ECE 598MR: Statistical Learning Theory

Maxim Raginsky

Homework 3

Assigned December 1, 2015; due December 8, 2015

---

1. In this problem, you will prove that the excess risk of ERM for binary classification can, in certain cases, be as low as $O(1/n)$, in contrast to the usual $O(1/\sqrt{n})$ behavior (here n is the size of the training set). For simplicity, we will only consider the case when the class $\mathscr{F}$ of candidate classifiers $f : \mathsf{X} \to \{0,1\}$ is a finite set.

   Thus, let $(X, Y) \in \mathsf{X} \times \{0,1\}$ be a random couple with distribution $P = P_{XY}$, and let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be n i.i.d. samples from P. Consider forming the usual empirical estimate of the loss $L(f) = \mathbb{P}(f(X) \neq Y)$ of every classifier $f \in \mathscr{F}$:

$$L_n(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{f(X_i) \neq Y_i\}},$$

   so that the ERM solution is

$$\widehat{f}_n = \operatorname*{arg\,min}_{f \in \mathscr{F}} L_n(f) \equiv \operatorname*{arg\,min}_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{f(X_i) \neq Y_i\}}.$$

   (a) Prove that, for any $f \in \mathscr{F}$,

$$L(f) \leq L_n(f) + \sqrt{\frac{2L(f)\log(1/\delta)}{n}} + \frac{2\log(1/\delta)}{3n}$$

   with probability at least $1 - \delta$.

   Hint: You may need the following version of *Bernstein's inequality* — if $U_1, \ldots, U_n$ are n i.i.d. Bernoulli(p) random variables, then

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} U_i < p - \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2p + 2\varepsilon/3}\right).$$

   (b) Use the result from part (a) to show that, for any $f \in \mathscr{F}$,

$$L(f) \leq L_n(f) + \sqrt{\frac{2L_n(f)\log(1/\delta)}{n}} + \frac{4\log(1/\delta)}{n}$$

with probability at least $1-\delta$. Use this to prove that if the ERM solution classifies every training example correctly, i.e., if $L_n(\widehat{f}_n) = 0$, then

$$L(\widehat{f}_n) \le \frac{4\log(|\mathscr{F}|/\delta)}{n}, \qquad \text{with probability at least } 1-\delta.$$

(In particular, this bound holds when the relationship between X and Y is deterministic, $Y = f(X)$, and the function f happens to lie in $\mathscr{F}$.)

Hint: You may need the fact that, for any three nonnegative numbers a, b, c, $a \le b + c\sqrt{a}$ implies $a \le b + c^2 + c\sqrt{b}$.

2. Let A be a learning algorithm with the following guarantee: for any ε and for any $n \ge n(\varepsilon)$,

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

for every distribution P on Z.

(a) Show that, for every $\delta \in (0,1)$, if $n \ge n(\varepsilon\delta)$, then

$$L(A(Z^n)) \le \inf_{f\in\mathscr{F}} L(f) + \varepsilon$$

with probability at least $1-\delta$.

(b) For every $\delta \in (0,1)$, define

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

Suppose that the loss function ℓ is bounded between 0 and 1. Show that one can use A to construct a learning algorithm $\tilde{A}$ that achives

$$L(\tilde{A}(Z^n)) \le \inf_{f\in\mathscr{F}} L(f) + \varepsilon \qquad \text{with probability } \ge 1-\delta$$

for all $n \ge n(\varepsilon,\delta)$.

Hint: Let $k = \lceil\log_2(2/\delta)\rceil$. Split the index set $I = \{1,\dots,n\}$ into $k+1$ disjoint subsets $T_1,\dots,T_k,V$, where $|T_j| = n(\varepsilon/4)$ for each $1 \le j \le k$. Let $Z^{(j)} = (Z_i)_{i\in I_j}$, $1 \le j \le k$, and $Z^v = (Z_i)_{i\in V}$. For each $1 \le j \le k$, let $f_j = A(Z^{(j)})$ be the output of A on the subsample $Z^{(j)}$. Prove that

$$\mathbb{P}\left[L(f_j) > \inf_{f\in\mathscr{F}} L(f) + \frac{\varepsilon}{2} \text{ for all } 1 \le j \le k\right] \le 2^{-k} \le \frac{\delta}{2}.$$

Finally, use the remaining subsample Z^v as the validation set (i.e., use it to select a suitable hypothesis among $f_1,\dots,f_k$).