Homework 6

36-465/665, Spring 2021

Due at 6 pm on Thursday, 18 March 2021

Agenda: Seeing that the growth function (which is distribution-free) upper-bounds the Rademacher-complexity (which is distribution-dependent).

An important result related to Rademacher complexity is called **Massart's lemma**, and goes as follows. Say $x_1, x_2, \ldots x_m$ are a fixed (non-random) and finite set of n-dimensional vectors, and $\rho = \max_{i \in 1:m} \|x_i\|$ (i.e. ρ is the length of the longest vector). Say σ is a n-dimensional vector of Rademacher random variables, so each coordinate of σ is ± 1 with equal probability and independent of the other coordinates. Then

$$\mathbb{E}\left[\max_{i \in 1:m} \frac{1}{n} \sigma \cdot x_i\right] \leq \frac{\rho\sqrt{2 \log m}}{n} \tag{1}$$

(This type of result is sometimes called a "maximal inequality".)

1. *The growth function limits Rademacher complexity.* In this question, we will assume that Massart's lemma holds, and use it to prove that, when $\mathcal{F}$ is a family of functions taking the values ± 1,

$$\mathcal{R}_n(\mathcal{F}) \leq \sqrt{\frac{2 \log \Pi_{\mathcal{F}}(n)}{n}} \tag{2}$$

as asserted in the slides for lecture 9. For the purposes of this problem, assume that $\mathcal{F}$ is "closed under negation", meaning that if $f \in \mathcal{F}$, then $-f \in \mathcal{F}$ as well.

   a. (5) Suppose all the entries of an n-dimensional vector are either -1 or 1. Show that the length of the vector is $\sqrt{n}$.

   b. (5) Suppose we see data points $z_1, z_2, \ldots z_n$. Explain how applying a function $f \in \mathcal{F}$ to these data points leads to an n-dimensional vector; what is the length of this vector?

   c. (6) Define $\Pi(z_1, \ldots z_n)$ to be the number of *distinct* vectors we can get by using different functions from $\mathcal{F}$ in the manner described in Q1b. Use Massart's lemma to show that

$$\hat{\mathcal{R}}_n(\mathcal{F}) \leq \sqrt{\frac{2 \log \Pi(z_1, \ldots z_n)}{n}} \tag{3}$$

   d. (6) Use Q1c to show that

$$\hat{\mathcal{R}}_n(\mathcal{F}) \leq \sqrt{\frac{2 \log \Pi_{\mathcal{F}}(n)}{n}} \tag{4}$$

   e. (5) Use (e) to show that

$$\mathcal{R}_n(\mathcal{F}) \leq \sqrt{\frac{2 \log \Pi_{\mathcal{F}}(n)}{n}} \tag{5}$$

2. *Proving Massart's lemma*

   a. (5) Explain why, for any $t > 0$,

$$\exp\left(t\mathbb{E}\left[\max_{i \in 1:m} \sigma \cdot x_i\right]\right) \leq \mathbb{E}\left[\exp\left(t \max_{i \in 1:m} \sigma \cdot x_i\right)\right] \tag{6}$$

b. (5) Using Q2a, show that

$$\exp\left(t\mathbb{E}\left[\max_{i\in 1:m} \sigma \cdot x_i\right]\right) \leq \sum_{i=1}^{m} \mathbb{E}\left[\exp\left(t\sigma \cdot x_i\right)\right] \tag{7}$$

c. (5) Use Q2b, the fact that coordinates of σ are IID, and the Hoeffding bounds, to show that

$$\exp\left(t\mathbb{E}\left[\max_{i\in 1:m} \sigma \cdot x_i\right]\right) \leq m\exp\left(t^2\rho^2/2\right) \tag{8}$$

d. (6) Use Q2c to show that

$$\mathbb{E}\left[\max_{i\in 1:m} \sigma \cdot x_i\right] \leq \frac{\log m}{t} + \frac{t\rho^2}{2} \tag{9}$$

e. (5) Use Q2d to prove Massart's lemma.

3. *Getting a feel for the VC bounds* In the slides, we derived generalization error bound using VC dimension. These involved a term $\sqrt{2\frac{\log(en/d)}{n/d}}$.

   a. (5) Plot this as a function of n over the range from 1 to 10^6 for $d = 1$, $d = 2$, $d = 10$, and $d = 1000$. (The full range may not be meaningful for every d — why not?) I suggest using a logarithmic scale on the horizontal axis.

   b. (5) Why do you think I'm only asking you to plot this term, and not the whole of the bound on the generalization error?

   c. (6) What lessons do you take from these curves about how much data you need to estimate models of different complexities to the same precision?

4. *"It is certainly not the least charm of a theory that it is refutable"*: You have very likely heard people talk about whether or not a certain idea is "falsifiable". The idea of falsifiability goes back to the philosopher Karl Popper in 1934, who asserted that a hypothesis is scientific only if it could be falsified by observations, i.e., if there is some possible data which could show that the hypothesis was false[1]. Suppose we're interested in the relationship between some binary variable Y and one or more other variables X. Our friend Fritz proposes that $Y = f(X)$, exactly, for some function f in a family of classifier functions $\mathcal{F}$. Fritz is vague about exactly which $f \in \mathcal{F}$, but, as he says, many scientific theories include parameters which have to be determined by experiment or observation.

   a. (5) If Fritz's family of classifiers $\mathcal{F}$ has VC dimension $d < \infty$, then Fritz's hypothesis is falsifiable. Explain why this is so, and explain why it would be possible to falsify it with as few as $d + 1$ data points.

   b. (5) Could Fritz's hypothesis be falsified by fewer than $d + 1$ data points, or do we need at least that many to have any hope of falsifying it? If you think the answer depends on further details about $\mathcal{F}$, beyond its VC dimension, explain.

   c. (5) Suppose instead that Fritz's family $\mathcal{F}$ has infinite VC dimension. Does this necessarily make Fritz's hypothesis unfalsifiable? Explain.

   d. (5) A true hypothesis can also be falsifiable. In that case, it will never actually be falsified, no matter how much data we see. Popper insisted that even if a hypothesis is falsifiable, and it is not falsified by a large amount of data, we still do not have reason to think the hypothesis is true. Many people have found this idea of his deeply wrong-headed. Does the generalization error bound using VC dimension show he was wrong?

[1] Popper didn't think every falsifiable idea was a *good* scientific hypothesis, but he did insist on falsifiability as a minimum for science. Also, in case you're wondering whether the falsifiability criterion is itself falsifiable, Popper offered it more as a proposal about how we should use words like "science", i.e., not a theory about the world.

5. (1) How much time did you spend on this problem set?

Presentation rubric (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision.