Machine Learning Theory
Lecture 6

Nicholas Harvey

September 25, 2018

1 The Fundamental Theorem of Statistical Learning

Theorem 1.1. Let $\mathcal{H}$ be a hypothesis class that can be infinite, but has $\mathrm{VCdim}(\mathcal{H}) = d$. Suppose m is sufficiently large as a function of d, ϵ and δ. If the training data S consists of m i.i.d. samples from $\mathcal{D}$, then

$$\Pr_S\left[\, |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon \;\forall h \in \mathcal{H}\,\right] \;\geq\; 1-\delta.$$

This is probably the most difficult theorem that we will do, at least in terms of probability theory. It uses several tricks. For people who work in high-dimensional probability, many of these tricks are standard. But many of you may be seeing them for the first time. A great place at UBC to learn these and other techniques is Yaniv Plan's graduate probability class (MATH 608D).

Tricks:

1. Let A_1, A_2 be random variables (not independent). Then $\max\{\mathrm{E}[A_1], \mathrm{E}[A_2]\} \leq \mathrm{E}[\max\{A_1, A_2\}]$. More generally, $\max\{\mathrm{E}[A_1], ..., \mathrm{E}[A_n]\} \leq \mathrm{E}[\max\{A_1, ..., A_n\}]$.

   This is a special case of Jensen's inequality: $f(\mathrm{E}[X]) \leq \mathrm{E}[f(X)]$ for any random vector $X \in \mathbb{R}^n$ and any convex function $f : \mathbb{R}^n \to \mathbb{R}$.

2. Switching back and forth between probability and expectation as convenient. Some things are more convenient with probabilities (Hoeffding). Other things are more convenient with expectations (like the previous trick). Specifically:

   Fact 1.2. Suppose Z is a non-negative random variable. If $\mathrm{E}[Z] \leq \epsilon\delta$ then $\Pr[Z \geq \epsilon] \leq \delta$.

   Fact 1.3. Suppose Z is a random variable that never exceeds 1. If $\Pr[Z \geq \alpha] \leq \alpha$ then $\mathrm{E}[Z] \leq 2\alpha$.

3. Symmetrization. This a bit hard to explain up front, but you will see it in action below.

Proof. To keep things simple, let's drop the absolute value. We want to show that:

$$\Pr_S\left[\max_{h\in\mathcal{H}}\big(L_S(h) - L_{\mathcal{D}}(h)\big) > \epsilon\right] \;\leq\; \delta.$$

Using a notational trick, let's rewrite that maximum as

$$\max_{h\in\mathcal{H}} \mathrm{E}_{x\sim\mathcal{D}} \left[\, L_S(h) - L_{\{x\}}(h) \,\right].$$

All we have done here is rewrite the true loss $L_{\mathcal{D}}(h)$ as $\mathrm{E}_{x\sim\mathcal{D}} \left[\, L_{\{x\}}(h) \,\right]$, where x is the test point (independent from S). Now we can employ Trick 1: for any fixed S,

$$\max_{h\in\mathcal{H}} \mathrm{E}_{x\sim\mathcal{D}} \left[\, L_S(h) - L_{\{x\}}(h) \,\right] \;\leq\; \mathrm{E}_{x\sim\mathcal{D}} \left[\max_{h\in\mathcal{H}} \big(L_S(h) - L_{\{x\}}(h) \big) \right].$$

Thus, reintroducing the probability with respect to S:

$$\Pr_S \left[\max_{h\in\mathcal{H}} \mathrm{E}_{x\sim\mathcal{D}} \left[\, L_S(h) - L_{\{x\}}(h) \,\right] > \epsilon \right] \;\leq\; \Pr_S \left[\underbrace{\mathrm{E}_x \left[\max_{h\in\mathcal{H}} \big(L_S(h) - L_{\{x\}}(h) \big) \right]}_{Z} > \epsilon \right].$$

Here comes Trick 2. If we took expectation over S rather than probability over S, then we would have a joint expectation with respect to S and x, which seems convenient. Fact 1.2 allows us to accomplish this. So our new goal is to prove

$$\mathrm{E}_{S,x} \left[\max_{h\in\mathcal{H}} \big(L_S(h) - L_{\{x\}}(h) \big) \right] \;\leq\; \epsilon\delta.$$

How to analyze $\max_h \big(L_S(h) - L_{\{x\}}(h) \big)$? In an earlier lecture we have seen how this can be done with a Hoeffding bound and union bound.

Trouble #1. In order to carry out that plan, we want to prove an exponentially small tail bound on $L_S(h) - L_{\{x\}}(h)$ using Hoeffding's inequality. Fortunately $L_S(h)$ is very concentrated because it's an average of m independent samples. Unfortunately $L_{\{x\}}(h)$ is not concentrated because x is just a single sample. For example, it could be 0 or 1 each with probability 1/2.

Fix #1. An easy way to make $L_{\{x\}}(h)$ more concentrated is to use more test points. Simply replace x with another set S' of m i.i.d. samples, and replace $L_{\{x\}}(h)$ with $L_{S'}(h)$. So now our goal is to show

$$\mathrm{E}_{S,S'} \left[\max_{h\in\mathcal{H}} \big(L_S(h) - L_{S'}(h) \big) \right] \;\leq\; \epsilon\delta. \tag{1.1}$$

Main Idea of Proof. Here is where we can use the VC-dimension of $\mathcal{H}$. Notice that h is only going to be evaluated on the points $S \cup S'$. So instead of taking the max over all hypotheses in $\mathcal{H}$, it seems we can just consider all labelings in $\mathcal{H}_{S\cup S'}$. This is a good idea, but actually carrying it out requires some care.

For example, when doing a union bound, the index set of the union bound is typically not random! One could condition on an event so that the index set becomes fixed, but then one must ensure that the conditional distribution still allows the Hoeffding bound to be used.

A natural thing to try would be to condition on the event $S \cup S' = C$, for an arbitrary set C, because then we only need to consider hypotheses in $\mathcal{H}_C$. So it would suffice to show

$$\mathrm{E}_{S,S'} \left[\max_{h\in\mathcal{H}_C} \big(L_S(h) - L_{S'}(h) \big) \;\big|\; S \cup S' = C \right] \;\leq\; \epsilon\delta.$$

Trouble #2. Unfortunately, after conditioning on $S \cup S' = C$, the distribution on S and S' is quite messy: the samples are no longer i.i.d., so Hoeffding cannot be applied. (For example, if you drew S and S' by sampling independently from C, then they would be unlikely to satisfy $S \cup S' = C$.) We need a new approach: how can we condition on an event so that $S \cup S' = C$ but the conditional distribution involves independent samples.

Fix #2. (This is basically Trick 3, symmetrization.) Here is a clever idea: suppose that the event ensures that each point in C belongs to one of S or S' *but does not reveal which it came from.* Formally, suppose that the training points in S are $(x_1, ..., x_m)$, and the test points in S' are $(x'_1, ..., x'_m)$. Then we can condition on the event

$$\mathcal{E}_C \;:=\; \bigwedge_{i=1}^{m} \{x_i, x'_i\} = \{c_i, c'_i\}$$

for any points $C = \{c_1, c'_1\}, ..., \{c_m, c'_m\}$. This event has some nice properties. Importantly, it does not determine whether $x_i = c_i$ or $x_i = c'_i$ (assuming $c_i \neq c'_i$). Furthermore, since S and S' are independent, the (conditional) probability that $x_i = c_i$ is $1/2$. Lastly, the events $x_i = c_i$ and $x_j = c_j$ are independent for $i \neq j$. Thus, we are in a good position to apply a Hoeffding bound.

Using the definition of training error, we have

$$L_S(h) - L_{S'}(h) \;=\; \frac{1}{m}\sum_{i=1}^{m}\Big(1_{h(x_i)\neq f(x_i)} - \sum_{i=1}^{m} 1_{h(x'_i)\neq f(x'_i)}\Big). \tag{1.2}$$

Then determining whether $\begin{bmatrix} x_i \\ x'_i \end{bmatrix} = \begin{bmatrix} c_i \\ c'_i \end{bmatrix}$ or $\begin{bmatrix} x_i \\ x'_i \end{bmatrix} = \begin{bmatrix} c'_i \\ c_i \end{bmatrix}$ just determines the sign of the i^{th} summand in (1.2). More specifically let $\sigma_i = +1$ if $x_i = c_i$, and otherwise $\sigma_i = -1$. Let

$$\begin{aligned} X_i &\;=\; \sigma_i\Big(1_{h(c_i)\neq f(c_i)} - 1_{h(c'_i)\neq f(c'_i)}\Big) \\ X &\;=\; \sum_{i=1}^{m} X_i/m. \end{aligned}$$

The main consequence of this definition is that

$$X \;=\; L_S(h) - L_{S'}(h)$$

Furthermore, the X_i are independent (conditional on $\mathcal{E}_C$), and have zero mean (again, conditional on $\mathcal{E}_C$). Thus, by a Hoeffding bound,

$$\mathrm{Pr}_{S,S'}\left[\, L_S(h) - L_{S'}(h) > \alpha \;\mid\; \mathcal{E}_C \,\right] \;=\; \mathrm{Pr}\left[\, X > \alpha \;\mid\; \mathcal{E}_C \,\right] \;\leq\; \exp(-\alpha^2 m/2).$$

If the number of samples m is at least $\ln(|\mathcal{H}_C|/\alpha)/\alpha^2$ and $\alpha = \epsilon\delta/2$, a union bound gives that

$$\mathrm{Pr}_{S,S'}\left[\max_{h\in\mathcal{H}_C}\big(L_S(h) - L_{S'}(h)\big) > \alpha \;\mid\; \mathcal{E}_C\right] \;\leq\; \sum_{h\in\mathcal{H}_C} \mathrm{Pr}_{S,S'}\left[\, L_S(h) - L_{S'}(h) > \alpha \;\mid\; \mathcal{E}_C \,\right] \;\leq\; \alpha.$$

Thus,

$$\begin{aligned}
\mathrm{E}_{S,S'}\left[\max_{h\in\mathcal{H}}\big(L_S(h)-L_{S'}(h)\big)\right] &= \sum_C \Pr\left[\,\mathcal{E}_C\,\right]\cdot \mathrm{E}_{S,S'}\left[\max_{h\in\mathcal{H}}\big(L_S(h)-L_{S'}(h)\big)\;\mid\;\mathcal{E}_C\right]\\
&= \sum_C \Pr\left[\,\mathcal{E}_C\,\right]\cdot \mathrm{E}_{S,S'}\left[\max_{h\in\mathcal{H}_C}\big(L_S(h)-L_{S'}(h)\big)\;\mid\;\mathcal{E}_C\right]\\
&\le \sum_C \Pr\left[\,\mathcal{E}_C\,\right]\cdot(2\alpha) \qquad \text{(by Fact 1.3)}\\
&= \epsilon\delta,
\end{aligned}$$

by definition of α. This proves (1.1).

One remaining detail is that we haven't fully determined the number of samples. Observe that $|\mathcal{C}| \le 2m$. By the Sauer-Shelah lemma, $|\mathcal{H}_C| \le cm^d$ for some constant c. Thus it suffices to have $m \ge d\ln(cm/\alpha)/\alpha^2$. This is satisfied by $m \ge cd\log(d/\alpha)/\alpha^2$, for some constant c. ■