Rademacher complexity properties 2: finite classes and margin losses

Administrative/Meta:

- Homework 2 is out; watch out for problems 3 and 4.
- Project writeups will probably need to only be 2 pages (or longer if you like); I'll update on the webpage once I'm 100% sure of this.

Overview

As we've discussed in the past few lectures, symmetrization / Rademacher complexity give us the following bound.

Theorem. Let functions $\mathcal{F}$ be given with $|f(z)| \leq c$ almost surely for every $f \in \mathcal{F}$. With probability $\geq 1-\delta$ over an i.i.d. draw $S := (Z_1, \ldots, Z_n)$, every $f \in \mathcal{F}$ satisfies

$$\mathbb{E}f \leq \hat{\mathbb{E}}f + 2\mathrm{Rad}(\mathcal{F}_{|S}) + 3c\sqrt{\frac{2}{n}\ln\left(\frac{2}{\delta}\right)}.$$

To make this meaningful to machine learning, we need to replace $\mathbb{E}f$ with some form of risk. Today will discuss three choices.

1. $\mathcal{R}_\ell$ where ℓ is Lipschitz. We covered this last time but will recap a little.
2. $\mathcal{R}_z(f) := \Pr[f(X) \neq Y]$; for this we'll use finite classes and discuss **shatter coefficients** and **VC dimension**.
3. $\mathcal{R}_\gamma(f) = \mathcal{R}_{\ell_\gamma}$ where $\ell_\gamma(z) := \max\{0, \min\{z/\gamma + 1, 1\}\}$ will lead to nice bounds for a number of methods, for instance boosting.

$\mathcal{R}_\ell$ recap

Last time we pointed out that bounded linear predictors $\mathcal{F} := \{x \mapsto \langle w, x\rangle : \|w\|_2 \leq W\}$ applied to bounded input values ($\|x\|_2 \leq X$) with a nondecreasing L-lipschitz loss (e.g., logistic loss is 1-Lipschitz) gives with probability $\geq 1-\delta$ for every $f \in \mathcal{F}$

$$\mathcal{R}_\ell(f) \leq \widehat{\mathcal{R}}_\ell(f) + LWX/\sqrt{n} + 3(LWX + \ell(0))\sqrt{\frac{2}{n}\ln\left(\frac{1}{\delta}\right)}.$$

Note moreover that regularization implies a choice of λ; namely, when $\lambda > 0$ and $\ell \geq 0$, minimization of

$$f(w) := \widehat{\mathcal{R}}(f) + \lambda\|w\|_2^2/2$$

implies that the ERM optimum w_λ satisfies

$$\lambda\|w_\lambda\|_2^2/2 \le f(w_\lambda) \le f(0) = \mathcal{R}(0),$$

thus we can take $W := \sqrt{2\mathcal{R}(0)/\lambda}$, and the earlier generalization terms with W become

$$\frac{LWX}{\sqrt{n}} = LX\sqrt{\frac{2\mathcal{R}(0)}{\lambda n}}.$$

Consequently, statistical learning theory typically recommends $\lambda \ge 1/n^{1-\varepsilon}$ for even $\varepsilon \le 1/2$ so that this bound goes to 0 quickly as n increases.

Remark. This is just a *sufficient* condition, not a *necessary* conditions.

$\mathcal{R}_{\mathbf{z}}$, VC dimension

Turning to $\mathcal{R}_{\mathbf{z}}$, we obtain a complexity term

$$\mathrm{Rad}(\{(x,y) \mapsto \mathbf{1}[\mathrm{sgn}(f(x)) \neq y] : f \in \mathcal{F}\}_{|S}\}).$$

The following definitions and lemma show how we can simplify this.

Now consider the sign patterns that arise from a set of real-valued predictors, meaning

$$\begin{aligned}
\mathrm{sgn}(\mathcal{F}) &:= \{x \mapsto \mathrm{sgn}(f(x)) : f \in \mathcal{F}\},\\
\mathrm{sgn}(U) &:= \{(\mathrm{sgn}(v_1), \ldots, \mathrm{sgn}(v_n)) : v \in U\}.
\end{aligned}$$

Define the **shatter coefficients** Sh and **VC dimension** VC as

$$\begin{aligned}
Sh(\mathcal{F}_{|S}) &:= \left|\mathrm{sgn}(\mathcal{F}_{|S})\right|,\\
Sh(\mathcal{F}; n) &:= \max_{S \in \mathcal{S}} \mathrm{Sh}(\mathcal{F}_{|S}),\\
\mathrm{VC}(\mathcal{F}) &:= \max\{i \in \mathbb{Z}_+ : \mathrm{Sh}(\mathcal{F}; i) = 2^i\}.
\end{aligned}$$

The following two lemmas show how to use these concepts in controlling $\mathcal{R}_{\mathbf{z}}$.

Lemma.

$$\mathrm{Rad}(\{(x,y) \mapsto \mathbf{1}[\mathrm{sgn}(f(x)) \neq y] : f \in \mathcal{F}\}_{|S}\}) \le \frac{1}{2}\mathrm{Rad}(\mathrm{sgn}(\mathcal{F})_{|S}).$$

Proof. For each coordinate i, define a map $f_i(z) := \mathbf{1}[z = y_i]$ for $z \in \{-1, +1\}$, and then extend this to an affine function by interpolation. Each f_i is 1/2-Lipschitz, thus the Lipschitz composition lemma gives the result. $\square$

Lemma (Sauer-Shelah, Vapnik-Chervonenkis?, Warren?)**.** Define $V := \mathrm{VC}(\mathcal{F})$ for convenience. Then

$$\mathrm{Sh}(\mathcal{F}; n) \le \begin{cases} 2^n & \text{when } n \le V,\\ \left(\frac{en}{V}\right)^V & \text{otherwise,}\end{cases}$$

and in general $\mathrm{Sh}(\mathcal{F}; n) \le n^V + 1$.

Proof. Omitted; this is taught in lots of machine learning classes… $\square$

Remark (historical). *[future matus: dig this all up properly.]*

1. Basically the definition of VC dimension appears in the Warren 1960s paper, who attributes it to an earlier paper by Shapiro. There is evidence Kolmogorov had a form of it decades earlier.
2. Warren roughly gives the Sauer-Shelah lemma in his 1960s paper.

Putting these pieces together, we get the following.

Theorem ("VC theorem")**.** With probability $\geq 1 - \delta$, every $f \in \mathcal{F}$ satisfies

$$\mathcal{R}_{\mathrm{z}}(f) \leq \hat{\mathcal{R}}_{\mathrm{z}}(f) + \mathrm{Rad}(\mathrm{sgn}(\mathcal{F})_{|S}) + 3\sqrt{\frac{2}{n}\ln\left(\frac{1}{\delta}\right)}$$

where

$$\mathrm{Rad}(\mathrm{sgn}(\mathcal{F})_{|S}) \leq \sqrt{\frac{8\ln(\mathrm{Sh}(\mathcal{F}_{|S}))}{n}} \qquad \text{and} \qquad \ln(\mathrm{Sh}(\mathcal{F}_{|S})) \leq \ln(\mathrm{Sh}(\mathcal{F}; n)) \leq \mathrm{VC}(\mathcal{F})\ln(n+1).$$

Remark (on optimization)**.**

1. As discussed many times, there are trivial cases where minimizing $\hat{\mathcal{R}}_{\mathrm{z}}$ is NP-hard.
2. Instead, it is common to minimize $\mathcal{R}_\ell$. This can be related as follows.

   Lemma. Suppose $\ell : \mathbb{R} \to \mathbb{R}_+$ is nondecreasing, and pick any $a \geq 0$ with $\ell(-a) > 0$. Then

   $$\mathcal{R}(\mathrm{sgn}(f)) \leq \Pr[f(X)Y \leq a] \leq \frac{\mathcal{R}_\ell(f)}{\ell(-a)}.$$

   Proof. By Markov's inequality,

   $$\begin{aligned} \mathcal{R}_{\mathrm{z}}(f) &= \Pr[\mathrm{sgn}(f(X)) \neq Y] \leq \Pr[Yf(X) \leq 0] \leq \Pr[Yf(X) \leq a] \\ &= \Pr[-Yf(X) \geq -a] \leq \Pr[\ell(-Yf(X)) \geq \ell(-a)] \\ &\leq \frac{\mathbb{E}(\ell(-Yf(X)))}{\ell(-a)}. \end{aligned}$$

   □

$\mathcal{R}_\gamma$

Define

$$\begin{aligned} l_\gamma(z) &:= \max\{0, \min\{1, 1 + z/\gamma\}\}, \\ \mathcal{R}_\gamma(f) &:= \mathcal{R}_{\ell_\gamma}(f) = \mathbb{E}(\ell_\gamma(-Yf(X))). \end{aligned}$$

These losses have a number of nice properties and are useful when analyzing boosting. Let's first get a few more Rademacher lemmas out of the way (not all of which we'll need here).

Lemma.

1. $\mathrm{Rad}(U) \geq 0$.
2. $\mathrm{Rad}(cU + \{v\}) \leq |c|\mathrm{Rad}(U)$.
3. $\mathrm{Rad}(\mathrm{conv}(U)) = \mathrm{Rad}(U)$.
4. Let sets of vectors $(U_i)_{i\geq 1}$ be given with $\sup_{v\in U_i} \langle \sigma, v\rangle_n \geq 0$ for every $\sigma \in \{-1, +1\}^n$. Then

   $$\mathrm{Rad}(\bigcup_i U_i) \leq \sum_i \mathrm{Rad}(U_i).$$

   (For instance, it suffices to have $U_i = -U_i$, or $0 \in U_i$.)

Proof.

1. $\mathbb{E}_\sigma \sup_{v\in U} \langle \sigma, v\rangle_n \geq \sup_{v\in U} \mathbb{E}_\sigma \langle \sigma, v\rangle_n = 0$.

2. Use the Lipschitz composition lemma with the $|c|$-lipschitz maps $f_i(z) := cz + v_0$.

3. As has been used a number of times in the course, optimizing a linear function over a polytope is achieved at a corner:

$$\begin{aligned}\mathbb{E}_\sigma \sup_{v \in \mathrm{conv}(U)} \langle \sigma, v \rangle_n &= \mathbb{E}_\sigma \sup_{k \geq 1} \sup_{v_1,\ldots,v_k \in U} \sup_{a_1,\ldots,a_k \in \Delta_k} \left\langle \sigma, \sum_j a_j v_j \right\rangle_n \\ &= \mathbb{E}_\sigma \sup_{v \in U} \langle \sigma, v \rangle_n .\end{aligned}$$

4. Thanks to the condition on each U_i,

$$\mathbb{E}_\sigma \sup_{v \in \bigcup_i U_i} \langle \sigma, v \rangle_n = \mathbb{E}_\sigma \sup_{i \geq 1} \sup_{v \in U_i} \langle \sigma, v \rangle_n \leq \mathbb{E}_\sigma \sum_i \sup_{v \in U_i} \langle \sigma, v \rangle_n = \sum_i \mathbb{E}_\sigma \sup_{v \in U_i} \langle \sigma, v \rangle_n .$$

□

Remark. In the case $\mathrm{Rad}(\cup_i U_i)$, some condition on U_i is needed; otherwise, any countable class U could be decposed into singletons U_i, and $0 \leq \mathrm{Rad}(U) \leq \sum_i \mathrm{Rad}(U_i) = 0$, which contradicts the existence of classes with cardinality 2 but positive Rademacher complexity.

These tools lead to the following control on $\mathcal{R}_\gamma$.

Theorem. Let some base class of functions $\mathcal{H}$ be given, and suppose

$$\mathcal{F} := W\mathrm{conv}(\mathcal{H} \cup -\mathcal{H}) = \left\{ \sum_{i=1}^k \alpha_i h_i : k \in \mathbb{Z}_+,\ \alpha \in \mathbb{R}^k, \|\alpha\|_1 \leq W, h_1, \ldots, h_k \in \mathcal{H} \right\}$$

With probability $\geq 1 - \delta$, every $f \in \mathcal{F}$ satisfies

$$\mathcal{R}_{\mathrm{z}}(f) \leq \mathcal{R}_\gamma(f) \leq \hat{\mathcal{R}}_\gamma(f) + \frac{2W}{\gamma} \mathrm{Rad}(\mathcal{H} \cup -\mathcal{H}) + 3\sqrt{\frac{2}{n} \ln\left(\frac{1}{\delta}\right)}.$$

Proof. Automatically, $\mathcal{R}_{\mathrm{z}}(f) \leq \mathcal{R}_\gamma(f)$, and the rest follows from Rademacher rules above. (Note the deviations have scaling 1 since they are controlled before unwrapping the Rademacher complexity.) □

Remark (on optimization, and the meaning of this bound)**.**

- First note the most valuable setting which justifies this theorem. Note that this theorem gives an excellent bound whenever $\widehat{\mathcal{R}}_\gamma(f)$ is small as well as W/γ. This property is referred to as the existence of a small-complexity hypothesis which separates the data with a significant *margin* γ. Under some conditions, boosting methods are known to output a classifier that satisfies this guarantee *[matus: include citation]*; consequently, this bound was heralded as a major achievement in understanding the success of boosting, which for a long time seemed to enjoy impossibly good generalization performance.
- To make more sense of this, consider the following alternative bound on the complexity of the predictors output by boosting. Namely, after t rounds of boosting, the output predictor is a linear combination of at most t predictors. Though it will not be proved here *[note to future matus : well I almost included it here..]*, a brute-force VC dimension upper bound is linear in t; therefore, the earlier bound miraculously seems to stay independent of t. Of course, one must consider the method of choosing W...
- Lastly, note that just as with $\mathcal{R}_{\mathrm{z}}$ and $\mathcal{R}_\ell$, it is easy to relate $\mathcal{R}_\gamma$ and $\mathcal{R}_\ell$; in particular, if $\ell'(0) > 0$,

  $$\ell_\gamma(z) \leq \frac{1}{\min\{\ell(0), \ell'(0)\gamma\}} \ell(z).$$

  This follows since $\ell(z) \geq 0$ and, setting $r := \min\{\ell(0), \ell'(0)\gamma$ for convenience,

  $$\ell(z) \geq \ell(0) + \ell'(0) = r\left(\frac{\ell(0)}{r} + \frac{\ell'(0)}{r} z\right) \geq r\left(1 + \frac{z}{\gamma}\right).$$

References