Slow Rates in ERM

August 5, 2020

Maybe it isn't surprising, but it can be shown that for any binary classification rule based on a dataset of a fixed size, there is some distribution for which the Bayes risk is 0, but the expected risk of the classification rule is large. This holds for ERM (as well as any other algorithm you might think of), but it does **not** mean that the risk is bounded away from zero for all n—notice the specification of a fixed sample size. The claim is that there is a "bad" distribution for **each** n, not that there is a distribution that is bad for all n (otherwise we wouldn't have consistency guarantees). In other words, one cannot find a classification rule that is guaranteed to have a certain performance across all distributions given a fixed number of samples. More precisely:

Theorem 1. For each $\epsilon > 0$ and binary classification rule ϕ_n based on a dataset $\mathcal{D} = \{(X_i, Y_i)\}^n$ of n samples, there exists a distribution over (X, Y) such that:

$$\inf_{\phi} \mathbb{P}(\phi(X) \neq Y) = 0$$

but:

$$\mathbb{P}(\phi_n(X) \neq Y) > \epsilon$$

Here is a somewhat more surprising result concerning the convergence *rate* of the risk for **all** n:

Theorem 2. Let $\{a_n\}$ be any sequence of positive numbers decreasing to 0 where $a_0 = 1/16$. Then, for any sequence of classification rules there is a distribution over (X, Y) such that for all n:

$$\inf_{\phi} \mathbb{P}(\phi(X) \neq Y) = 0$$

but:

$$\mathbb{P}(\phi_n(X) \neq Y) > a_n$$

In other words, even though a classification rule is consistent, one can always find a distribution such that the error probability decreases to 0 arbitrarily slowly. Of course, this means that in order to investigate the convergence rates associated with some algorithm, we have to make some assumptions on the distribution (X, Y). A particular example is a "low-noise" condition specifying that the posterior distribution is regular at the boundary $\eta(x) = \mathbb{P}(Y = 1|X = x) = 1/2$ [1].

Now a proof of the first theorem. For each n, we just need to show that at least a single "bad" distribution exists. As a simple example, let $X \sim \text{Uni}([i]_{i=0}^{k-1})$ for some positive integer k and let the distribution over (X, Y) be parameterized by the value of $B \sim \text{Uni}[0, 1]$ independent of X such that $Y = \text{dyadic}(B; X)$ where $\text{dyadic}(b; i)$ gives the value in the *ith* position of the binary expansion of b (so $\text{dyadic}(1/2; 0) = 1$).

Since we have a one-to-one mapping from X to Y, the Bayes error rate is 0 (how to deal with numbers with two different binary expansions is left as an exercise to the reader). Consider the expected error probability for a classifier ϕ_n based on a dataset of n pairs $\{(X_i, Y_i)\}^n$:

$$\begin{aligned} \mathbb{P}(\phi_n(X; \{(X_i, Y_i)\}^n) \neq Y) &\geq 1/2\mathbb{P}(X \notin \{X_i\}^n) + 0 && (1) \\ &= 1/2(1 - 1/k)^n && (2) \end{aligned}$$

Where the inequality follows from conditioning on the event $\{X \notin \{X_i\}^n\}$ and recognizing that in the best case either no error is made (since we already know the corresponding value of dyadic$(B; X_i)$) or we make a mistake with probability 1/2 since each element dyadic$(B; i) \sim$ Bernoulli(1/2) and is independent of X for $X \neq i$.

The theorems and discussion are mostly from chapter 7 of Devroye et al. [2].

References

[1] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, “Convexity, Classification, and Risk Bounds,” *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 138–156, 2006.

[2] L. Devroye, L. Gyorfi, and G. Lugosi, “A Probabilistic Theory of Pattern Recognition,” *Discrete Applied Mathematics*, vol. 73, no. 2, pp. 192–194, 1997.