Lecture Notes 3
Uniform Bounds

1 Introduction

Recall that, if $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ and $\widehat{p}_n = n^{-1}\sum_{i=1}^n X_i$ then, from Hoeffding's inequality,

$$\mathbb{P}(|\widehat{p}_n - p| > \epsilon) \le 2e^{-2n\epsilon^2}.$$

Sometimes we want to say more than this.

Example 1 *Suppose that $X_1, \ldots, X_n$ have cdf F. Let*

$$F_n(t) = \frac{1}{n}\sum_{i=1}^n I(X_i \le t).$$

We call F_n the **empirical cdf**. *How close is F_n to F? From Hoeffding's inequality, we have for each t, that*

$$\mathbb{P}(|F_n(t) - F(t)| > \epsilon) \le 2e^{-2n\epsilon^2}.$$

But how big is $\sup_t |F_n(t) - F(t)|$? We would like a bound of the form

$$\mathbb{P}\left(\sup_t |F_n(t) - F(t)| > \epsilon\right) \le \text{ something small.}$$

Example 2 *Suppose that $X_1, \ldots, X_n \sim P$. Let*

$$P_n(A) = \frac{1}{n}\sum_{i=1}^n I(X_i \in A).$$

How close is $P_n(A)$ to $P(A)$? That is, how big is $|P_n(A) - P(A)|$? From Hoeffding's inequality,

$$\mathbb{P}(|P_n(A) - P(A)| > \epsilon) \le 2e^{-2n\epsilon^2}.$$

But that is only for one set A. How big is $\sup_{A\in\mathcal{A}} |P_n(A) - P(A)|$ for a class of sets $\mathcal{A}$? We would like a bound of the form

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| > \epsilon\right) \le \text{ something small.}$$

Example 3 *(Classification.) Suppose we observe data $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $Y_i \in \{0, 1\}$. Let (X, Y) be a new pair. Suppose we observe X. Now we want to predict Y. A* *classifier h is a function $h(x)$ which takes values in $\{0,1\}$. When we observe X we predict Y with $h(X)$. The classification error, or risk, is the probability of an error:*

$$R(h) = \mathbb{P}(Y \neq h(X)).$$

The training error is the fraction of errors on the observed data $(X_1, Y_1), \ldots, (X_n, Y_n)$:

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

By Hoeffding's inequality,

$$\mathbb{P}(|\widehat{R}(h) - R(h)| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

How do we choose a classifier? One way is to start with a set of classifiers $\mathcal{H}$. Then we define $\widehat{h}$ to be the member of $\mathcal{H}$ that minimizes the training error. Thus

$$\widehat{h} = \text{argmin}_{h \in \mathcal{H}} \widehat{R}(h).$$

An example is the set of linear classifiers. Suppose that $x \in \mathbb{R}^d$. A linear classifier has the form $h(x) = 1$ of $\beta^T x \geq 0$ and $h(x) = 0$ of $\beta^T x < 0$ where $\beta = (\beta_1, \ldots, \beta_d)^T$ is a set of parameters.

Although $\widehat{h}$ minimizes $\widehat{R}(h)$, it does not minimize $R(h)$. Let h_* minimize the true error $R(h)$. A fundamental question is: how close is $R(\widehat{h})$ to $R(h_*)$? We will see later than $R(\widehat{h})$ is close to $R(h_*)$ if $\sup_h |\widehat{R}(h) - R(h)|$ is small. So we want

$$\mathbb{P}\left(\sup_h |\widehat{R}(h) - R(h)| > \epsilon\right) \leq \quad \text{something small.}$$

More generally, we can state out goal as follows. Let $\mathcal{A}$ be a class of sets. We want a bound of the form

$$\mathbb{P}\left(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \epsilon\right) \leq c_1 \kappa(\mathcal{A}) e^{-c_2 n \epsilon^2}$$

where $P_n(A) = n^{-1} \sum_{i=1}^n I(X_i \in A)$. Bounds like these are called *uniform bounds* since they hold uniformly over a class of functions or over a class of sets.

2 Finite Classes

Let $\mathcal{A} = \{A_1, \ldots, A_N\}$. We will make use of the *union bound*. Recall that

$$\mathbb{P}\left(B_1 \bigcup \cdots \bigcup B_N\right) \leq \sum_{j=1}^{N} \mathbb{P}(B_j).$$

Let B_j be the event that $|P_n(A_j) - P(A_j)| > \epsilon$. From Hoeffding's inequality, $\mathbb{P}(B_j) \leq 2e^{-2n\epsilon^2}$. Then

$$\begin{aligned}\mathbb{P}\left(\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| > \epsilon\right) &= \mathbb{P}\left(B_1 \bigcup \cdots \bigcup B_N\right) \\ &\leq \sum_{j=1}^{N} \mathbb{P}(B_j) \leq \sum_{j=1}^{N} 2e^{-n\epsilon^2} = 2Ne^{-2n\epsilon^2}.\end{aligned}$$

Thus we have shown that

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| > \epsilon\right) \leq 2\kappa e^{-n\epsilon^2}$$

where $\kappa = |\mathcal{A}|$.

To extend these ideas to infinite classes like $\mathcal{A} = \{(-\infty, t] : \ t \in \mathbb{R}\}$ we need to introduce a few more concepts.

3 Shattering

Let $\mathcal{A}$ be a class of sets. Some examples are:

1. $\mathcal{A} = \{(-\infty, t] : \ t \in \mathbb{R}\}$.
2. $\mathcal{A} = \{(a, b) : \ a \leq b\}$.
3. $\mathcal{A} = \{(a, b) \cup (c, d) : \ a \leq b \leq c \leq d\}$.
4. $\mathcal{A} =$ all discs in $\mathbb{R}^d$.
5. $\mathcal{A} =$ all rectangles in $\mathbb{R}^d$.
6. $\mathcal{A} =$ all half-spaces in $\mathbb{R}^d = \{x : \ \beta^T x \geq 0\}$.
7. $\mathcal{A} =$ all convex sets in $\mathbb{R}^d$.

Let $F = \{x_1, \ldots, x_n\}$ be a finite set. Let G be a subset of F. Say that $\mathcal{A}$ **picks out** G if

$$A \cap F = G$$

for some $A \in \mathcal{A}$. For example, let $\mathcal{A} = \{(a, b) : \ a \leq b\}$. Suppose that $F = \{1, 2, 7, 8, 9\}$ and $G = \{2, 7\}$. Then $\mathcal{A}$ picks out G since $A \cap F = G$ if we choose $A = (1.5, 7.5)$ for example.

Let $S(\mathcal{A}, F)$ be the number of these subsets picked out by $\mathcal{A}$. Of course $S(\mathcal{A}, F) \leq 2^n$.

Example 4 *Let* $\mathcal{A} = \{(a, b) : \ a \leq b\}$ *and* $F = \{1, 2, 3\}$. *Then* $\mathcal{A}$ *can pick out:*

$$\emptyset, \ \{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 2, 3\}.$$

So $s(\mathcal{A}, F) = 7$. *Note that* $7 < 8 = 2^3$. *If* $F = \{1, 6\}$ *then* $\mathcal{A}$ *can pick out:*

$$\emptyset, \ \{1\}, \{6\}, \{1, 6\}.$$

In this case $s(\mathcal{A}, F) = 4 = 2^2$.

We say that F is **shattered** if $s(\mathcal{A}, F) = 2^n$ where n is the number of points in F.

Let $\mathcal{F}_n$ denote all finite sets with n elements.

Define the **shatter coefficient**

$$s_n(\mathcal{A}) = \sup_{F \in \mathcal{F}_n} s(\mathcal{A}, F).$$

Note that $s_n(\mathcal{A}) \leq 2^n$.

The following theorem is due to Vapnik and Chervonenis. The proof is beyond the scope of the course. (If you take 10-702/36-702 you will learn the proof.)

Theorem 5 *Let* $\mathcal{A}$ *be a class of sets. Then*

$$\mathbb{P}\left(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \epsilon \right) \leq 8 \ s_n(\mathcal{A}) \ e^{-n\epsilon^2/32}. \tag{1}$$

This partly solves one of our problems. But, how big can $s_n(\mathcal{A})$ be? Sometimes $s_n(\mathcal{A}) = 2^n$ for all n. For example, let $\mathcal{A}$ be all polygons in the plane. Then $s_n(\mathcal{A}) = 2^n$ for all n. But, in many cases, we will see that $s_n(\mathcal{A}) = 2^n$ for all n up to some integer d and then $s_n(\mathcal{A}) < 2^n$ for all $n > d$.

Class $\mathcal{A}$	VC dimension $V_{\mathcal{A}}$
$\mathcal{A} = \{A_1, \ldots, A_N\}$	$\leq \log_2 N$
Intervals $[a, b]$ on the real line	2
Discs in $\mathbb{R}^2$	3
Closed balls in $\mathbb{R}^d$	$\leq d + 2$
Rectangles in $\mathbb{R}^d$	$2d$
Half-spaces in $\mathbb{R}^d$	$d + 1$
Convex polygons in $\mathcal{R}^2$	∞
Convex polygons with d vertices	$2d + 1$

Table 1: The VC dimension of some classes $\mathcal{A}$.

Example 6 *Let $\mathcal{A} = \{(a, b) : \ a, b \in \mathbb{R}, a \leq b\}$. Then we have:*

n	2^n	s_n
1	*2*	*2*
2	*4*	*4*
3	*8*	*7*
4	*16*	*11*
$\vdots$	$\vdots$	$\vdots$

So $s_n = 2^n$ for $n = 1, 2$. For $n > 2$ we have $s_N < 2^n$.

The **Vapnik-Chervonenkis (VC) dimension** is

$$d = d(\mathcal{A}) = \quad \text{largest n such that } s_n(\mathcal{A}) = 2^n.$$

In other words, d is the size of the largest set that can be shattered.

Thus, $s_n(\mathcal{A}) = 2^n$ for all $n \leq d$ and $s_n(\mathcal{A}) < 2^n$ for all $n > d$. The VC dimensions of some common examples are summarized in Table 1. Now here is an interesting question: for $n > d$ how does $s_n(\mathcal{A})$ behave? It is less than 2^n but how much less?

Theorem 7 (Sauer's Theorem) *Suppose that $\mathcal{A}$ has finite VC dimension d. Then, for all $n \geq d$,*

$$s(\mathcal{A}, n) \leq (n + 1)^d. \tag{2}$$

Sauer's Theorem is very surprising. It says there is a phase transition from exponential to polynomial. We conclude that:

Theorem 8 *Let $\mathcal{A}$ be a class of sets with VC dimension $d < \infty$. Then*

$$\mathbb{P}\left(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \epsilon\right) \leq 8\ (n+1)^d\ e^{-n\epsilon^2/32}. \tag{3}$$

Example 9 *Let's return to our first example. Suppose that $X_1, \ldots, X_n$ have cdf F. Let*

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq t).$$

We would like to bound $\mathbb{P}(\sup_t |F_n(t) - F(t)| > \epsilon)$. Notice that $F_n(t) = P_n(A)$ where $A = (-\infty, t]$. Let $\mathcal{A} = \{(-\infty, t] :\ t \in \mathbb{R}\}$. This has VC dimension $d = 1$. So

$$\mathbb{P}(\sup_t |F_n(t) - F(t)| > \epsilon) = \mathbb{P}\left(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > \epsilon\right) \leq 8\ (n+1)\ e^{-n\epsilon^2/32}.$$

In fact, there is a tighter bound in this case called the DKW (Dvoretsky-Kiefer-Wolfowitz) inequality:

$$\mathbb{P}(\sup_t |F_n(t) - F(t)| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$