CS281B/Stat241B (Spring 2006) Statistical Learning Theory **Lecture: 14**

Risk Bounds and Uniform Convergence

Lecturer: Peter Bartlett *Scribe: John DeNero*

1 Risk Bounds

This lecture continues our previous discussion of the relationship between the empirical risk and true risk of a classifier. In particular, we are interested in the asymptotic convergence of the empirical risk as the classifier is presented with increasing amounts of training data.

Consider the following definitions from early in the course for a classifier $f : \mathcal{X} \mapsto \{+1, -1\}$, where $f \in F$, so class of functions.

$$\begin{aligned} L(f) &= \mathbb{E}l(f(\mathcal{X}), \mathcal{Y}) \\ \hat{L}(f) &= \hat{\mathbb{E}}l(f(\mathcal{X}), \mathcal{Y}) = \frac{1}{n}\sum_{i=1}^{n} l(f(X_i), Y_i) \\ \hat{f} &= \arg\min_{f\in F} \hat{L}(f) \end{aligned}$$

And for any one function $f \in F$, we know that $\hat{L}(f) \to L(f)$ almost surely as $n \to \infty$ from the law of large numbers. What we would like to show, however is something stronger: that the risk of the function we find through empirical risk minimization is in fact the function with the lowest risk. This stronger result will require more than the law of large numbers.

Example. Consider $F = F_+ \cup F_-$, where

$$\begin{aligned} F_+ &= \{f \in \{+1, -1\}^{\mathcal{X}} : |\{x : f(x) = -1\}| < \infty\} \\ F_- &= \{f \in \{+1, -1\}^{\mathcal{X}} : |\{x : f(x) = 1\}| < \infty\} \end{aligned}$$

Note that for any finite sequence, we can choose f from either F_+ or F_- to explain it. Now, suppose we have a distribution P such that $P(Y = 1|X) = 0.9$ almost surely, and for all x, $P(X = x) = 0$. Then, we have

$$\begin{aligned} f \in F_+ \quad &\text{implies} \quad L(f) = L^* = 0.1 \\ f \in F_- \quad &\text{implies} \quad L(f) = 0.9 > L^* \end{aligned}$$

where L^* is the Bayes risk.

However, for any finite sample there is an $f \in F$ with $\hat{L}(f) = 0$ but $L(f) - L^* = 0.8$. So, choosing a function from a class via empirical risk minimization does not guarantee risk minimization with such a rich class. Restated,

$$L(\arg\min_{f\in F} \hat{L}(f)) \neq \inf_{f\in F} L(f)$$

The goal for today will be to develop tools to show properties relating the empirical risk minimizer (LHS) and the minimal risk (RHS).

2 Concentration Inequalities

We will be interested not only in whether $L(\arg\min_{f\in F}\hat{L}(f)) \to \inf_{f\in F} L(f)$, but how fast this convergence happens, called the rate of convergence. One way to show this is via uniform convergence, that is, a statement of the form

$$\sup_{f\in F} |\hat{L}_n(f) - L(f)| \leq \epsilon$$

2.1 Classic bounds

For this, several classic inequalities are useful that impose upper bounds on the total probability mass contained within the tail of a distribution.

Theorem 2.1. (Markov's Inequality) If $X \geq 0$ a.s. and $t > 0$, then $Pr(X \geq t) \leq \frac{\mathbb{E}X}{t}$.

Proof. $\mathbb{E}X \geq \mathbb{E}\left[X \cdot \mathbb{1}(X \geq t)\right] \geq t \cdot Pr(X \geq t) + 0 \cdot Pr(X < t)$ □

Theorem 2.2. (Chebyshev's Inequality) If $t > 0$, then $Pr(|X - \mathbb{E}X| \geq t)) \leq \frac{\mathrm{Var(X)}}{t^2}$.

Proof. Apply Markov's inequality to $(X - \mathbb{E}X)^2$. □

2.2 Independent Variables

These upper bounds are not necessarily tight, of course.

Example. For instance, the probability mass in the tail of a Gaussian is bounded by an exponential function while Chebyshev's will impose only a quadratically shrinking bound.

For a binomial random variable sampled n times independently, we can approximate its sum with a Gaussian. Consider $Z_i \in \{0, 1\}$ with identical $P(Z_i = 1) = p$ for all i and the quantities $S_n = \sum_{i=1}^{n} Z_i$, $\sigma^2 = \mathrm{Var(Z_i)} = \mathrm{p(1-p)}$, $\mathrm{Var(S_n)} = \mathrm{n}\sigma^2$. From Chebyshev's inequality, choosing $t = n\epsilon$, we have

$$P\left(\left|\frac{S_n}{n} - \frac{\mathbb{E}\left[S_n\right]}{n}\right| > \epsilon\right) = P\left(\left|\frac{S_n}{n} - p\right| > \epsilon\right) = \frac{\sigma^2}{n\epsilon^2}$$

But this bound on the empirical risk error is in fact quite weak. We can additionally show that this quantity has an exponentially decreasing bound by appealing to the central limit theorem:

$$\lim_{n\to\infty} Pr\left(\sqrt{\frac{n}{\sigma^2}}\left|\frac{S_n}{n} - p\right| > t\right) = 1 - \Phi(t) \leq \frac{c}{t}\exp\left(\frac{-t^2}{2}\right)$$

This tight bound comes from the independence of the summed variables.

2.3 Hoeffding's inequality

We can show concentration inequalities for sums of independent random variables more generally. Note that the following bounds leverage independence, but don't require identical distributions among the variables involved.

Consider independent $X_1, ..., X_n$ and their sum, $S_n = \sum_{i=1}^n X_i$. The quantity of interest that we intend to bound is

$$Pr(S_n - \mathbb{E}[S_n] \geq t) \tag{1}$$

Via a monotonic transformation, we arrive at the identical quantity for $s > 0$.

$$Pr(e^{s(S_n - \mathbb{E}[S_n])} \geq e^{st})$$

Exponentiation gives us a positive random variable, so we applying Markov's inequality to get

$$\begin{aligned} Pr(S_n - \mathbb{E}[S_n] \geq t) &= Pr(e^{s(S_n - \mathbb{E}[S_n])} \geq e^{st}) \\ &\leq e^{-st}\, \mathbb{E}\left[e^{s(S_n - \mathbb{E}[S_n])}\right] \\ &= e^{-st}\, \mathbb{E}\left[\exp(s \sum_{i=1}^n (X_i - \mathbb{E}X_i))\right] \\ &= e^{-st} \prod_{i=1}^n \mathbb{E}\left[e^{s(X_i - \mathbb{E}X_i)}\right] \ , \end{aligned}$$

where the last step uses independence of the X_i's. So we see that quantity (1) above can be bounded by an expression involving only the product of moment generating functions $\mathbb{E}\left[e^{s(X_i - \mathbb{E}X_i)}\right]$. This fact gives rise to a general approach: by bounding these moment generating functions we will bound the original quantity of interest.

Lemma 2.3. If X has $\mathbb{E}X = 0$ and $a \leq X \leq b$, then for $s > 0$,

$$\mathbb{E}[e^{sx}] \leq e^{\frac{s^2(b-a)^2}{8}}$$

Proof. Only a brief sketch of a proof will be provided. First, we note that the mapping $x \rightarrow e^{sx}$ is convex for $s > 0$. Thus, we have

$$e^{sx} \leq \frac{x-a}{b-a} e^{sb} + \frac{b-x}{b-a} e^{sa}$$

Hence, we can bound the expected value of this quantity by

$$\mathbb{E}[e^{sx}] \leq \frac{be^{sa} - ae^{sb}}{b-a}$$

The remainder of the proof and the appearance of the 8 constant in the denominator of the exponent arise from taking the Taylor series expansion of the log of the right hand side of the above equation about $s = 0$. □