ECE 543: Statistical Learning Theory

Maxim Raginsky

April 15, 2021

Homework 4

Assigned April 15; due April 27, 2021

Note: natural logarithms are used throughout.

1. **Intrinsic limitations of learning.** In our analysis of regression with quadratic loss, we have focused on the ERM algorithm and developed high-probability bounds on its excess loss. In this problem, we will see that there are certain intrinsic limitations any learning algorithm will face even in the realizable case when $Y = f(X)$ (with probability one) and the function f is a member of the chosen hypothesis class $\mathcal{F}$.

   Let μ be the marginal probability distribution of X, and for each $f \in \mathcal{F}$ let $Y^f = f(X)$. Let $\mathbf{P}_f$ denote the joint distribution of (X, Y^f). That is, under $\mathbf{P}^f$ we have

   $$\mathbf{P}_f(A \times B) = \int_A \mu(\mathrm{d}x) \mathbf{1}_{\{f(x) \in B\}}$$

   for all measurable sets $A \subset \mathsf{X}$ and all $B \subset \mathbb{R}$. Consider a learning algorithm $\mathcal{A}_n$ that receives a sequence of i.i.d. training samples $Z_i^f = (X_i, Y_i^f)$, $1 \le i \le n$, drawn from $\mathbf{P}_f$, where $f \in \mathcal{F}$ is unknown. Consider also the following *random* subset of $\mathcal{F}$:

   $$\mathcal{V}_n(f) := \left\{ h \in \mathcal{F} : h(X_i) = f(X_i),\, 1 \le i \le n \right\}.$$

   This set, called the *version space*, consists of all functions $h \in \mathcal{F}$ that agree with the unknown target function f on the training data. Let $D_n(f)$ denote the *diameter* of the version space in $L^2(\mu)$ norm:

   $$D_n(f) := \sup_{h,h' \in \mathcal{V}_n(f)} \|h - h'\|_{L^2(\mu)} \equiv \sup_{h,h' \in \mathcal{V}_n(f)} \left(\int_{\mathsf{X}} \left| h(x) - h'(x) \right|^2 \mu(\mathrm{d}x) \right)^{1/2}.$$

   Note that $D_n(f)$ is a random variable, since it depends on the training data. Our goal is to prove that, no matter how sophisticated $\mathcal{A}_n$ is, it cannot attain better performance than a constant multiple of $D_n^2(f)$.

   (a) Suppose that $\mathcal{A}_n$ is the ERM algorithm: upon receiving the training data $Z^n = (Z_1, \ldots, Z_n)$ with $Z_i = (X_i, Y_i)$, $1 \le i \le n$, it outputs

   $$\widehat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (Y_i - f(X_i))^2.$$

Prove that if Z^n are i.i.d. samples from $\mathbf{P}_{f^*}$ for some $f^* \in \mathcal{F}$, then

$$L(\widehat{f}_n) \equiv \int_{\mathsf{X}} \left(\widehat{f}_n(x) - f^*(x)\right)^2 \mu(\mathrm{d}x) \le D_n^2(f^*).$$

(b) Now we will prove the following converse result: for an arbitrary learning algorithm $\mathcal{A}_n$, there exists at least one $f \in \mathcal{F}$, such that

$$\mathbf{P}_f^n \left(L(\tilde{f}_n) \ge \frac{D_n^2(f)}{16} \right) \ge \frac{1}{2}, \tag{1}$$

where $\tilde{f}_n = \mathcal{A}_n\left(Z^{f,n}\right)$ is the output of $\mathcal{A}_n$ given training data $Z^n = Z^{f,n}$ drawn i.i.d. from $\mathbf{P}_f$. We will prove this in several steps.

i. Given $f \in \mathcal{F}$, consider the version space $\mathcal{V}_n(f)$ and let $h_{0,f}, h_{1,f} \in \mathcal{V}_n(f)$ be such that $\|h_{0,f} - h_{1,f}\|_{L^2(\mu)} = D_n(f)$. Let ε be a Bernoulli(1/2) random variable independent of X^n, and define the random function

$$h_f := (1-\varepsilon)h_{0,f} + \varepsilon h_{1,f}.$$

That is, if $\varepsilon = 0$, then $h_f = h_{0,f}$; if $\varepsilon = 1$, then $h_f = h_{1,f}$. Prove that, for any realization of ε, $D_n(f) = D_n(h_f)$.

ii. Prove that, for any realization of ε,

$$\sup_{f \in \mathcal{F}} \mathbf{P}_f^n \left(\left\| \mathcal{A}_n(Z^{n,f}) - f \right\|_{L^2(\mu)} \ge \frac{D_n(f)}{4} \right) \ge \sup_{f \in \mathcal{F}} \mathbf{P}_f^n \left(\left\| \mathcal{A}_n(Z^{n,h_f}) - h_f \right\|_{L^2(\mu)} \ge \frac{D_n(f)}{4} \right). \tag{2}$$

iii. Let Π_n denote the quantity on the right-hand side of (2). Note that Π_n is a random variable that depends on ε. Prove that

$$\mathbf{E}_\varepsilon \Pi_n \ge \frac{1}{2} \sup_{f \in \mathcal{F}} \left(\mu^n(A_{0,f}) + \mu^n(A_{1,f}) \right), \tag{3}$$

where, for $b \in \{0, 1\}$, we have defined the event

$$A_{b,f} := \left\{ \left\| \mathcal{A}_n(Z^{n,h_{b,f}}) - h_{b,f} \right\|_{L^2(\mu)} \ge \frac{D_n(f)}{4} \right\}.$$

iv. Prove that the union of the events $A_{0,f}$ and $A_{1,f}$ occurs with μ-probability one, and conclude from this and from (3) that $\mathbf{E}_\varepsilon \Pi_n \ge 1/2$.

Hint: Use the fact $\|h_{0,f} - h_{1,f}\|_{L^2(\mu)} = D_n(f)$, and that both $h_{0,f}$ and $h_{1,f}$ are in the version space $\mathcal{V}_n$, and therefore the function output by the learning algorithm $\mathcal{A}_n$ upon seeing the training data

$$(X_1, h_{0,f}(X_1)), \dots, (X_n, h_{0,f}(X_n))$$

is the same as the function output by $\mathcal{A}_n$ upon seeing the training data

$$(X_1, h_{1,f}(X_1)), \dots, (X_n, h_{1,f}(X_n))$$

with the *same* i.i.d. input sequence $X_1, \dots, X_n \sim \mu$.

v. Finally, use all of the above to prove that there exists at least one $f \in \mathcal{F}$, such that (1) holds true.

The moral of the story is: even if there is no noise in the data, the best performance of any learning algorithm is controlled by the richness of the function class $\mathcal{F}$. In particular, if $\mathcal{F}$ is very rich, the version space is likely to be large (as measured by the $L^2(\mu)$ norm) because there will be many functions that can match the target function on a given sample. This limitation is there even if we design our algorithm with full knowledge that the target function f is in our hypothesis class, and even if we know the marginal distribution μ of X ahead of time.

2. **Amplifying weak learning algorithms.** Let $\mathcal{F}$ be a class of functions from some space Z into $[0,1]$. Let a learning algorithm A be given with the following property: for any $\varepsilon > 0$, there exists $n(\varepsilon) \in \mathbb{N}$, such that, for any probability distribution P on Z,

$$\mathbf{E}[L(A(Z^n))] \le \inf_{f \in \mathcal{F}} L(f) + \varepsilon$$

for all $n \ge n(\varepsilon)$. Here, $A(Z^n)$ is the (random) element of $\mathcal{F}$ returned by A upon receiving an n-tuple $Z^n = (Z_1, \dots, Z_n)$ of i.i.d. samples from P, and $L(f) := \mathbf{E}_P[f(Z)]$.

(a) Prove that, for any distribution P and any $\delta \in [0,1]$,

$$\mathbf{P}\left\{ L(A(Z^n)) > \inf_{f \in \mathcal{F}} L(f) + \varepsilon \right\} \le \delta, \qquad \text{if } n \ge n(\varepsilon\delta).$$

(b) Let $Z^n(1), \dots, Z^n(k)$ be a collection of k independent n-tuples $Z^n(1), \dots, Z^n(k)$ of i.i.d. draws from P. For each $j \in [k]$, let $\widehat{f}_j = A(Z^n(j))$ — that is, we run the algorithm A independently on each of the k training sets. Prove that, if $n \ge n(\varepsilon\eta)$ for some $\eta \in [0,1]$, then

$$\mathbf{P}\left\{ \min_{1 \le j \le k} L(\widehat{f}_j) > \inf_{f \in \mathcal{F}} L(f) + \varepsilon \right\} \le \eta^k.$$

(c) Use the result of Part (b) to show that one can use A to design another learning algorithm $\tilde{A}$ with the following property: for any distribution P on Z,

$$\mathbf{P}\left\{ L(\tilde{A}(Z^n)) > \inf_{f \in \mathcal{F}} L(f) + \varepsilon \right\} \le \delta$$

with

$$n = n(\varepsilon/4) \left\lceil \log_2(2/\delta) \right\rceil + \left\lceil \frac{8}{\varepsilon^2} \left(\log(4/\delta) + \log \left\lceil \log_2(2/\delta) \right\rceil \right) \right\rceil.$$

Hint: Split the sample Z^n into $k+1$ disjoint subsamples, where the first k subsamples each have size $n(\varepsilon/4)$. Run A independently on each of these first k subsamples to generate $\widehat{f}_1, \dots, \widehat{f}_k \in \mathcal{F}$. Now use the remaining subsample to select a suitable hypothesis among $\{\widehat{f}_1, \dots, \widehat{f}_k\}$.

(d) In your own words, explain the conceptual idea behind the result of part (c).