum_{j=1}^{n}\ell_f(\pi(Z_j))-\frac{1}{n}\sum_{j=1}^{n}\ell_f(\pi(Z'_j))\right|>\varepsilon/2\right).
\end{aligned}$$

Now, suppose that there are at least $r = n\varepsilon/2$ mistakes among $2n$ points. How many permutations are there in which no mistakes occur in the first half of the permuted double sample? There are $n(n-1)\cdots(n-r+1)$ ways to arrange the r mistake points in the second half, and $(2n-r)(2n-r-1)\cdots 1$ ways to arrange the remaining points thereafter. On the other hand, if we are unrestricted in where the mistakes are placed, then the first product is $2n(2n-1)\cdots(2n-r+1)$. Therefore, the fraction of the permutations where no mistakes occur in the first half is at most

$$\frac{n}{2n}\frac{n-1}{2n-1}\cdots\frac{n-r+1}{2n-r+1} \le 2^{-r} \le 2^{-n\varepsilon/2}.$$

Therefore,

$$\begin{aligned}\Pr\left(\sup_{f\in\mathcal{F}:\,\hat{R}(f)=0}\left|\frac{1}{n}\sum_{j=1}^{n}\ell_f(Z_j)-\frac{1}{n}\sum_{j=1}^{n}\ell_f(Z'_j)\right|>\varepsilon/2\right) &\le \mathsf{E}\left[|\mathcal{F}_{2n}|2^{-n\varepsilon/2}\right]\\ &\le \Pi_{\mathcal{F}}(2n)2^{-n\varepsilon/2}.\end{aligned}$$

□

References

Vladimir N. Vapnik and Alexey Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.