CS281B/Stat241B (Spring 2008) Statistical Learning Theory **Lecture: 13**

Concentration Inequalities: Hoeffding and McDiarmid

Lecturer: Peter Bartlett *Scribe: Galen Reeves*

1 Recap

For a function $f \in F$ the empirical risk function is

$$\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i)$$

and the empirical risk minimizing function is

$$f_n = \arg\min_{f \in F} \hat{R}(f).$$

Often, we are interested in the true performance of f_n

$$R(f_n) = \mathbb{E}[\ell(f_n(X), Y)].$$

For example, this might correspond to the performance of the hinge-loss cost function using an SVM.

Previously, we showed that for finite classes F the statement

$$\forall f \in F \qquad P\left(R(f) \geq \hat{R}(f) + \epsilon\right) \leq e^{-c\epsilon^2 n}$$

implies that

$$P\left(\exists f \in F \ : \ R(f) \geq \hat{R}(f) + \epsilon\right) \leq |F| e^{-c\epsilon^2 n}$$

or equivalently, w.p. $\geq 1 - \delta$, $\forall f \in F$,

$$R(f) \leq \hat{R}(f) + c\sqrt{\frac{\log |F|}{n} + \frac{\log(1/\delta)}{n}}$$

1.1 Recap of Inequalities

We want to show that the expected risk $R(f)$ is close to the sample average $\hat{R}(f)$. To do so we use concentration inequalities; two simple inequalities are the following:

- Markov's Inequality: For $X \geq 0$, $P(X \geq t) \leq \frac{\mathbb{E}X}{t}$
- Chebyshev's Inequality: $P\left(|X - \mathbb{E}X| \geq t\right) \leq \frac{\mathrm{Var}(X)}{t^2}$

Although the above inequalities are very general, we want bounds which give us stronger (exponential) convergence. This lecture introduces Hoeffding's Inequality for sums of independent bounded variables and shows that exponential convergence can be achieved. Then, a generalization of Hoeffding's Inequality called McDiarmid's (or Bounded Differences or Hoeffding/Azuma) Inequality is presented.

2 Hoeffding's Inequality

Consider the sum $S_n = \sum_{i=1}^n X_i$ of independent random variables, $X_1, \cdots, X_n$. Then, we have for all $s > 0$,

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

where the inequality is true through the application of Markov's Inequality, and the second equality follows from the independence of X_i. Note that $\mathbb{E}e^{s(X_i - \mathbb{E}X_i)}$ is the moment-generating function of $X_i - \mathbb{E}X_i$.

Lemma 2.1 (Hoeffding)**.** For a random variable X with $\mathbb{E}X = 0$ and $a \leq X \leq b$ then for $s > 0$

$$\mathbb{E}e^{sX} \leq e^{s^2(b-a)^2/8}$$

Proof. Note that e^{sx} is convex in x and is thus uniformly bounded as

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

Taking expectation yields

$$\mathbb{E}e^{sX} \leq \frac{be^{sa} - ae^{sb}}{b-a},$$

and taking the Taylor series expansion of $\log \mathbb{E}e^{sX}$ about $s = 0$ reveals that

$$\log \mathbb{E}e^{sX} \leq \frac{s^2(b-a)^2}{8}.$$

□

By combining (1) and Lemma 2.1 we see that if $X_i \in [a_i, b_i]$ then

$$\begin{aligned} P\left(S_n - \mathbb{E}S_n \geq t\right) &\leq \inf_{s>0}\left(e^{-st}\prod_{i=1}^n e^{s^2(b_i-a_i)^2/8}\right) \\ &= \inf_{s>0}\exp\left(-st + \frac{s^2}{8}\sum_{i=1}^n (b_i - a_i)^2\right) \\ &= \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \end{aligned}$$

where the minimizing value of s is given by $s^2 = 4t/\sum_{i=1}^n (b_i - a_i)^2$. Thus we have just proved the following theorem.

Theorem 2.2 (Hoeffding's Inequality)**.** For bounded random variables $X_i \in [a_i, b_i]$ where $X_1, \cdots, X_n$ are independent, then

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

and

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

Remark. With Hoeffding's Inequalities the tails of the error probability are starting to look more Gaussian, i.e. they decay exponentially with t^2, and they correspond to the worst case variance given the bounds a_i and b_i.

Example. Consider minimizing $\hat{R}(f)$ with $\ell(\hat{y}, y) \in [0, 1]$. We have

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

This is what the Central Limit Theorem would suggest with $\sigma^2 = \frac{1}{2}(1-\frac{1}{2})$, which corresponds to the variance of a Bernoulli(1/2) variable.

3 McDiarmid's Inequality

We now look at a generalization of Hoeffding's inequality where the quantity of interest is some function of the data, i.e. $S_n = \phi(X_1, X_2, \cdots, X_n)$. Some restrictions on ϕ are required to get exponential bounds. The following theorem makes this precise (The critical property of ϕ required is that each component (X_i) cannot influence the outcome too much).

Theorem 3.1 (McDiarmid's (or Bounded Differences or Hoeffding/Azuma) Inequality)**.** Consider independent random variables $X_1, \cdots, X_n \in \mathcal{X}$ and a mapping $f : \mathcal{X}^n \to \mathbb{R}$. If, for all $i \in \{1, \cdots, n\}$, and for all $x_1, \cdots, x_n, x_i' \in \mathcal{X}$, the function ϕ satisfies

$$|\phi(x_1, \cdots, x_{i-1}, x_i, x_{i+1}, \cdots, x_n) - \phi(x_1, \cdots, x_{i-1}, x_i', x_{i+1}, \cdots, x_n)| \leq c_i \tag{2}$$

then

$$P\left(\phi(X_1, \cdots, X_n) - \mathbb{E}\phi \geq t\right) \geq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n} c_i^2}\right)$$

and

$$P\left(\phi(X_1, \cdots, X_n) - \mathbb{E}\phi \leq -t\right) \geq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n} c_i^2}\right)$$

Before looking at the proof of Theorem 3.1 we consider some examples.

Example. Hoeffding's Inequality

Example. Leave-one-out error estimate of 1-nearest neighbor. Given a data set, we label based upon the closest point in the data. For a leave-one-out error estimate, for $i = 1, \cdots n$ we classify the i^{th} sample based on all the other data to create the estimate $\hat{Y}_i$. Thus the error for each sample is $\epsilon_i = \ell(\hat{Y}_i, Y_i)$, and the total error estimate is given by

$$\phi\left((X_1, Y_1), \cdots, (X_n, Y_n)\right) = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i$$

If we change (X_i, Y_i) to (X_i', Y_i') then we affect both ϵ_i and potentially other ϵ_j's for $j \neq i$. If we assume that the geometry is such that at most k other sample errors can be affected, then the total affect is less than k/n and ϕ obeys the necessary condition.

Example. Consider the traveling salesman problem, where we desire to find the shortest path through all cities. We note that by changing one path, the total distance can only increase by at most twice the diameter. Thus, for random choice of the city locations in a bounded region, the length of the shortest path is tightly concentrated about its expectation.

Example. Consider

$$\phi(X_1, \cdots, X_n) = \max_{f \in F} |\mathbb{E}f - \mathbb{E}_n f|$$

where $\mathbb{E}_n f$ is the sample average. If for all $f : \mathcal{X} \to [a, b]$, then $c_i = (b-a)/n$. This means that

$$P\left(\sup_{f\in F} |\mathbb{E}f - \mathbb{E}_n f| - \mathbb{E}\left(\sup_{f\in F} |\mathbb{E}f - \mathbb{E}_n f|\right) \geq t\right) \leq e^{\frac{-2nt^2}{(b-a)^2}}.$$

We now give the proof of McDiarmid's Inequality

Proof. We will think of a martingale sequence. We define

$$V_i = \mathbb{E}\left[\phi | X_1, \cdots, X_i\right] - \mathbb{E}\left[\phi | X_1, \cdots, X_{i-1}\right],$$

and note that $\mathbb{E}V_i = 0$ and

$$\sum_{i=1}^{n} V_i = \mathbb{E}[\phi | X_1, \cdots, X_n] - \mathbb{E}\phi = \phi(X_1, \cdots, X_n) - \mathbb{E}\phi. \tag{3}$$

We further define upper and lower bounds

$$L_i = \inf_x \mathbb{E}\left[\phi | X_1, \cdots, X_{i-1}, x\right] - \mathbb{E}\left[\phi | X_1, \cdots, X_{i-1}\right]$$
$$U_i = \sup_x \mathbb{E}\left[\phi | X_1, \cdots, X_{i-1}, x\right] - \mathbb{E}\left[\phi | X_1, \cdots, X_{i-1}\right]$$

and note that $L_i \leq V_i \leq U_i$. We can use the independence of X_i to show that $U_i - L_i \leq c_i$.

As in Hoeffding, we have

$$P(\phi - \mathbb{E}\phi \geq t) \leq \inf_{s>0} e^{-st} \mathbb{E}\left(\prod_{i=1}^{n} e^{sV_i}\right).$$

Furthermore,

$$\begin{aligned}
\mathbb{E}\left(\prod_{i=1}^{n} e^{sV_i}\right) &= \mathbb{E}\,\mathbb{E}\left[\prod_{i=1}^{n-1} e^{sV_i} s^{sV_n} | X_1, \cdots, X_{n-1}\right] \\
&= \mathbb{E}\prod_{i=1}^{n-1} e^{sV_i} \mathbb{E}\left[e^{sV_n} | X_1, \cdots, X_{n-1}\right] \\
&\leq \mathbb{E}\left(\prod_{i=1}^{n-1} e^{sV_i}\right) e^{s^2 c_n^2/8} \\
&\cdots \\
&\leq \exp(s^2 \sum_{i=1}^{n} c_i^2/8)
\end{aligned}$$

More details of this proof can be found on the course website. □

4 Concluding Remarks

So far, the bound we have discussed corresponds to the worst case variance for the given constraints. We can say more if we have some bound on the variance.

Example. Let $R(f) = \mathbb{E}(Y - f(X))^2$ and then

$$R(f) - R(f^*) = \mathbb{E}\left[(Y - f(X))^2 - (Y - f^*(X))^2\right],$$

where $f^* = \arg\min_{f \in F} R(f)$. If F is convex, then the variance decreases as the risk decreases.

Also, we see that summations are the worst case for bounded differences inequalities (Hoeffing's Inequality gives the same result).

Lastly, for

$$\hat{f} = \arg\min_{f \in F} \hat{R}(f)$$
$$f^* = \arg\min_{f \in F} R(f)$$

we may compare $R(\hat{f})$ and $R(f^*)$. We have

$$\begin{aligned} R(\hat{f}) - R(f^*) &= R(\hat{f}) - \hat{R}(\hat{f}) && (4) \\ &\quad + \hat{R}(\hat{f}) - \hat{R}(f^*) \\ &\quad + \hat{R}(f^*) - R(f^*) \\ &\leq 2 \sup_{f \in F} |R(f) - \hat{R}(f)| && (5) \end{aligned}$$

because the second term on the right hand side of (4) is non-positive.