EECS 598: Statistical Learning Theory, Winter 2014 **Topic 3**

Hoeffding's Inequality

Lecturer: Clayton Scott *Scribe: Andrew Zimmer*

1 Introduction

Suppose we are given training data $(X_i, Y_i) \overset{i.i.d.}{\sim} P_{XY}$ and a classifier $h : \mathcal{X} \to \mathcal{Y}$. Define the *empirical risk* of h to be

$$\widehat{R}_n(h) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{h(X_i) \neq Y_i\}}.$$

Notice that $n\widehat{R}_n(h) \sim \text{binom}(n, R(h))$ and so $\mathbb{E}\left[\widehat{R}_n(h)\right] = R(h)$. We would like to understand how accurate $\widehat{R}_n(h)$ is as an estimate of $R(h)$. Thankfully there are many well known *concentration inequalities* that provide us with quantitative answers to this question. The goal of this lecture is to establish one such bound: Hoeffding's inequality [2]. This inequality was originally proved in the 1960's and will imply that

$$\Pr\left(\left|\widehat{R}_n(h) - R(h)\right| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2}. \tag{1}$$

Along the way we will prove Markov's inequality, Chebyshev's inequality, and Chernoff's bounding method. A key point to notice is that the probability in (1) is with respect to the draw of the training data.

2 Markov's Inequality

Proposition 1. *If U is a non-negative random variable on $\mathbb{R}$, then for all $t > 0$*

$$\Pr\left(U \geq t\right) \leq \frac{1}{t}\mathbb{E}\left[U\right].$$

Proof. Notice that

$$\Pr\left(U \geq t\right) = \mathbb{E}\left[\mathbf{1}_{\{U \geq t\}}\right] \leq \mathbb{E}\left[\frac{U}{t}\mathbf{1}_{\{U \geq t\}}\right] = \frac{1}{t}\mathbb{E}\left[U\mathbf{1}_{\{U \geq t\}}\right] \leq \frac{1}{t}\mathbb{E}\left[U\right]$$

where both inequalities use the fact that U is nonnegative. □

2.1 Chebyshev's Inequality

Our first concentration inequality is an easy consequence of Markov's inequality.

Corollary 1. *If Z is a random variable on $\mathbb{R}$ with mean μ and variance σ^2 then*

$$\Pr\left(|Z - \mu| \geq \sigma t\right) \leq \frac{1}{t^2}.$$

Proof. If we apply Markov's inequality to the random variable $(Z-\mu)^2$ we obtain:

$$\Pr\left(|Z-\mu| \geq \sigma t\right) = \Pr\left((Z-\mu)^2 \geq \sigma^2 t^2\right) \leq \frac{1}{\sigma^2 t^2}\mathbb{E}\left[(Z-\mu)^2\right] = \frac{\sigma^2}{\sigma^2 t^2} = \frac{1}{t^2}.$$

□

Since $n\widehat{R}_n(h) \sim \text{binom}(n, R(h))$, we see that $\widehat{R}_n(h)$ is a random variable with mean $\mu = R(h)$ and variance $\sigma^2 = R(h)(1-R(h))/n$. So applying Chebyshev's inequality to $\widehat{R}_n(h)$ we obtain:

$$\Pr\left(\left|\widehat{R}_n(h) - R(h)\right| \geq \epsilon\right) \leq \frac{R(h)(1-R(h))}{n\epsilon^2}.$$

This is nice, but as mentioned in the introduction we can actually get exponential decay as a function of n.

2.2 Chernoff's Bounding Method

A second corollary of Markov's inequality is known as Chernoff's bounding method [1]. We will use this to prove Hoeffding's inequality.

Corollary 2. *Let Z be a random variable on $\mathbb{R}$. Then for all $t > 0$*

$$\Pr\left(Z \geq t\right) \leq \inf_{s>0} e^{-st} M_Z(s)$$

where M_Z is the moment-generating function of Z.

Proof. For any $s > 0$ we can use Markov's inequality to obtain:

$$\Pr\left(Z \geq t\right) = \Pr\left(sZ \geq st\right) = \Pr\left(e^{sZ} \geq e^{st}\right) \leq e^{-st}\mathbb{E}\left[e^{sZ}\right] = e^{-st}M_Z(s).$$

Since $s > 0$ was arbitrary the corollary follows. □

3 Hoeffding's Inequality

Theorem 1. *Let $Z_1, \ldots, Z_n$ be independent random variables on $\mathbb{R}$ such that $a_i \leq Z_i \leq b_i$ with probability one. If $S_n = \sum_{i=1}^n Z_i$ then for all $t > 0$*

$$\Pr\left(S_n - \mathbb{E}\left[S_n\right] \geq t\right) \leq e^{-2t^2/\sum(b_i-a_i)^2}$$

and

$$\Pr\left(S_n - \mathbb{E}\left[S_n\right] \leq -t\right) \leq e^{-2t^2/\sum(b_i-a_i)^2}.$$

Remark. By combining the two bounds we obtain:

$$\Pr\left(|S_n - \mathbb{E}\left[S_n\right]| \geq t\right) \leq 2e^{-2t^2/\sum(b_i-a_i)^2}.$$

Remark. If $Z_i \overset{i.i.d.}{\sim} \text{Ber}(p)$ then $a_i = 0$, $b_i = 1$, and $S_n \sim \text{binom}(n,p)$. Since $\mathbb{E}[S_n] = np$, Hoeffding's theorem specializes to *Chernoff's bound*

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i - p\right| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$

There are actually several Chernoff bounds, and all of apply Chernoff's bounding method in different ways. Some of the bounds additionally incorporate the variance of the random variable, which allows for tighter bounds. In our case, the variance involves $p = R(f)$ which is unknown, making the bound not computable. See the expository article [3] for more on Chernoff bounds.

The proof of Hoeffding's theorem will use Chernoff's Bounding Method and the next lemma:

Lemma 1. *Let V be a random variable on $\mathbb{R}$ with $\mathbb{E}[V] = 0$ and suppose $a \leq V \leq b$ with probability one. Then for all $s > 0$*

$$\mathbb{E}\left[e^{sV}\right] \leq e^{s^2(b-a)^2/8}.$$

Proof. If $x \in [a, b]$ then the convexity of the function $x \to e^{sx}$ implies that

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

Using the fact that $\mathbb{E}[V] = 0$ we then obtain the bound:

$$\mathbb{E}\left[e^{sV}\right] \leq \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb}. \tag{2}$$

Now let $p = b/(b-a)$ and $u = (b-a)s$ and consider the function:

$$\begin{aligned}\varphi(u) &= \log\left(pe^{sa} + (1-p)e^{sb}\right)\\ &?= sa + \log\left(p + (1-p)e^{s(b-a)}\right)\\ &= (p-1)u + \log\left(p + (1-p)e^{u}\right).\end{aligned}$$

This function is smooth and hence by Taylor's theorem for any $u \in \mathbb{R}$ there exists $\xi = \xi(u) \in \mathbb{R}$ such that

$$\varphi(u) = \varphi(0) + \varphi'(0)u + \frac{1}{2}\varphi''(\xi)u^2.$$

Now $\varphi(0) = 0$ and

$$\varphi'(u) = (p-1) + \frac{(1-p)e^u}{p + (1-p)e^u} = (p-1) + 1 - \frac{p}{p + (1-p)e^u}.$$

Hence $\varphi'(0) = 0$ and

$$\varphi''(u) = \frac{p(1-p)e^u}{(p + (1-p)e^u)^2}.$$

Using calculus one can then show that $\varphi''(\xi) \leq 1/4$ and so

$$\varphi(u) \leq u^2/8 = s^2(b-a)^2/8. \tag{3}$$

Combining the bounds in Eqns. (2) and (3) completes the proof of the lemma. □

Proof of Theorem 1. First apply Chernoff's bounding method to the random variable $S_n - \mathbb{E}[S_n]$ to obtain

$$\Pr\left(S_n - \mathbb{E}[S_n] \geq t\right) \leq \min_{s>0} e^{-st}\mathbb{E}\left[e^{s(S_n - \mathbb{E}[S_n])}\right].$$

Since the Z_i are independent we have:

$$e^{-st}\mathbb{E}\left[e^{s(S_n - \mathbb{E}[S_n])}\right] = e^{-st}\prod_{i=1}^{n}\mathbb{E}\left[e^{s(Z_i - \mathbb{E}[Z_i])}\right]$$

and so, by applying the lemma above,

$$\Pr\left(S_n - \mathbb{E}[S_n] \geq t\right) \leq \min_{s>0} e^{-st + (s^2/8)\sum(b_i - a_i)^2}.$$

Since the function $s \to -st + (s^2/8)\sum(b_i - a_i)^2$ describes a parabola we know that the minimizer is at $s = 4t/\sum(b_i - a_i)^2$ and hence

$$\Pr\left(S_n - \mathbb{E}\left[S_n\right] \geq t\right) \leq e^{-2t^2/\sum(b_i - a_i)^2}.$$

To obtain the second bound simply apply the first bound to the random variables $-Z_1, \ldots, -Z_n$. □

4 KL divergence and hypothesis testing

In this final section we will apply Hoeffding's inequality to hypothesis testing. The set up is as follows: let $\mathcal{Y} = \{0, 1\}$ and assume P_{XY} is a distribution on $\mathcal{X} \times \mathcal{Y}$. In addition let's assume that

- the prior probabilities $\pi_y = P_Y(Y = y)$ are equal,
- $P_{X|Y=y}$ has *known* density p_y,
- The supports of p_0 and p_1 are the same, i.e., $\{x : p_0(x) > 0\} = \{x : p_1(x) > 0\}$,
- $0 < \alpha \leq p_y(x) \leq \beta < +\infty$ for all x such that $p_y(x) > 0$, $y = 0, 1$.

Now suppose we observe $X_1, \ldots, X_n \overset{i.i.d.}{\sim} p_y$ where $y \in \{0, 1\}$ is unknown. Can we determine y? How good will our guess be?

We can think of this as a classification problem where $(X_1, \ldots, X_n)$ is the feature vector. The density of this vector is $\prod_{i=1}^n p_y(X_i)$. The optimal classifier is given by the likelihood ratio test (see previous notes):

$$\widehat{h}_n(x) = \begin{cases} 1 & \text{if } \frac{\prod_{i=1}^n p_1(X_i)}{\prod_{i=1}^n p_0(X_i)} \geq \frac{\pi_0}{\pi_1} = 1 \\ 0 & \text{otherwise} \end{cases}$$

Since the natural logarithm is strictly increasing, this optimal classifier can be expressed

$$\widehat{h}_n(x) = \begin{cases} 1 & \text{if } \hat{S}_n(X_1, \ldots, X_n) \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

where

$$\hat{S}_n(X_1, \ldots, X_n) = \sum_{i=1}^n \log \frac{p_0(X_i)}{p_1(X_i)}.$$

Now for a classifier $h : \mathcal{X}^n \to Y$ and $y \in \{0, 1\}$ define

$$R_y(h) = \Pr(h(X) \neq y | Y = y).$$

Then

$$\pi_0 R_0(\widehat{h}_n) + \pi_1 R_1(\widehat{h}_n)$$

is the probability that our classifier returns the wrong answer after n observations.

It turns out that we can use Hoeffding's inequality to bound this probability. Let $Z_i := \log \frac{p_0(X_i)}{p_1(X_i)}$. Then

$$\log \frac{\alpha}{\beta} \leq Z_i \leq \log \frac{\beta}{\alpha}$$

with probability one (with respect to either probability measure). Moreover,

$$R_0(\widehat{h}_n) = \Pr\left(S_n \geq 0 \middle| Y = 0\right) = \Pr\left(S_n - \mathbb{E}\left[S_n | Y = 0\right] \geq -\mathbb{E}\left[S_n | Y = 0\right] \middle| Y = 0\right)$$

and

$$\mathbb{E}\left[S_n | Y = 0\right] = n\mathbb{E}\left[Z_1 | Y = 0\right] = n \int \log\left(\frac{p_1(x)}{p_0(x)}\right) p_0(x) dx = -n \int \log\left(\frac{p_0(x)}{p_1(x)}\right) p_0(x) dx = -nD(p_0 || p_1)$$

where $D(p_0||p_1)$ is the *Kullback-Leibler divergence* of p_0 from p_1. Finally applying Hoeffding's inequality gives the following bound:

$$R_0(\widehat{h}_n) \leq e^{-2nD(p_0||p_1)^2/c^2} \text{ where } c = 4\left(\log\beta - \log\alpha\right)^2.$$

A similar analysis gives an exponential bound on $R_1(\widehat{h}_n)$ and thus we see that the probability that our classifier returns the wrong answer after n observations decays to zero exponentially and the rate of exponential decay depends on $D(p_0||p_1)$, $D(p_1||p_0)$, and α, β. The Kullback-Leibler divergence measures how close two probability distributions are, so our bound makes intuitive sense: the more distinct the two distributions, the easier it should be to determine the distribution being observed.

Remark. The Kullback-Leibler divergence belongs to a family of functions D_f defined on certain pairs of probability distributions. Given a convex function $f : \mathbb{R} \to \mathbb{R}$ with $f(0) = 1$ and two densities p and q, define the *f-divergence* of q from p to be

$$D_f(q||p) = \int f\left(\frac{p(x)}{q(x)}\right) q(x)dx.$$

Notice that the Kullback-Leibler divergence is the special case when $f(x) = -\log(x)$ and that we can repeat the above argument to obtain exponential bounds in terms of $D_f(p_0||p_1)$ and $D_f(p_1||p_0)$.

Exercises

1. (a) Apply Chernoff's bounding method to obtain an exponential bound on the tail probability $\Pr(Z \geq t)$ for a Gaussian random variable $Z \sim \mathcal{N}(\mu, \sigma^2)$.
   (b) Appealing to the central limit theorem, use part (a) to give an approximate bound on the binomial tail. This should not only match the exponential decay given by Hoeffding's inequality, but also reveal the dependence on the variance of the binomial.
2. Can you remove the assumption in Section 4 that $0 < \alpha \leq p_y(x)$? Consider other restrictions on p_y, other concentration inequalities, or other f-divergences.

References

[1] H. Chernoff, "A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations," *Annals of Mathematical Statistics*, vol. 23, no. 4, pp. 493-507, 1952.

[2] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13-30, 1963.

[3] T. Hagerup and C. Rüb, "A guided tour of Chernoff bounds," *Information Processing Letters*, vol. 33, pp. 305-308, 1990.