MLT Homework 6

Ana Borovac
Jonas Haslbeck
Bas Haver

October 29, 2018

Raw 0/1-loss with real-valued predictions:

$$l'(y, \hat{y}) = \frac{1 - \operatorname{sign}(y \cdot \hat{y})}{2}, \quad \operatorname{sign}(u) = \begin{cases} 1 & \text{if } u \geq 0, \\ -1 & \text{if } u < 0 \end{cases}$$

Question 1

Suppose that we have a loss function l based on a raw l' as above that is bounded, i.e. $\sup_{\hat{y} \in \hat{Y}, y \in Y} |l'(y, \hat{y})| < \infty$.

Subquestion 1.1

Suppose that $\mathcal{H}$ consists of just one hypothesis, $\mathcal{H} = \{h\}$. Show that for all samples S, we have that $R(l \circ \mathcal{H} \circ S) = 0$.

Solution

Let's recall the definition of Rademacher complexity:

$$R(l \circ \mathcal{H} \circ S) = \frac{1}{m} \mathbb{E}_{(y_1, \ldots, y_m) \in \{-1,1\}^m} \left[\sup_{h \in \mathcal{H}} \sum_{y_i} y_i \, l(h, x_i) \right]$$

In our case we have:

$$R(l \circ \mathcal{H} \circ S) = \frac{1}{m} \mathbb{E}_{(y_1, \ldots, y_m) \in \{-1,1\}^m} \left[\sum_{y_i} y_i \, l(h, x_i) \right]$$

With linearity of expected value:

$$R(l \circ \mathcal{H} \circ S) = \frac{1}{m} \sum_{y_i} l(h, x_i) \, \mathbb{E}_{(y_1, \ldots, y_m) \in \{-1,1\}^m} [y_i]$$

Let calculate the expected value of y_i:

$$\mathbb{E}[y_i] = -1 \cdot P(y_i = -1) + 1 \cdot P(y_i = 1) = -p + p = 0$$

where $P(y_i = 1) = P(y_i = -1) = p$. Now we can conclude:

$$R(l \circ \mathcal{H} \circ S) = \frac{1}{m} \sum_{y_i} l(h, x_i) \cdot 0 = 0$$

Subquestion 1.2

Consider a hypothesis class $\mathcal{H}$ and some hypothesis h'. Show that

$$R(l \circ \mathcal{H} \cup \{h'\} \circ S) \geq R(l \circ \mathcal{H} \circ S),$$

and explain why it follows that Rademacher complexity is always nonnegative.

Solution

First we are going to write what is $R(l \circ \mathcal{H} \cup \{h'\} \circ S)$:

$$R(l \circ \mathcal{H} \cup \{h'\} \circ S) = \frac{1}{m} \mathbb{E}_{(y_1, \ldots, y_m) \in \{-1,1\}^m} \left[\sup_{h \in \mathcal{H} \cup \{h'\}} \sum_{y_i} y_i \, l(h, x_i) \right]$$

When comparing this with the rademacher complexity without the added h', we now take a supremum over a set containing the original class $\mathcal{H}$, so this rademacher complexity of the set with h' must be at least as big as the one without h'.
In the previous subquestion we proved that Rademacher complexity of hypothesis class with one hypothesis equals 0, now we proved that if we add a hypothesis to a hypothesis class we get at least as much as we did before. So, Rademacher complexity is nonnegative.

Subquestion 1.3

Rademacher complexity is often used as a tool to show that some given hypothesis class $\mathcal{H}$ is PAC-learnable. Show that Theorem 26.5 in the book implies agnostic PAC-learnability for $\mathcal{H}$ whenever we can prove that expected Rademacher complexity satisfies

$$\lim_{m \to \infty} \sup_{D} \mathbb{E}_{S \sim D^m} \left[R(l \circ \mathcal{H} \circ S) \right] = 0$$

where the supremum is over all probability distributions D that can be defined on $Z = X \times Y$.

Solution

We use the third statement of Theorem 26.5, which says that for any h^*, with probability of at least $1-\delta$,

$$L_{\mathcal{D}}(\mathrm{ERM}_{\mathcal{H}}(S)) - L_{\mathcal{D}}(h^*) \leq 2R(l \circ \mathcal{H} \circ S) + 5c\sqrt{\frac{2\ln(8/\delta)}{m}}.$$

Now since we assume

$$\lim_{m\to\infty} \sup_{\mathcal{D}} \mathbb{E}_{\S\sim\mathcal{D}^m}[R(l \circ \mathcal{H} \circ S)] = 0,$$

we also can pick, by the non-negativity as proved by the second subquestion, an $m_1(\epsilon,\delta)$ such that for all $m \geq m_1(\epsilon,\delta)$ we have $R(l \circ \mathcal{H} \circ S) \leq \epsilon/3$. Furthermore we can also pick an $m_2(\epsilon,\delta)$ such that for every $m \geq m_2(\epsilon,\delta)$ we have $5c\sqrt{\frac{2\ln(8/\delta)}{m}} \leq \epsilon/3$. Now when we choose $m(\epsilon,\delta) = \max(m_1(\epsilon,\delta), m_2(\epsilon,\delta))$ we have for choosing ERM, by lemma 26.5.3

$$L_{\mathcal{D}}(\mathrm{ERM}_{\mathcal{H}}(S)) - L_{\mathcal{D}}(h^*) \leq 2\epsilon/3 + \epsilon/3 = \epsilon,$$

which is what we need to have agnostic PAC-learnability.

Question 2

(Lipschitz Losses) Consider a raw loss function l' as above with prediction set $\hat{Y}$; here $\hat{Y}$ can either be $\mathbb{R}$ or a closed and bounded interval in $\mathbb{R}$. We say that l' is $\rho -$ Lipschitz if for all $a, b \in \hat{Y}$ we have

$$\sup_{y\in Y} |l'(y,a) - l'(y,b)| \leq \rho|a-b|$$

We say that l' is Lipschitz if it is ρ-Lipschitz for some $\rho < \infty$. Intuitively, if a loss function is Lipschitz then small changes in the predictions imply small changes in the loss. It turns out that analysis with Rademacher complexity can be done for (raw) loss functions that are either bounded or Lipschitz. For loss functions that are neither bounded nor Lipschitz, it is much more problematic. We will now consider Lipschitzness for some of the most commonly appearing loss functions.

Subquestion 2.1

Show that the 0/1-loss for real-valued predictions is not Lipschitz.

Solution

0/1-loss for real-valued predictions is defined as:

$$l'(y,\hat{y}) = \frac{1-\mathrm{sign}(y\cdot\hat{y})}{2}, \quad \mathrm{sign}(u) = \begin{cases} 1 & \text{if } u \geq 0, \\ -1 & \text{if } u < 0 \end{cases}$$

It follows:

$$\begin{aligned}
\sup_{y\in Y}|l'(y,a)-l'(y,b)| &= \sup_{y\in Y}\left|\frac{1-\operatorname{sign}(y\cdot a)}{2}-\frac{1-\operatorname{sign}(y\cdot b)}{2}\right| \\
&= \frac{1}{2}\sup_{y\in Y}|-\operatorname{sign}(y\cdot a)+\operatorname{sign}(y\cdot b)| \\
&= \frac{1}{2}\sup_{y\in Y}\left|(a-b)\frac{\operatorname{sign}(y\cdot a)-\operatorname{sign}(y\cdot b)}{a-b}\right| \\
&= \frac{1}{2}|a-b|\sup_{y\in Y}\left|\frac{\operatorname{sign}(y\cdot a)-\operatorname{sign}(y\cdot b)}{a-b}\right| \\
&= \frac{1}{2}|a-b|\frac{\sup_{y\in Y}|\operatorname{sign}(y\cdot a)-\operatorname{sign}(y\cdot b)|}{|a-b|}
\end{aligned}$$

If 0/1-loos would be Lipschitz, we would be able to bound:

$$\frac{\sup_{y\in Y}|\operatorname{sign}(y\cdot a)-\operatorname{sign}(y\cdot b)|}{2|a-b|}$$

We know that the value of numerator is between 0 and 2. If a and b are close together, the whole fraction becomes very large. From that we can conclude that 0/1-loss is not Lipschitz.

Subquestion 2.2

Is the absolute loss on a bounded domain $Y=\hat{Y}=[-1,1]$ Lipschitz?

Solution

The absolute loss is defined as:

$$l'(y,\hat{y})=|y-\hat{y}|$$

It follows:

$$\sup_{y\in Y}|l'(y,a)-l'(y,b)|=\sup_{y\in Y}||y-a|-|y-b||$$

If $a=b$ we get $\sup_{y\in Y}|l'(y,a)-l'(y,b)|=0$, so it is bounded with $\rho|a-b|$.

Without loss of generality we can say $a<b$. Now, we have three possible situations:

- $y\le a<b$: If follows $(y-a)\le 0$ and $(y-b)<0$, so:

$$\sup_{y\in Y}||y-a|-|y-b||=\sup_{y\in Y}|-y+a+y-b|\le\underbrace{1}_{\rho}\cdot|a-b|$$

- $a < y \leq b$: If follows $(y-a) > 0$ and $(y-b) \leq 0$, so:

$$\begin{aligned}
\sup_{y \in Y} ||y-a| - |y-b|| &= \sup_{y \in Y} |y - a + y - b| \\
&= \sup_{y \in Y} \left| (a-b) \left(\frac{y-a}{a-b} + \frac{y-b}{a-b} \right) \right| \\
&= |a-b| \sup_{y \in Y} \left| \left(\frac{y-a}{a-b} + \frac{y-b}{a-b} \right) \right| \\
&\leq |a-b| \sup_{y \in Y} \left(\left| \frac{y-a}{a-b} \right| + \left| \frac{y-b}{a-b} \right| \right) \\
&= |a-b| \sup_{y \in Y} \left(\frac{|y-a|}{|a-b|} + \frac{|y-b|}{|a-b|} \right)
\end{aligned}$$

From $a < y \leq b$ we know:

$$\begin{aligned}
|y-a| < |a-b| \quad &\Rightarrow \quad \frac{|y-a|}{|a-b|} < 1 \\
|y-b| \leq |a-b| \quad &\Rightarrow \quad \frac{|y-b|}{|a-b|} \leq 1
\end{aligned}$$

So:

$$\begin{aligned}
\sup_{y \in Y} ||y-a| - |y-b|| &= |a-b| \sup_{y \in Y} \left(\frac{|y-a|}{|a-b|} + \frac{|y-b|}{|a-b|} \right) \\
&\leq |a-b| \left(\sup_{y \in Y} \frac{|y-a|}{|a-b|} + \sup_{y \in Y} \frac{|y-b|}{|a-b|} \right) \\
&\leq |a-b|(1+1) \\
&= \underbrace{2}_{\rho} \cdot |a-b|
\end{aligned}$$

- $a < b < y$: If follows $(y-a) > 0$ and $(y-b) > 0$, so:

$$\sup_{y \in Y} ||y-a| - |y-b|| = \sup_{y \in Y} |y - a - y + b| \leq \underbrace{1}_{\rho} \cdot |a-b|$$

To sum up, we can conclude that:

$$\sup_{y \in Y} |l'(y,a) - l'(y,b)| \leq \underbrace{2}_{\rho} \cdot |a-b|$$

So, yes, the absolute loss is Lipschitz.

Subquestion 2.3

Is the unbounded absolute loss on unbounded domain $Y = \hat{Y} = \mathbb{R}$ Lipschitz?

Solution

Yes, we can repeat the proof from bounded domain.

Subquestion 2.4

Is the squared error loss $l'(y, \hat{y}) = (y - \hat{y})^2$, defined on bounded domain $Y = \hat{Y} = [-1, 1]$ Lipschitz?

Solution

From the below:

$$\begin{aligned}
\sup_{y \in Y} |l'(y, a) - l'(y, b)| &= \sup_{y \in Y} |(y - a)^2 - (y - b)^2| = \\
&= \sup_{y \in Y} |y^2 - 2ay + a^2 - y^2 + 2by - b^2| = \\
&= \sup_{y \in Y} |a^2 - b^2 - 2y(a - b)| = \\
&= \sup_{y \in Y} |(a - b)(a + b) - 2y(a - b)| = \\
&= \sup_{y \in Y} |(a - b)(a + b - 2y)| = \\
&= \sup_{y \in Y} |a - b||a + b - 2y| = \\
&= |a - b| \sup_{y \in Y} |a + b - 2y| \leq \\
&\leq |a - b| \sup_{y \in Y} (|a + b| + |2y|) \leq \\
&\leq |a - b|(|a + b| + \sup_{y \in Y} |2y|) \leq \\
&\leq |a - b|(2 + 2) = \\
&= \underbrace{4}_{\rho} \cdot |a - b|
\end{aligned}$$

it follows that the squared error loss defined on bounded domain is Lipschitz.

Subquestion 2.5

Is the unbounded squared error loss defined on unbounded domain $Y = \hat{Y} = \mathbb{R}$ Lipschitz?

Solution

No. If it was, we should be able to bound with a constant $< \infty$:

$$\sup_{y \in Y} |a + b - 2y|$$

which we can not, since Y is not bounded.

Question 3

(Simplifying Rademacher) To make Rademacher complexity analysis easier and a bit less abstract, it would be very convenient to work directly with the hypotheses $h(x)$ rather than the derived functions $f_h(x, y) := l(h, (x, y)) = l'(y, h(x))$, as is done in the book. If we can do this we can completely ignore the Y-values in our analyss, thus making things simpler (this is illustrated in the next exercise). Show that we can indeed work with h instead of f_h whenever the loss is Lipschitz. To be precise, show that, if $l(h, (x, y)) = l'(y, h(x))$ as above and l' is Lipschitz, then there exists a constant $0 < c < \infty$ such that for every $x_1, \ldots, x_m \in \mathcal{X}^m$, every $y_1, \ldots, y_m \in \mathcal{Y}^m$: $R(l \circ \mathcal{H} \circ S) \leq c \cdot R(\mathcal{H} \circ S_{\mathcal{X}}$, where $S = ((x_1, y_1), \ldots, (x_m, y_m))$ and $S_{\mathcal{X}} := (x_1, \ldots, x_m)$.
HINT: this is straightforward based on one of the results in the chapter on Rademacher complexity in the book.

Solution

As the hint reveals, we can make use of one of the results from the book. In our case we will use Lemma 26.9, the Contraction Lemma, which actually gives us the desired result once we show that our situaton is from the form as demanded from the Lemma.
First, we can choose the A from the lemma to be $\mathcal{H} \circ S_{\mathcal{X}}$. When we choose ϕ_i to be $l'(y_i, h(x))$ than it is as in the lemma (assuming that there was made a typo in the book such that $(\phi_1(a_1), \ldots, \phi_m(y_m))$ should be $(\phi_1(a_1), \ldots, \phi_m(a_m))$ in order to make sence as no y is defined in that lemma, but in our case it actually is a y), and Lipschitz by assumption. Because $l(h, (x, y)) = l'(y, h(x))$, we can also conclude that $R(l \circ \mathcal{H} \circ S) = R(l' \circ \mathcal{H} \circ S_{\mathcal{X}})$ and the result follows from the lemma when we choose c to be equal to any value for which l' is c-Lipschitz.