CS281B/Stat241B: Advanced Topics in Learning & Decision Making

Statistical Learning Theory—Finite Case II

Lecturer: Michael I. Jordan *Scribes: Samson Cheung*

1 Optimistic case for finite function classes

In this lecture, we consider the "optimistic" bound for a finite function class Λ. This is the case where there is at least one function $\alpha \in \Lambda$ with zero risk, i.e. $R(\alpha) = 0$. For such an α, the empirical risk $R_{\text{emp}}(\alpha)$ must also be zero. Since the goal is to minimize risk, we only need to consider those α's with zero empirical risk. The probability we attempt to bound can thus be defined as follows:

$$P\left\{\sup_{1\le k\le N}(|R(\alpha_k) - R_{\text{emp}}(\alpha_k)|)\,\hat{\theta}(R_{\text{emp}}(\alpha_k)) > \epsilon\right\},$$

where $\hat{\theta}(R_{\text{emp}}(\alpha_k))$ is the following indicator function:

$$\hat{\theta}(u) \triangleq \begin{cases} 1 & \text{if } u = 0, \\ 0 & \text{otherwise.} \end{cases}$$

We can compute the bound with the following series of inequalities:

$$\begin{aligned} & P\left\{\sup_{1\le k\le N}(|R(\alpha_k) - R_{\text{emp}}(\alpha_k)|)\,\hat{\theta}(R_{\text{emp}}(\alpha_k)) > \epsilon\right\} \\ \le\ & \sum_{k=1}^{N} P(|R(\alpha_k) - R_{\text{emp}}(\alpha_k)|)\,\hat{\theta}(R_{\text{emp}}(\alpha_k)) > \epsilon) && (1) \\ \le\ & (N-1)\sup_{1\le k\le N} P(|R(\alpha_k) - R_{\text{emp}}(\alpha_k)|)\,\hat{\theta}(R_{\text{emp}}(\alpha_k)) > \epsilon) && (2) \\ \le\ & (N-1)P_\epsilon. && (3) \end{aligned}$$

Inequality (1) is based on the union bound. The $N-1$ in (2) is due to the optimistic assumption that at least one α has zero risk, thus making at least one the probabilities in the sum of (1) zero. P_ϵ in (3) is the probability that any function with probability of error larger than ϵ has zero empirical risk. For any such function, the probability that it produces zero empirical risk for a single trial is less than or equal to $1-\epsilon$. For l independent trials, we have

$$P_\epsilon \le (1-\epsilon)^l.$$

Substituting this bound of P_ϵ into (3), we obtain

$$P\left\{\sup_{1\le k\le N}(|R(\alpha_k) - R_{\text{emp}}(\alpha_k)|)\,\hat{\theta}(R_{\text{emp}}(\alpha_k)) > \epsilon\right\} \le (N-1)(1-\epsilon)^l. \quad (4)$$

Let $\eta = (N-1)(1-\epsilon)^l$ and solve for ϵ in terms of η for $N > 1$:

$$\begin{aligned} \epsilon &= 1 - \exp\left\{-\frac{\ln(N-1) - \ln\eta}{l}\right\} \\ &\le \frac{\ln(N-1) - \ln\eta}{l} \end{aligned}$$

The last inequality holds because $x \geq 1 - e^{-x}$. Substituting this bound back into (4), we obtain our final bound,

$$R(\alpha) \quad \leq \quad R_{\text{emp}}(\alpha) + \frac{\ln(N-1) - \ln \eta}{l} \tag{5}$$

which holds with probability $1-\eta$, for all α in Λ. In particular, it holds for α_l which is the α that minimizes the empirical risk. The bound in (5) decays at rate $1/l$, much faster than $1/\sqrt{l}$ in the pessimistic case.

2 General case for finite function classes

By using the multiplicative Chernoff bound for Bernoulli variables,

$$P\left(\frac{p - \frac{1}{n}S_n}{\sqrt{p}} > \epsilon\right) \quad \leq \quad \exp\left\{-\frac{\epsilon^2 l}{2}\right\} \tag{6}$$

we can compute a general bound with both pessimistic and optimistic bounds as special cases. The probability we want to bound is

$$P\left\{\sup_{1 \leq k \leq N} \frac{R(\alpha_k) - R_{\text{emp}}(\alpha_k)}{\sqrt{R(\alpha_k)}} > \epsilon\right\}.$$

By using (6) in a similar way to the pessimistic case, we obtain the following bound for $R(\alpha_l)$,

$$R(\alpha_l) \quad \leq \quad R_{\text{emp}}(\alpha_l) + \left(\frac{\ln N - \ln \eta}{l}\right)\left(1 + \sqrt{1 + 2\frac{R_{\text{emp}}(\alpha_l) l}{\ln N - \ln \eta}}\right) \tag{7}$$

which holds with probability $1-\eta$. In fact, the bound (7) holds for all $\alpha_k \in \Lambda$. Note that when the empirical risk equals zero, the bound (7) differs from the bound (5) only by the factor of two. When the value of the empirical risk is close to $\frac{1}{2}$ and l is rather large, (7) is close to the pessimistic bound,

$$R(\alpha_l) \leq R_{\text{emp}}(\alpha_l) + \sqrt{\frac{\ln N - \ln \eta}{2l}}.$$

To estimate how close the risk $R(\alpha_l)$ is to the minimal risk $R(\alpha_o)$, we can combine (7) with the following bound obtained in the pessimistic case,

$$R(\alpha_o) > R_{\text{emp}}(\alpha_o) - \sqrt{\frac{-\ln \eta}{2l}},$$

which also holds with probability $1-\eta$. Thus, with probability $1 - 2\eta < (1-\eta)^2$, we obtain,

$$\begin{aligned}
R(\alpha_l) \quad &\leq \quad R_{\text{emp}}(\alpha_l) + \left(\frac{\ln N - \ln \eta}{l}\right)\left(1 + \sqrt{1 + 2\frac{R_{\text{emp}}(\alpha_l) l}{\ln N - \ln \eta}}\right) \\
&\leq \quad R_{\text{emp}}(\alpha_o) + \left(\frac{\ln N - \ln \eta}{l}\right)\left(1 + \sqrt{1 + 2\frac{R_{\text{emp}}(\alpha_l) l}{\ln N - \ln \eta}}\right) \\
&< \quad R(\alpha_o) + \sqrt{\frac{-\ln \eta}{2l}} + \left(\frac{\ln N - \ln \eta}{l}\right)\left(1 + \sqrt{1 + 2\frac{R_{\text{emp}}(\alpha_l) l}{\ln N - \ln \eta}}\right).
\end{aligned}$$

In the next lecture, we will consider the case when the function class Λ is infinite. We need an alternative notion to measure the capacity of the function class in this case. Even though there are infinitely many functions in Λ, for a given set of samples, only a finite number of clusters of events is distinguishable. We define the total number of such equivalent events as $N^\wedge(l) \equiv N^\wedge(z_1, z_2, \ldots, z_l) \leq 2^l$ and introduce the concept of entropy as follows:

$$H^\wedge(l) \triangleq \int \ln N^\wedge(z_1, \ldots, z_l)\, dF(z_1, \ldots, z_l).$$

The key to measuring the capacity is to compute the growth rate of $H^\wedge(l)$. If $H^\wedge(l)$ grows linearly with l, we cannot learn anything about the problem using functions from Λ. The surprising fact shown by Vapnik is that $H^\wedge(l)$ can grow either linearly or logarithmically with l. To show this result, we need the following bound which will be proved in the next lecture:

$$\begin{aligned} P\left\{\sup_{\alpha\in\Lambda} |R(\alpha) - R_{\text{emp}}(\alpha)| > \epsilon\right\} \quad &\leq \quad 4\exp\{(\frac{H^\wedge_{\text{ann}}(2l)}{l} - (\epsilon - \frac{1}{l})^2)l\} \\ &= \quad 4E(N^\wedge(2l))\exp\left\{-(\epsilon - \frac{1}{l})^2 l\right\}, \end{aligned} \tag{8}$$

where $H^\wedge_{ann}(l) \triangleq \ln E(N^\wedge(l))$ is called the *Annealed Entropy.* Comparing (8) with the bound obtained in the finite case using Chernoff bounds, we see that the function count N is now replaced by the annealed entropy.