CS281B/Stat241B (Spring 2008) Statistical Learning Theory **Lecture: 12**

Ada Boost, Risk Bounds, Concentration Inequalities

Lecturer: Peter Bartlett *Scribe: Subhransu Maji*

1 AdaBoost and Estimates of Conditional Probabilities

We continue with our discussion on AdaBoost and derive risk bounds of the classifier. Recall that for a function f, we have the following relationship between the expected excess risk and the excess ϕ approximation risk for a loss function ϕ,

$$\Psi\left(R(f) - R^*\right) \leq R_\phi(f) - R_\phi^*$$

where, R^* is the optimal Bayes Risk, R_ϕ^* is the risk of the optimal f i.e.$R_\phi^* = \inf_f R_\phi^*(f)$ and $H(\eta)$ is the function

$$H(\eta) = \inf_{\alpha \in R} \left[\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\right]$$

and $\Psi(\theta)$ is the function

$$\begin{aligned}\Psi(\theta) &= H^-\left(\frac{1+\theta}{2}\right) + H\left(\frac{1+\theta}{2}\right) \\ &= \phi(0) - \inf_{\alpha \in R}\left[\frac{1+\theta}{2}\phi(\alpha) + \frac{1-\theta}{2}\phi(-\alpha)\right]\end{aligned}$$

In the context of AdaBoost the loss function $\phi(\alpha) = e^{-\alpha}$ is convex and classification calibrated. Thus,

$$H(\eta) = \inf_{\alpha \in R}\left[\eta e^{-\alpha} + (1-\eta)e^{\alpha}\right]$$

Differentiating w.r.t. α and setting to zero gives us the optimal

$$\alpha(\eta) = \ln\sqrt{\frac{\eta}{1-\eta}}$$

This suggests that if we could choose $f(x)$ separately for each x, it would be a monotonically transformed version of conditional probability(see next section). Plugging this α^* into H yields

$$H(\eta) = 2\sqrt{\eta(1-\eta)},$$

which is concave and symmetric around 1/2. Then $\Psi(\theta)$ simplifies to

$$\Psi(\theta) = 1 - \sqrt{(1+\theta)(1-\theta)} = 1 - \sqrt{1-\theta^2}.$$

Finally, plugging this in to the original inequality yields

$$1 - \sqrt{1 - (R(f) - R^*)^2} \leq R_\phi(f) - R_\phi^*.$$

Examining the Taylor series of the left side about 0 shows that this is equivalent, for some constant c, to

$$R(f) - R^* \leq c\sqrt{(R_\phi(f) - R_\phi^*)}$$

when the excess ϕ-risk is sufficiently small. Thus, driving the excess ϕ-risk to zero will drive the discrete loss to zero as well, which justifies AdaBoost's use of this particular convex loss function.

2 Relationship to logistic regression

It turns out that we can interpret the value of $F(x)$ (where F is the boosted classifier returned by AdaBoost) as a transformed estimate of $\Pr(Y = 1|X = x)$. Consider a logistic model where

$$\Pr(Y = 1|X = x) = \frac{1}{1 + e^{-2f(x)}} = \frac{e^{f(x)}}{e^{f(x)} + e^{-f(x)}},$$

a rescaled version of the logistic function. In this model, the log loss (negative log likelihood) takes the form

$$\begin{aligned} -\ln \prod_{i=1}^{n} \Pr(Y = y_i|X = x_i) &= -\sum_{y_i=1} \ln \frac{1}{1 + e^{-2f(x)}} - \sum_{y_i=-1} \ln \left(1 - \frac{1}{1 + e^{-2f(x)}}\right) \\ &= \sum_{y_i=1} \ln \left(1 + e^{-2f(x)}\right) + \sum_{y_i=-1} \ln \left(1 + e^{2f(x)}\right) \\ &= \sum_{i=1}^{n} \ln \left(1 + e^{-2y_i f(x_i)}\right). \end{aligned}$$

Thus, the maximum likelihood logistic regression solution attempts to minimize the sample average of $\phi(\alpha) = \ln\left(1 + e^{-2\alpha}\right)$. This is closely related to AdaBoost, which minimizes the sample average of $\phi(\alpha) = e^{-\alpha}$. To see the connection, note that the first few terms of the Taylor expansion of $\ln\left(1 + e^{-2\alpha}\right) + 1 - \ln 2$ about 0,

$$1 - \alpha + \frac{\alpha^2}{2} - \ldots,$$

are identical to those of $e^{-\alpha}$.

While the two functions are very similar near zero, their asymptotic behavior is very different. In general we have that

$$\ln\left(1 + e^{-2\alpha}\right) \leq e^{-\alpha};$$

furthermore, the former grows linearly as α approaches $-\infty$, whereas the latter grows exponentially. Thus, we can view AdaBoost as approximating the maximum likelihood logistic regression solution, except with (sometimes exponentially) larger penalties for mistakes. A further similarity between the methods is that the α^* for $\phi(\alpha) = \ln\left(1 + e^{-2\alpha}\right)$ is the same as for AdaBoost.

3 Risk Bounds and Uniform Convergence

So far, we've looked at algorithms (including AdaBoost) that optimize over a set of training samples:

$$\min_{f \in F} \hat{R}(f) = \hat{\mathbb{E}} l(y, f(x)) = \frac{1}{n} \sum_{i=1}^{n} l(y_i, f(x_i)).$$

If the empirical minimizer is $\hat{f}$, we are interested in bounding the true loss $R(\hat{f}) = \mathbb{E} l(y, \hat{f}(x))$ under this function. In particular, we hope that $\hat{R}(\hat{f})$ will converge to $\inf_{f \in F} R(f)$ as $n \to \infty$

For the (trivial) case where our function class F contains only a single function, we can simply appeal to the law of large numbers. For example, in the case of discrete loss, the Chernoff bound gives an upper bound on $\Pr(|\hat{R}(f) - R(f)| > \epsilon)$ that shrinks exponentially in n for any given ϵ.

This argument, however, fails when F is not a singleton. We cannot simply apply the law of large numbers to each $f \in F$ and then argue that the desired property holds when minimizing over all of F. The problem

is that we are considering $R(\text{argmin}_{f\in F}\hat{R}(f))$, where the inner part depends on the data. In particular, if F is such that for any n and data set there are functions $f \in F$ with small $\hat{R}(f)$ but large $R(f)$, then choosing an f that minimizes $\hat{R}(f)$ may not tend to minimize $R(f)$.

Example. Let $F = F_+ \cup F_-$ with

$$F_+ = \{x \mapsto f(x) : |\{x : f(x) = +1\}| < \infty\}$$
$$F_- = \{x \mapsto f(x) : |\{x : f(x) = -1\}| < \infty\}$$

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 $P(Y = 1|X) = 0.95$ almost surely, and for all x, $P(X = x) = 0$. Then, we have

$$f \in F_+ \Rightarrow R(f) = 0.05 = R^*$$
$$f \in F_- \Rightarrow R(f) = 0.95 > R^*$$

where, R^* is the Bayes risk. However, for any finite sample there is an $f \in F_+$ with $R(f) = 0$ but $R(f) - R^* = 0.9$. So, choosing a function from a class via empirical risk minimization does not guarantee risk minimization with such a rich class. Restated,

$$R(\text{argmin}_{f\in F}R(f)) \neq \inf_{f\in F} R(f)$$

Example. If the set of functions $F \in \{+1, -1\}^X$ is finite, then we *can* say something about the true risk given that the empirical risk is zero. The following theorem makes this explicit.

Theorem 3.1.

$$\Pr(\exists f \in F \text{ and } \hat{R}(f) = 0 \ \&\ R(f) \geq \epsilon) \leq |F|e^{-\epsilon n}$$

i.e., with probability at least $1 - \delta$,

$$\text{if, } \hat{R}(f) = 0, \text{ then } R(f) \leq \frac{\log|F|}{n} + \frac{\log 1/\delta}{n}$$

Proof. To show this we use the properties of the exponential functions and union bounds. For any $f \in F$ with $R(f) \geq \epsilon$, we have

$$\begin{aligned}\Pr(\hat{R}(f) = 0) &\leq (1-\epsilon)^n \\ &= \exp(n\log(1-\epsilon)) \\ &\leq \exp(-n\epsilon) \end{aligned} \tag{1}$$

Using the union bound (Boole's inequality: the probability of a union of events is no more than the sum of their probabilities), we have

$$\begin{aligned}\Pr\left(\bigcup\{f \in F : R(f) \geq \epsilon \ \&\ \hat{R}(f) = 0\}\right) &\leq \sum_{f\in F} \Pr\left(R(f) \geq \epsilon \ \&\ \hat{R}(f) = 0\right) \\ &\leq |F|e^{-\epsilon n} \end{aligned} \tag{2}$$

□

Example. (Decision Trees) Consider the class of decision trees of finite number of nodes N over $x \in \{+1, -1\}^d$. Thus $|F| \leq (d + 2)^N$, because we can specify the tree by listing, in breadth-first order, the N nodes of the tree, and each can be either one of the covariates or outputs $\{+1, -1\}$. Thus, if $\hat{R}(f) = 0$, then with probability $\geq 1 - \delta$,

$$R(f) \leq \frac{N\log(d+2)}{n} + \frac{\log 1/\delta}{n}$$

Example. F is parameterized using N bits, i.e. $F = \{x \mapsto \phi(x,b), b \in \{0,1\}^N\}$ with $f_b(x) = \phi(x,b)$. $|F| = 2^N$ and thus if $\hat{R}(f) = 0$, then with probability $\geq 1-\delta$,

$$R(f) \leq \frac{N}{n} + \frac{\log 1/\delta}{n}$$

Typically when we learn classifiers on training data the empirical risk is small but not zero and the above theorem can not be applied directly. In the next few sections we will be developing tools to show properties relating the empirical risk minimizer and the minimal risk.

4 Concentration Inequalities

We will be interested not only in whether $R(\arg\min_{f\in F}\hat{R}(f)) \to \inf_{f\in F} R(f)$, but how fast this convergence happens, called the rate of convergence.

4.1 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 4.1. (Markov's Inequality) If $X \geq 0$ a.s. and $t > 0$, then $\Pr(X \geq t) \leq \frac{\mathbb{E}(X)}{t}$

Proof. $\mathbb{E}X \geq \mathbb{E}[X1(X \geq t)] \geq t\Pr(X \geq t) + 0\Pr(X < t) = t\Pr(X \geq t)$ □

Theorem 4.2. (Chebyshev's Inequality) If $t > 0$, then $\Pr(|X - \mathbb{E}X| \geq t)) \leq \frac{Var(X)}{t^2}$

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

These upper bounds are not necessarily tight as seen in the following example

Example. Let $Z_i = \{0,1\}$ be i.i.d with $\Pr(Z_i = 1) = p$. Denote $S_n = \sum_i^n z_i$, then using Chebyshev's inequality on the variable S_n/n we have

$$\Pr\left(\left|\frac{S_n}{n} - \frac{\mathbb{E}S_n}{n}\right| > t\right) \leq \frac{Var(S_n/n)}{t^2} = \frac{p(1-p)}{nt^2}$$

On the other hand, the central limit theorem says

$$\sqrt{\frac{n}{\sigma^2}}\left(\frac{S_n}{n} - p\right) \to N(0,1)$$

Thus,

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

where $\Phi(t)$ is the cumulative distribution function of $N(0,1)$. So, $\Pr(\frac{S_n}{n} - p \geq \epsilon)$ should decrease as $exp\left(\frac{-\epsilon^2 n}{2\sigma^2}\right)$, which is much faster than the rate implied by Chebyshev's inequality.

4.2 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.

Theorem 4.3. (Hoeffding's Inequality) Consider independent $X_i \in [a_i, b_i]$ and their sum, $S_n = \sum_{i=1}^n X_i$. Then,

$$\Pr(S_n - \mathbb{E}S_n \geq t) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

Proof. A monotonic transformation and exponentiation using $s > 0$, gives us a positive random variable. Applying Markov's inequality we get,

$$\begin{aligned}
\Pr(S_n - \mathbb{E}S_n \geq t) &= \Pr(e^{s(S_n - \mathbb{E}S_n)} \geq e^{st}) \\
&\leq e^{-st}\mathbb{E}\left[e^{s(S_n - \mathbb{E}S_n)}\right] \text{ (Markov's Inequality)} \\
&= e^{-st}\mathbb{E}\left[\exp(s(\sum_{i=1}^n (X_i - \mathbb{E}X_i)))\right] \\
&= e^{-st}\mathbb{E}\prod_{i=1}^n \left[e^{s(X_i - \mathbb{E}X_i)}\right]
\end{aligned} \tag{3}$$

where, the last inequality uses the independence of the variables. We will see in the next lecture a bound on the last inequality, which will complete the proof.

Example. Let $X_i = \{0, 1\}$ denote the loss on the $i'th$ example. Then, $S_n = n\hat{R}(f)$ and $\mathbb{E}S_n = nR(f)$. Applying Hoeffding's inequality we get

$$P(|\hat{R}(f) - R(f)| > \epsilon) \leq 2\exp\left(\frac{-2\epsilon^2 n^2}{\sum_{i=1}^n 1^2}\right) = 2\exp(-2n\epsilon^2)$$

Note that though the rate is right and this bound is tighter than Markov's, there is still a factor of σ^2 missing compared to the bounds one would expect from central limit theorem. □