CSC 588: Homework 2

Chicheng Zhang

February 26, 2022

Please complete the following exercises and read the following instructions carefully.

- Your solutions to these problems will be graded based on both correctness and clarity. Your arguments should be clear: there should be no room for interpretation about what you are writing. Otherwise, I will assume that they are wrong, and grade accordingly.
- If you feel unable to make progress on any of the questions, you can post your questions on Piazza. Try posing your questions to be as general as possible, so that it can promote discussion among the class.
- You are encouraged to discuss the homework questions with your classmates, but the discussions should only be at a high level, and you should write your solutions in your own words. For every question you have had discussions on, please mention explicitly whom you have discussed with; otherwise it may be counted as academic integrity violation.
- For detailed homework policies, please read the course syllabus, available on the course website.

This homework is due on Mar 15, 2022, 5pm MST, on gradescope.

Problem 1 (10pts)

Solve the following exercises from the lectures:

1. Show that the class of non-homogenenous linear classifiers

   $$\mathcal{H} = \left\{ h_{w,b}(x) = I(\langle w, x \rangle + b > 0) : w \in \mathbb{R}^d, b \in \mathbb{R} \right\}$$

   has VC dimension $d + 1$.

2. Suppose $\mathcal{H} \subset (\mathcal{X} \to \{-1, +1\})$ is a set of binary classifiers, and $\mathcal{F} = \{\ell_h : h \in \mathcal{H}\} \subset (\mathcal{X} \times \{-1, +1\} \to \{0, 1\})$ (where $\ell_h(x, y) = I(h(x) \neq y)$) is its induced zero-one loss function class. Prove that for all natural numbers n, $\mathcal{S}(\mathcal{F}, n) = \mathcal{S}(\mathcal{H}, n)$.

3. In the proof of the uniform convergence theorem, we used McDiarmid's Inequality and double sampling trick to reduce bounding the maximum deviation of empirical average to population mean for functions f in class $\mathcal{F} \subset (\mathcal{Z} \to \{0, 1\})$, i.e.,

   $$\sup_{f \in \mathcal{F}} (\mathbb{E}_S f(Z) - \mathbb{E}_D f(Z)) \tag{1}$$

   to bounding the expected maximum deviation of empirical average to another "vaildation" empirical average

   $$\mathbb{E}_{S \sim D^m, S' \sim D^m} \left[\sup_{f \in \mathcal{F}} (\mathbb{E}_S f(Z) - \mathbb{E}_{S'} f(Z)) \right]. \tag{2}$$

(a) Use Massart's finite lemma to directly prove that

$$\mathbb{E}_{S\sim D^m, S'\sim D^m}\left[\sup_{f\in\mathcal{F}}(\mathbb{E}_S f(Z) - \mathbb{E}_{S'} f(Z))\right] \leq \sqrt{\frac{2\ln \mathcal{S}(\mathcal{F}, 2n)}{n}}.$$

(b) If we use this (instead of going over the lecture's proof that relates (2) to $\mathcal{F}$'s Rademacher complexity), can we still obtain an upper bound on (1) that holds with probability $1-\delta$? If yes, what is your upper bound; if not, why?

Problem 2

In this exercise, we explore the relationship between the Rademacher complexity of a hypothesis class $\mathcal{H}$ of binary classifiers and its induced zero-one loss function class $\mathcal{F}$ (here we re-use the notations introduced in Problem 1.2).

1. Show that given a binary classifier $h : \mathcal{X} \to \{-1, +1\}$ and example $(x, y) \in \mathcal{X} \times \{-1, +1\}$,

$$I(h(x) \neq y) = \frac{1}{2}\big(1 - yh(x)\big).$$

2. Given a set of labeled examples $S = ((x_1, y_1), \ldots, (x_n, y_n))$, denote by $U = (x_1, \ldots, x_n)$ its unlabeled part. Show that

$$\mathrm{Rad}_S(\mathcal{F}) = \frac{1}{2}\mathrm{Rad}_U(\mathcal{H}).$$

(This implies that, in the binary classification setup, $\mathrm{Rad}_U(\mathcal{H})$, the ability of classifiers in $\mathcal{H}$ to fit random ±1 labels, is tightly connected to the uniform deviation between training and generalization error of classifiers in $\mathcal{H}$.)

3. Use the above insights to show that, if $T = (z_1, \ldots, z_n)$ is a set of distinct elements, and $\mathcal{B}$ is the set of all functions from T to $\{0, 1\}$, then $\mathrm{Rad}_T(\mathcal{B}) = \frac{1}{2}$. (Hint: if $\mathcal{B}$ were the set of all functions from T to $\{-1, 1\}$ instead, what would be the value of $\mathrm{Rad}_T(\mathcal{B})$?)

Problem 3

Consider the following extension of a result in the class. Fix $v, k \geq 2$. Suppose there is a base hypothesis class $\mathcal{B}$ with $\mathrm{VC}(\mathcal{B}) = v$, and $\mathcal{F}$ is a collection of functions from $\{-1, +1\}^k$ to $\{-1, +1\}$. Now, define

$$\mathcal{H} = \big\{f(b_1(x), \ldots, b_k(x)) : f \in \mathcal{F}, b_1, \ldots, b_k \in \mathcal{B}\big\}.$$

1. Suppose $\mathcal{F}$ is finite. Show that there exists some constant $c > 0$ such that

$$\mathrm{VC}(\mathcal{H}) \leq c \cdot (kv \ln(kv) + \ln|\mathcal{F}|).$$

(This shows that, allowing the "aggregating function" f to vary within a class $\mathcal{F}$ affects the VC dimension of the composite class $\mathcal{H}$, but only mildly (incurring an additive $\ln|\mathcal{F}|$ factor)).

2. Suppose $\mathcal{F}$ is infinite but has a finite VC dimension d, can you still give a good upper bound on $\mathrm{VC}(\mathcal{H})$? If yes, what is your upper bound; if not, why?

Hint: Same as the in-class example, use Sauer's Lemma, along with the relationship between VC dimension and growth function.

Problem 4

Project proposal: in a few paragraphs, describe your plans for the course project. Detailed instructions can be found in the course website: `https://zcc1307.github.io/courses/csc588sp22/project.html`.