CSE 598 TEL - Homework 3

MJT

November 28, 2016

Instructions. (Same as homework 2.)

- Everyone must submit an individual write-up.
- You may discuss with up to 3 other people. State their NetIDs clearly on the first page. Outside of office hours, you should not discuss with anyone but these three.
- Homework is due **Wednesday, December 14, at 11:00am**; no late homework accepted.
- Each part of each question will count for 1 point.
- Please consider using the provided LaTeX file as a template (apologies for its weird indentation).
- Policies on the webpage still apply (where not in conflict with the policies above).

1. **(The Dudley entropy integral.)**

Lecture 22 presented[1] the following bound.

Theorem 1. *Given function class $\mathcal{F}$ and finite sample S of size n,*

$$\mathrm{Rad}(\mathcal{F}_{|S}) \leq \inf_{\epsilon>0}\left(\frac{\epsilon}{n^{1/p}} + \frac{\sup_{f\in\mathcal{F}} 2\|f(S)\|_2\sqrt{2\ln(\mathcal{N}_p(\mathcal{F};\epsilon,S))}}{n}\right).$$

This question will develop a tighter bound: the *Dudley entropy integral*. This bound has the following benefits.

- We'll use it for an upper bound on Rademacher complexity, but it can also give a lower bound.
- It shaves a $\ln(n)$ in the "VC Theorem" (which originally appears via Sauer-Shelah).
- Dudley has a nice theme[2].

This question will use the following notation.

$$\begin{aligned}
S &\quad \text{sample of size } n,\\
\mathcal{F} &\quad \text{function class},\\
\mathcal{F}_0 := \{x \mapsto 0\} &\quad \text{coarsest cover; scale } r_0 := \sup_{f\in\mathcal{F}}\|f\|_p,\\
(\mathcal{F}_i)_{i=1}^k &\quad \text{(minimal) covers at scale } r_i := 2^{-i}r_0,\\
f_i \in \mathcal{F}_i &\quad \text{closest approximant to a given } f\in\mathcal{F},\\
C_n &\quad \text{constant (given } n\text{) with } \|v\|_2 \leq C_n\|v\|_p \text{ for all } v\in\mathbb{R}^n.
\end{aligned}$$

By these definitions, for any $f\in\mathcal{F}$ and any $i\in\{0,\ldots,k\}$, there exists $f_i\in\mathcal{F}_i$ with $\|f(S)-f_i(S)\|_p \leq r_i$.

(a) Using the representation $f = f - f_k + \sum_{i=1}(f_i - f_{i-1})$ (which motivates $f_0 = 0$), prove

$$\mathrm{Rad}(\mathcal{F}_{|S}) \leq 2^{-k}r_0 n^{-1/p} + \frac{6C_n}{n}\sum_{i=1}^{k} r_i\sqrt{\ln|\mathcal{F}_i|}.$$

Remark: C_n really shouldn't be there, and is an artifact of ℓ_2 nature of the Massart finite lemma. This problem is open in the literature, so feel free to ping me about it.

(b) Starting from the previous problem, now prove

$$\mathrm{Rad}(\mathcal{F}_{|S}) \leq \inf_{\alpha\leq 1}\frac{2\alpha r_0}{n^{1/p}} + \frac{12C_n}{n}\int_{\alpha r_1}^{r_1}\sqrt{\ln(\mathcal{N}_p(\mathcal{F};r,S))}dr.$$

This is the Dudley entropy integral.

(c) Suppose $p=2$, $|f(x)|\leq 1$ for all $f\in\mathcal{F}$, and $\ln(\mathcal{N}_2(\mathcal{F};\epsilon,S)) \leq C\ln(n)\sqrt{n}/\epsilon$ for some $C>0$. Use Theorem 1 to derive an upper bound on $\mathrm{Rad}(\mathcal{F}_{|S})$. (Your upper bound should not contain ϵ.)

Note. Don't stress over the constants in the final bound, just make sure the exponent on n is optimal given Theorem 1. Also, note that this bound on $\mathcal{N}_2$ is not artificial, for instance it holds for nondecreasing univariate functions.

(d) Suppose the setting of the previous part, but now use the Dudley entropy integral to upper bound $\mathrm{Rad}(\mathcal{F}_{|S})$. **Note:** your dependence on n should now be better.

[1] `http://mjt.web.engr.illinois.edu/courses/f2016/mltheory/lec22.html`.
[2] `https://www.youtube.com/watch?v=jUs-ITmi_u4`.

Solution.

2. **(Statistical guarantees for least squares.)**

This problem will develop a generalization story for least squares, complementing the optimization analysis from last time.

First recall the basic Rademacher bound from class.

Theorem 2. *Suppose the distribution over Z has support U. Then with probability at least $1-\delta$ over a sample $S := (z_1, \ldots, z_n)$, every $g \in \mathcal{G}$ satisfies*

$$\mathbb{E}(g) \leq \hat{\mathbb{E}}(g) + 2\mathrm{Rad}(\mathcal{G}_{|S}) + 3 \sup_{\substack{g\in\mathcal{G}\\ z\in U}} |g(z)| \sqrt{\frac{2}{n} \ln\left(\frac{2}{\delta}\right)}.$$

Throughout this problem, assume the following.

- The setting is classification over $\mathbb{R}^d$, meaning $z_i = (x_i, y_i)$ where $x_i \in \mathbb{R}^d$ with $\|x_i\|_2 \leq 1$ (for simplicity) and $y_i \in \{-1, +1\}$. For convenience, combine these two sets into $\mathcal{Z} := \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\} \times \{-1, +1\}$.
- The predictors are linear: $\mathcal{F} := \{x \mapsto \langle w, x\rangle : w \in \mathcal{W}\}$, with $\mathcal{W} := \{w \in \mathbb{R}^d : \|w\|_2 \leq B\}$, where $B \geq 0$ can be treated as a constant throughout the problem.
- We are concerned with risk-minimization of a margin loss, meaning $\mathcal{R}_\ell(f) = \mathcal{R}_\ell(w) = \mathbb{E}(\ell(-yf(x))) = \mathbb{E}(\ell(-y\langle w, x\rangle))$, where ℓ is L-Lipschitz over $[-B, +B]$.

To start the problem, first consider the case of least squares.

(a) Further assume the least squares loss

$$\ell_{\mathrm{sq}}(w, (x, y)) := (\langle w, x\rangle - y)^2/2 = (1 - y\langle w, x\rangle)^2/2.$$

Use Theorem 2 and the Rademacher lemmas from class to upper bound $\mathcal{R}_{\ell_{\mathrm{sq}}} - \widehat{\mathcal{R}}_{\ell_{\mathrm{sq}}}$. Your answer can not use L, but should instead use B.

(b) Can you use the preceding analysis to put a bound on the standard *empirical* solution (from homework 2) $\hat{w} := \hat{\mathbb{E}}(xx^\top)^\dagger \hat{\mathbb{E}}(xx^\top y)$? Why, or why not?

The rest of the problem will show how to get the correct rate $\mathcal{O}(1/n)$, albeit with suboptimal constants. This analysis will make use of a *peeling argument*. Throughout this problem, suppose the following.

- The risk $\mathcal{R}_\ell$ is λ-strongly-convex (with respect to w). **Note:**
  - No assumption is made about $\widehat{\mathcal{R}}_\ell$; it can fail to be strongly convex.
  - Note that Lipschitz contradicts strongly convex over all of $\mathbb{R}^d$, but here Lipschitz was assumed only for $[-B, +B]$.
- Let $\bar{w} \in \mathbb{R}^d$ denote an optimum of $\mathcal{R}_\ell$, which exists and is unique following derivations in homework 2. Assume B is large enough so that $\|\bar{w}\|_2 \leq B$. **Note:** similarly to the previous note, $\bar{w}$ will not in general minimize $\widehat{\mathcal{R}}_\ell$.

Define the following objects, treating $r > 0$ as a fixed scalar whose value will be determined later in the problem.

$$\begin{aligned}
\mathcal{E}_\ell(w) &:= \mathcal{R}_\ell(w) - \inf_{v\in\mathbb{R}^d} \mathcal{R}_\ell(v) = \mathcal{R}_\ell(w) - \mathcal{R}_\ell(\bar{w}),\\
\widehat{\mathcal{E}}_\ell(w) &:= \widehat{\mathcal{R}}_\ell(w) - \inf_{v\in\mathbb{R}^d} \widehat{\mathcal{R}}_\ell(v),\\
k_w &:= \min\left\{k \in \mathbb{Z}_+ : \mathcal{E}(w) \leq r4^k\right\},\\
f_w(z) = f_w((x, y)) &:= \ell(-y\langle w, x\rangle) - \ell(-y\langle \bar{w}, x\rangle)\\
g_w(z) &:= 4^{-k_w} f_w(z),\\
\mathcal{G} &:= \{g_w : w \in \mathcal{W}\}.
\end{aligned}$$

The quantity $\mathcal{E}_\ell(w)$ is the *excess risk*. The starting point of the proof is the following bound, which follows from applying Theorem 2 to $\mathcal{G}$: with probability at least $1-\delta$, every $g_w \in \mathcal{G}$ satisfies

$$\mathbb{E}g_w \le \hat{\mathbb{E}}g_w + 2\mathrm{Rad}(\mathcal{G}_{|S}) + 3 \sup_{\substack{g_w \in \mathcal{G} \\ z \in \mathcal{Z}}} |g_w(z)| \sqrt{\frac{2}{n} \ln\left(\frac{2}{\delta}\right)}. \tag{1}$$

The following questions are for the general margin loss ℓ, not the special case ℓ_{sq}.

(c) Prove $\sup_{\substack{g_w \in \mathcal{G} \\ z \in \mathcal{Z}}} |g_w(z)| \le L\sqrt{2r/\lambda}$.

(d) Define, for every $a \ge 0$, two helper classes

$$\mathcal{F}(a) := \{f_w : w \in \mathcal{W}, \mathcal{E}_\ell(w) \le a\},$$
$$\tilde{\mathcal{F}}(a) := \{f_w : w \in \mathcal{W}, \|w - \bar{w}\|_2 \le \sqrt{2a/\lambda}\}.$$

Prove

$$\mathrm{Rad}(\mathcal{F}(a)_{|S}) \le \mathrm{Rad}(\tilde{\mathcal{F}}(a)_{|S}) \le L\sqrt{\frac{4a}{\lambda n}}.$$

Hint: strong convexity gives a relationship between $\|w - \bar{w}\|_2$ and $\mathcal{E}_\ell(w)$.

(e) Using the preceding part, prove

$$\mathrm{Rad}(\mathcal{G}_{|S}) \le 4L\sqrt{\frac{r}{\lambda n}}.$$

Hint: use $\mathcal{G} \subseteq \cup_{k \ge 0} 4^{-k}\mathcal{F}(r4^k)$; this is where the "peeling" is most apparent.

(f) Now use the preceding parts, together with eq. (1), the choice

$$r := \frac{2^{16}L^2 \ln(2e/\delta)}{\lambda n},$$

and perhaps also the elementary inequality $\sqrt{a} + \sqrt{b} \le \sqrt{2(a+b)}$ (for $a, b \ge 0$) to prove: with probability at least $1-\delta$, every $w \in \mathcal{W}$ satisfies

$$\mathcal{E}_\ell(w) \le \widehat{\mathcal{E}}_\ell(w) + \frac{r4^{k_w}}{8}.$$

Note: r does not have optimal constants, it has been made large to give you some wiggle room.

(g) By separately considering the cases $k_w = 0$ and $k_w > 0$, prove: with probability at least $1-\delta$, every $w \in \mathcal{W}$ satisfies

$$\mathcal{E}_\ell(w) \le 2\widehat{\mathcal{E}}_\ell(w) + \frac{r}{8} = 2\widehat{\mathcal{E}}_\ell(w) + \frac{2^{13}L^2 \ln(2e/\delta)}{\lambda n}.$$

Lastly, returning to ℓ_{sq}.

(h) Instantiate the previous bound for ℓ_{sq}, but replace L and λ with expressions involving B, $\mathbb{E}(xx^\top)$ and $\mathbb{E}(xy)$.

(i) Can you use the preceding analysis to put a bound on the standard *empirical* solution $\hat{w} := \hat{\mathbb{E}}(xx^\top)^\dagger \hat{\mathbb{E}}(xx^\top y)$? Why, or why not?

Solution.

3. **(Epilogue.)**
   Are we still friends? (**Hard mode:** tell the truth.)

   Solution.