STATISTICAL LEARNING THEORY

In these notes we will cover basic performance guarantees for classification.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be iid realizations of (X, Y), where $X \in \mathbb{R}^d$, $Y \in \{0, 1\}$.

Let $f: \mathbb{R}^d \to \{0,1\}$ be a classifier. Define the <u>risk</u>

$$R(f) := \Pr\{f(X) \neq Y\}$$

$$= E\left[\mathbb{1}_{\{f(X) \neq Y\}}\right]$$

and the <u>empirical risk</u>

$$\hat{R}_n(f) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{f(X_i) \neq Y_i\}}$$

of f. Notice

(A) $$n\hat{R}_n(f) \sim$$

Hoeffding's Inequality

Theorem Let $Z_1, \ldots, Z_n$ be independent, bounded RVs such that $Pr\{Z_i \in [a_i, b_i]\} = 1$. Set $S_n = \sum_{i=1}^{n} Z_i$. Then $\forall t > 0$,

$$Pr\{S_n - ES_n \geq t\} \leq e^{-2t^2 / \sum_{i=1}^{n} (b_i - a_i)^2}$$

and

$$Pr\{S_n - ES_n \leq -t\} \leq e^{-2t^2 / \sum_{i=1}^{n} (b_i - a_i)^2}$$

∎

Remarks

- We may combine the two statements to obtain

$$Pr\{|S_n - ES_n| \geq t\} \leq 2e^{-2t^2 / \sum_{i=1}^{n} (b_i - a_i)^2}$$

- In the special case where Z_i are iid Bernoulli(p), then $b_i = 1$, $a_i = 0$, and S_n is binom(n, p), and we recover <u>Chernoff's bound</u>:

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

- Hoeffding's is an example of a <u>concentration inequality</u>.

Proof

LEMMA 2.1. *Let V be a random variable with* $\mathbf{E}V = 0$, $a \le V \le b$. *Then for* $s > 0$:

$$\mathbf{E}\{e^{sV}\} \le e^{s^2(b-a)^2/8}.$$

← Devroye and Lugosi, Combinatorial Methods in Density Estimation, Springer 2001.

PROOF. Note that by convexity of the exponential function

$$e^{sv} \le \frac{v-a}{b-a}e^{sb} + \frac{b-v}{b-a}e^{sa} \qquad \text{for } a \le v \le b.$$

Exploiting $\mathbf{E}V = 0$, and introducing the notation $p = -a/(b-a)$, we get

$$\begin{aligned}\mathbf{E}\{e^{sV}\} &\le \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} \\ &= (1-p+pe^{s(b-a)})e^{-ps(b-a)} \\ &\stackrel{\text{def}}{=} e^{\phi(u)},\end{aligned}$$

where $u = s(b-a)$ and $\phi(u) = -pu + \log(1-p+pe^u)$. But by straightforward calculation it is easy to see that the derivative of ϕ is

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

and therefore $\phi(0) = \phi'(0) = 0$. Moreover,

$$\phi''(u) = \frac{p(1-p)e^{-u}}{(p+(1-p)e^{-u})^2} \le 1/4.$$

Thus, by Taylor's theorem, for some $\theta \in [0, u]$:

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2}\phi''(\theta) \le \frac{u^2}{8} = \frac{s^2(b-a)^2}{8}. \quad \square$$

Lemma (Markov's Inequality) If U is a nonnegative random variable, then for all $t > 0$,

$$\Pr\{U \ge t\} \le \frac{EU}{t}$$

Proof:

$$\begin{aligned}\Pr\{U \ge t\} &= E\left[1_{\{U \ge t\}}\right] \\ &\le E\left[\frac{U}{t}1_{\{U \ge t\}}\right] \\ &= \frac{1}{t}E\left[U 1_{\{U \ge t\}}\right] \le \frac{1}{t}E[U]\end{aligned}$$

Now, for any $s > 0$, we have

$$\Pr\{S_n - ES_n \geq t\} = \Pr\{s(S_n - ES_n) \geq st\}$$

$$= \Pr\{e^{s(S_n - ES_n)} \geq e^{st}\}$$

$$\leq e^{-st} \cdot E\left[e^{s(S_n - ES_n)}\right] \qquad \text{(Markov's inequality)}$$

$$= e^{-st} E\left[e^{s \cdot \sum_{i=1}^{n}(Z_i - EZ_i)}\right]$$

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

$$= e^{-st} \prod_{i=1}^{n} E\left[e^{s(Z_i - EZ_i)}\right] \qquad \text{(independence)}$$

$$\leq e^{-st} \prod_{i=1}^{n} e^{s^2 (b_i - a_i)^2 / 8} \qquad \text{(by the lemma)}$$

$$= e^{-st} e^{s^2 \sum_{i=1}^{n} (b_i - a_i)^2 / 8}$$

$$= e^{-2t^2 / \sum_{i=1}^{n} (b_i - a_i)^2} \qquad \left(s = 4t \big/ \Sigma (b_i - a_i)^2\right)$$

$\blacksquare$

Returning to classification, by Hoeffding's / Chernoff's bound we know that for any classifier f

$$\Pr\left\{ \hat{R}_n(f) \geq R(f) + \epsilon \right\} \leq e^{-2n\epsilon^2},$$

which $\to 0$ exponentially fast as $n \to \infty$ (ϵ fixed).

Uniform Deviation Bounds

In reality, we don't know the best classifier a priori. One way to overcome this is to prove a performance guarantee that holds for many classifiers simultaneously.

Let $\mathcal{F} = \{f_1, \ldots, f_M\}$.

<u>Theorem</u> For any $\epsilon > 0$,

$$\Pr\left\{ \max_{f \in \mathcal{F}} \left| \hat{R}_n(f) - R(f) \right| \geq \epsilon \right\} \leq 2Me^{-2n\epsilon^2}$$

Proof:

$$= \Pr\left\{ \text{for some } m,\ \left| \hat{R}_n(f_m) - R(f) \right| \geq \epsilon \right\}$$

$$\leq \sum_{m=1}^{m} \Pr\left\{ \left| \hat{R}_n(f_m) - R(f_m) \right| \geq \epsilon \right\} \quad \left(\text{union bound}\right)$$

$$\leq 2Me^{-n\epsilon^2}.$$

Empirical Risk Minimization

Let's turn this result into a classification rule with a performance guarantee.

Denote

$$R(\mathcal{F}) = \inf_{f \in \mathcal{F}} R(f)$$

and define the rule

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \hat{R}_n(f)$$

Theorem | Let $\epsilon > 0$. With probability at least $1 - 2Me^{-2n\epsilon^2}$,

$$R(\hat{f}_n) \le R(\mathcal{F}) + 2\epsilon.$$

Proof: With prob $\geq 1 - 2Me^{-2n\epsilon^2}$, we have

$\sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| < \epsilon$. In this event, for any f,

$$R(\hat{f}_n) \le \hat{R}_n(\hat{f}_n) + \epsilon$$

$$\le \hat{R}_n(f) + \epsilon \qquad (\text{def of } \hat{f}_n)$$

$$\le R(f) + \epsilon$$

Since f is arbitrary, the result follows. $\boxtimes$

Key point | The above result is distribution free, meaning it makes no assumptions on the distribution of (X, Y).

Note that the proof did not depend on $\mathcal{F}$ being finite, only on the existence of a uniform deviation bound for $\mathcal{F}$. Such bounds also exist in cases where $\mathcal{F}$ is infinite.

VC Bounds

Let $\mathcal{F}$ now be an arbitrary collection of classifiers, perhaps uncountably infinite.

Definition | Given points $x_1, \dots, x_n \in \mathbb{R}^d$, let $N_{\mathcal{F}}(x_1, \dots, x_n)$ be the number of distinct vectors $(f(x_1), \dots, f(x_n)) \in \{0,1\}^n$ as f ranges over $\mathcal{F}$.

Now define the n^{th} shatter coefficient of $\mathcal{F}$

$$S(\mathcal{F}, n) = \max_{x_1, \dots, x_n} N_{\mathcal{F}}(x_1, \dots, x_n).$$

Cleary $S(\mathcal{F}, n) \le 2^n$ for every n. If $S(\mathcal{F}, n) = 2^n$, then $N_{\mathcal{F}}(x_1, \dots, x_n) = 2^n$ for some $x_1, \dots, x_n$, and we say $\mathcal{F}$ <u>shatters</u> $x_1, \dots, x_n$.

<u>Definition</u> Assume $|\mathcal{F}| \ge 1$. The largest k such that $S(\mathcal{F}, k) = 2^k$ is called the Vapnik-Chervonenkis (VC) dimension of $\mathcal{F}$. If no such k exists, we set $VCdim(\mathcal{F}) = \infty$.

It can be shown that if $\mathcal{F}$ has VC dim. $V \ge 2$, then $\forall n$,

$$S(\mathcal{F}, n) \le n^V$$

The following result is due to Vapnik & Chervonenkis.

<u>Theorem</u> For any $\epsilon > 0$

$$P\left\{ \sup_{f \in \mathcal{F}} |\hat{R}_n(f) - R(f)| > \epsilon \right\} \le 8 S(\mathcal{F}, n) e^{-n\epsilon^2/32}$$

Proof: See Devroye, Györfi, and Lugosi, A Probabilistic Theory of Pattern Recognition.

Again, this is a <u>distribution-free</u> result.

<u>Corollary</u> For empirical risk minimimization,

$$P\left\{ R(\hat{f}_n) \geq \inf_{f \in \mathcal{F}} R(f) + 2\epsilon \right\} \leq 8 S(\mathcal{F}, n) e^{-n\epsilon^2/32}$$

<u>Key Point</u> If VC dim $(\mathcal{F}) = V < \infty$, and

we use $S(\mathcal{F}, n) \leq n^V$, then we see that

the "failure probability" is bounded by

$$8 n^V e^{-n\epsilon^2/32}$$

which $\to 0$ exponentially fast as $n \to \infty$ (ϵ fixed).

So which $\mathcal{F}$ have <u>finite</u> VC dimension?

VC Classes

<u>Rectangles</u> Suppose $\mathcal{F}$ is the collection of

classifiers of the form $\mathbb{1}_{\{x \in R\}}$,

where R ranges over all rectangles in $\mathbb{R}^d$.

What is the VC dim. of $\mathcal{F}$?

Claim: $V = 2d$.

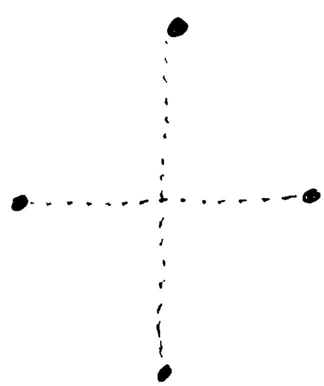
Need to show (a) $\exists$ $2d$ points shattered by $\mathcal{F}$,
(b) $\mathcal{F}$ cannot shatter any collection of $> 2d$ points.

For (a), take the $2d$ points

$$(1,0,\dots,0),\ (0,1,0,\dots,0),\ \dots,\ (0,\dots,0,1)$$
$$(-1,0,\dots,0),\ (0,-1,0,\dots,0),\ \dots,\ (0,\dots,0,-1)$$

For (b), consider any set of $> 2d$ points.
Then there exists a subset of at most $2d$ "extreme" points, that are the min or max along at least on dimension. Clearly no R contains all these points but not the others.

$d=2$



The following general result allows us to bound VC dims for many classes.

<u>Theorem</u> Let G be a vector space of functions with $\dim(G) = r$. If $\mathcal{F}$ is the set of classifiers of the form

$$x \longmapsto 1_{\{g(x) \geq 0\}}, \quad g \in G$$

then $\mathrm{VCdim}(\mathcal{F}) \leq r$.

PROOF. It suffices to show that no set of size $m = 1 + r$ can be shattered by sets of the form $\{x : g(x) \geq 0\}$. Fix m arbitrary points $x_1, \ldots, x_m$, and define the linear mapping $L : \mathcal{G} \to \mathcal{R}^m$ as ← DGL

$$L(g) = (g(x_1), \ldots, g(x_m)).$$

Then the image of $\mathcal{G}$, $L(\mathcal{G})$, is a linear subspace of $\mathcal{R}^m$ of dimension not exceeding the dimension of $\mathcal{G}$, that is, $m - 1$. Then there exists a nonzero vector $\gamma = (\gamma_1, \ldots, \gamma_m) \in \mathcal{R}^m$, that is orthogonal to $L(\mathcal{G})$, that is, for every $g \in \mathcal{G}$

$$\gamma_1 g(x_1) + \ldots + \gamma_m g(x_m) = 0.$$

We can assume that at least one of the γ_i's is negative. Rearrange this equality so that terms with nonnegative γ_i stay on the left-hand side:

$$\sum_{i:\gamma_i \geq 0} \gamma_i g(x_i) = \sum_{i:\gamma_i < 0} -\gamma_i g(x_i).$$

Now, suppose that there exists a $g \in \mathcal{G}$ such that the set $\{x : g(x) \geq 0\}$ picks exactly the x_i's on the left-hand side. Then all terms on the left-hand side are nonnegative, while the terms on the right-hand side must be negative, which is a contradiction, so $x_1, \ldots, x_m$ cannot be shattered, and the proof is completed. □

Linear Classifiers

Suppose $\mathcal{F}$ = all f of the form $f(x) = \mathrm{sign}\{w^T x + b\}$, $w \in \mathbb{R}^d$, $b \in \mathbb{R}$.

What is V?

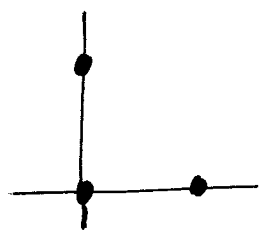
Claim: $V = d+1$.

By the above theorem, we have $V \le d+1$, taking G to be the space spanned by

$$\varphi^{(1)}(x) = x^{(1)}, \ \ldots, \ \varphi^{(d)}(x) = x^{(d)}, \ \varphi^{(d+1)}(x) = 1$$

Furthermore, $\mathcal{F}$ shatters

$$(0,\ldots,0), (1,0,\ldots,0), (0,1,0,\ldots,0), \ \ldots, \ (0,\ldots,0,1)$$



Exercise | Determine the VC dimension of

- $\mathcal{F}$ = all classifiers of the form

  $$f(x) = \mathbb{1}_{\{x \in B(a,b)\}}$$

  where $B(a,b) = \{x : \|x - a\| \le b\}$, $a \in \mathbb{R}^d, b \in \mathbb{R}$.

- $\mathcal{F}$ = all classifiers of the form

  $$f(x) = \mathbb{1}_{\{x \in C\}}$$

  where C is a convex polygon in $\mathbb{R}^2$.

Neural Networks

For neural networks with k hidden units and w tunable weights, Karpinski and Macintyre (1994) showed

$$V \le \frac{kw(kw-1)}{2} + w(1+2k) + w(1+3k)\log(3w+6kw+3),$$

assuming the standard sigmoid function.

Summary

The above results tell us performance guarantees for many class. For example, for the empirical risk minimizing linear classifier $\hat{f}_n$

$$\Pr\left\{ R(\hat{f}_n) - \inf_{f \in \mathcal{F}} R(f) \ge 2\epsilon \right\} \le 8 n^{d+1} e^{-n\epsilon^2/32}$$

Unfortunately, empirical risk minimization is (provably) not computational feasible over most classes of interest. Therefore these results are largely of theoretical importance. Furthermore, the bounds tend to be very loose (often > 1) in practice.

PAC Learning and Sample Complexity

An algorithm $\hat{f}_n$ is said to be an (ϵ,δ)-learning algorithm for $\mathcal{F}$ if $\exists$ a function $N(\epsilon,\delta)$ such that, $\forall \epsilon,\delta > 0$,

$$n \geq N(\epsilon,\delta) \implies \Pr\left\{ R(\hat{f}_n) - \inf_{f\in\mathcal{F}} R(f) > \epsilon \right\} \leq \delta$$

for all distributions of (X,Y).

Terminology

- $N(\epsilon,\delta)$ is called the sample complexity
- $\mathcal{F}$ is said to be uniformly learnable
- $\hat{f}_n$ is said to be "probably approximately correct" (PAC)

We have seen that if $\mathrm{VCdim}(\mathcal{F}) < \infty$, then

- $\mathcal{F}$ is uniformly learnable
- ERM is an (ϵ,δ)-learning algorithm
- $N(\epsilon,\delta) = O\left(\max\left\{\frac{V}{\epsilon^2}\log\frac{V}{\epsilon^2}, \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right\}\right)$

$$\Updownarrow$$

$$8n^V e^{-n\epsilon^2/128} \leq \delta$$

$\boxed{\text{Key}}$ A. $n\hat{R}_n(f) \sim \mathrm{binom}(n, R(f))$