Homework 3
APPM 4490/5490 Theory of Machine Learning, Spring 2024

Due date: Friday, Feb 9 '24, before 11 AM, via paper or via Gradescope

Instructor: Prof. Becker
Revision date: 2/3/2024

Theme: Rademacher complexity

Instructions Collaboration with your fellow students is OK and in fact recommended, although direct copying is not allowed. The internet is allowed for basic tasks (e.g., looking up definitions on wikipedia) but it is not permissible to search for proofs or to *post* requests for help on forums such as `http://math.stackexchange.com/` or to look at solution manuals. Please write down the names of the students that you worked with.

An arbitrary subset of these questions will be graded.

Reading You are responsible for reading chapter 3.1 (about Rademacher complexity) in Foundations of Machine Learning, 2nd edition, by Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar (MIT Press, 2018, ISBN-13: 978-0262039406, \$50 on Amazon). The authors host a free PDF of the book at their website.

Problem 1: Let $\mathcal{H} = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle \mid \|\mathbf{w}\|_2 \le 1\}$ be a set of linear classifiers. Let $S_x = (\mathbf{x}_1, \ldots, \mathbf{x}_m) \subset \mathbb{R}^n$ be a collection of vectors, and define the Frobenius norm[1] of this set as $\|S_x\|_F^2 = \sum_{i=1}^m \|\mathbf{x}_i\|_2^2$. We define $\mathcal{H} \circ S_x \overset{\text{def}}{=} \{\mathbf{a} = (a_1, \ldots, a_m) \mid h \in \mathcal{H}, a_i = h(\mathbf{x}_i)\}$.

a) Show the (empirical) Rademacher complexity is bounded as follows:

$$\widehat{\mathfrak{R}}(\mathcal{H} \circ S_x) \le \frac{1}{m} \mathbb{E}_{\boldsymbol{\sigma}} \left\| \sum_{i=1}^m \sigma_i \mathbf{x}_i \right\|_2$$

(In fact, this is an equality, but you don't need to show that)

b) Simplify the bound to the following:

$$\widehat{\mathfrak{R}}(\mathcal{H} \circ S_x) \le \frac{1}{m} \|S_x\|_F$$

(Hint: use Jensen's inequality)

c) Suppose $\mathcal{D}$ is the multivariate normal distribution $\mathcal{N}(0, I_{n\times n})$ on $\mathbb{R}^n$. Building on your work from part (b), compute a bound on the (expected) Rademacher complexity $\mathfrak{R}_m(\mathcal{H}) = \mathbb{E}_{S_x \sim \mathcal{D}^m} \widehat{\mathfrak{R}}(\mathcal{H} \circ S_x)$. The bound should only depend on m and n.

Note: It is not straightforward to compute $\widehat{\mathfrak{R}}(\ell \circ \mathcal{H} \circ S_x)$ where ℓ is the $0-1$ loss function composed with a thresholding function, since ℓ is not Lipschitz continuous.

Background facts If $\|\cdot\|$ is some norm, we define its *dual norm* (sometimes written $\|\cdot\|_*$) as

$$\|\mathbf{a}\|_* = \sup_{\|\mathbf{x}\| \le 1} \langle \mathbf{x}, \mathbf{a} \rangle.$$

For $1 \le p \le \infty$, the "p-norms" on $\mathbb{R}^n$ are defined

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p}, \quad 1 \le p < \infty; \quad \|\mathbf{x}\|_\infty = \max_{i \in [n]} |x_i|, \quad p = \infty$$

[1] See end of HW for background on norms if you are not familiar

and if $p = 2$ we call this the Euclidean norm (this is often our "default" norm). If $\frac{1}{p} + \frac{1}{q} = 1$ (we say p and q are "Hölder conjugates") then $\|\cdot\|_q$ is the dual norm of $\|\cdot\|_p$ and vice-versa, and we have Hölder's inequality

$$|\langle \mathbf{a}, \mathbf{x} \rangle| \le \|\mathbf{a}\|_p \cdot \|\mathbf{x}\|_q$$

which reduces to Cauchy-Schwarz when $p = q = 2$.

For a matrix $A \in \mathbb{R}^{m \times n}$, if we view this as a reshaped vector of size mn, then the most natural norm is the Frobenius (aka Hilbert-Schmidt) norm, $\|A\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^m |A_{ij}|^2}$, whereas if we view it as a linear operator from $\mathbb{R}^n \to \mathbb{R}^m$ then the most natural norm is the appropriate operator norm, in this case usually called the spectral norm, written either as $\|A\|_2$ or just $\|A\|$, defined as $\|A\| = \sup_{\mathbf{x} \neq 0} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$, which is equivalent to the largest singular value. So be careful, $\|\cdot\|_2$ often means the Euclidean norm if the input is a vector, and the spectral norm if the input is a matrix (in particular, Matlab follows this convention).