

# CSC 665: Homework 2

Chicheng Zhang

September 29, 2019

Please complete the following set of exercises. You must write down your solutions **on your own**. If you have discussed with your classmates on any of the questions, please indicate so in your solutions. The homework is due **on Oct 10, 12:30pm, on Gradescope**. You are free to cite existing theorems from the textbook and course notes.

## Problem 1

Do Exercise 2.3 in (Shalev-Shwartz and Ben-David, 2014). For item 2, you can assume that the joint distribution of  $(X_1, X_2)$  is continuous over  $\mathbb{R}^2$ .

## Problem 2

1. Show the following inequality: for positive  $a, b$  and  $x$ , if  $x > 2a \ln(2a) + 2b$ , then  $x > a \ln x + b$ .
2. Show the following basic inequality: for  $n, d$  such that  $n \geq 2$  and  $n \geq d$ ,  $\binom{n}{\leq d} \leq n^{d+1}$ .
3. Consider  $l$  hypothesis classes  $\mathcal{H}_1, \mathcal{H}_2, \dots, \mathcal{H}_l$ , where  $\text{VC}(\mathcal{H}_i) = v \geq 1$ . Define  $\mathcal{H} \triangleq \cup_{i=1}^l \mathcal{H}_i$ . Show that there exists some constant  $c > 0$  such that

$$\text{VC}(\mathcal{H}) \leq c \cdot (v \ln(v) + \ln(l)).$$

4. Let  $\mathcal{H} = \{\text{sign}(\langle w, x \rangle) : w \in \mathbb{R}^d, |\{i : w_i \neq 0\}| = k\}$  be the set of  $k$ -sparse homogenous linear classifiers in  $\mathbb{R}^d$ , where  $k \leq d$ . Show that there exists some constant  $c > 0$  such that

$$\text{VC}(\mathcal{H}) \leq c \cdot (k \ln d).$$

5. Consider  $l$  hypothesis classes  $\mathcal{H}_1, \dots, \mathcal{H}_l$ , where  $\text{VC}(\mathcal{H}_i) = d_i \geq 1$ . Suppose  $f$  is a function from  $\{\pm 1\}^l$  to  $\{\pm 1\}$  (for example, the majority function  $f(z_1, \dots, z_l) = \text{sign}(\sum_{i=1}^l z_i)$  or the parity function  $f(z_1, \dots, z_l) = \prod_{i=1}^l z_i$ ). Define  $\mathcal{H} \triangleq \{f(h_1(x), \dots, h_l(x)) : h_1 \in \mathcal{H}_1, \dots, h_l \in \mathcal{H}_l\}$ . Show that there exists some constant  $c > 0$  such that

$$\text{VC}(\mathcal{H}) \leq c \left( \sum_{i=1}^l d_i \right) \ln \left( \sum_{i=1}^l d_i \right).$$

### Problem 3

In this exercise, we will show that, under the *realizable setting*, with hypothesis class  $\mathcal{H}$  having VC dimension  $d$ , ERM (in fact, the consistency algorithm) will have a PAC sample complexity of  $O\left(\frac{1}{\epsilon}(d \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta})\right)$ . Suppose  $S = \{Z_1, \dots, Z_m\}$  a set of  $m$  training examples drawn iid from distribution  $D$ , where each  $Z_i = (X_i, Y_i)$  is a labeled example. In addition,  $\mathcal{F} = \{\mathbf{1}(h(x) \neq y) : h \in \mathcal{H}\}$  is the zero-one loss function class. Our proof will mostly follow the steps for showing agnostic PAC sample complexity given in the lecture.

1. **Double Sampling Trick.** Fix a training set  $S$ . Suppose  $\mathbb{E}_S f(Z) = 0$  and  $\mathbb{E}_D f(Z) \geq \epsilon$ . Show that for a fresh set of random examples  $S'$  of size  $m$  ( $m \geq \frac{16}{\epsilon}$ ) sampled iid from  $D$ :

$$\mathbb{P}_{S' \sim D^m} \left( \mathbb{E}_{S'} f(Z) \geq \frac{\epsilon}{2} \right) \geq \frac{1}{2}.$$

2. **Conditioning.** Denote by events

$$E' \triangleq \left\{ \text{there exists } f \in \mathcal{F}, \mathbb{E}_S f(Z) = 0, \mathbb{E}_{S'} f(Z) \geq \frac{\epsilon}{2} \right\},$$

$$E \triangleq \left\{ \text{there exists } f \in \mathcal{F}, \mathbb{E}_S f(Z) = 0, \mathbb{E}_D f(Z) \geq \epsilon \right\}.$$

Show  $\mathbb{P}_{S, S' \sim D^m}(E' | E) \geq \frac{1}{2}$ , and conclude that  $\mathbb{P}_{S \sim D^m}(E) \leq 2\mathbb{P}_{S, S' \sim D^m}(E')$ .

3. **Symmetrization.** Introduce  $\sigma = (\sigma_1, \dots, \sigma_m)$  where each  $\sigma_i \in \{\pm 1\}$ . Denote by

$$(W_i, W'_i) = \begin{cases} (Z_i, Z'_i) & \sigma_i = +1, \\ (Z'_i, Z_i) & \sigma_i = -1. \end{cases}$$

Show that

$$\mathbb{P}_{S, S' \sim D^m}(E') = \mathbb{P}_{S, S' \sim D^m, \sigma \sim R^n} \left( \text{exists } f \in \mathcal{F}, \sum_{i=1}^m f(W_i) = 0, \sum_{i=1}^m f(W'_i) \geq \frac{m\epsilon}{2} \right),$$

where  $R$  is the Rademacher distribution, i.e. uniform in  $\{\pm 1\}$ .

4. **The randomness in Rademacher random variables.** Fix two size  $m$  training sets  $S$  and  $S'$ . Show that for a fixed classifier  $f$  in  $\mathcal{F}$ ,

$$\mathbb{P}_{\sigma \sim R^n} \left( \sum_{i=1}^m f(W_i) = 0, \sum_{i=1}^m f(W'_i) \geq \frac{m\epsilon}{2} \right) \leq \exp\left(-\frac{m\epsilon}{4}\right).$$

5. Use the above items to conclude that for  $m \geq \frac{16}{\epsilon}$ ,

$$\mathbb{P}_{S \sim D^m}(\text{there exists } f \in \mathcal{F}, \mathbb{E}_S f(Z) = 0, \mathbb{E}_D f(Z) \geq \epsilon) \leq 2\mathcal{S}(\mathcal{F}, 2n) \exp\left\{-\frac{m\epsilon}{4}\right\}.$$

In addition, show that ERM has a PAC sample complexity of  $O\left(\frac{1}{\epsilon}(d \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta})\right)$ .

## Problem 4

In this exercise, we develop sample complexity lower bounds for *realizable* PAC learning using Le Cam's method and Assouad's method. Suppose hypothesis class  $\mathcal{H}$  has VC dimension  $d \geq 2$ , and it shatters examples  $z_0, z_1, \dots, z_{d-1}$ . In addition, suppose  $\epsilon, \delta \in (0, \frac{1}{8})$  are target error and target failure probability. A learning algorithm  $\mathcal{A}$  is a mapping from training set  $S$  to  $\{\pm 1\}$ . In the following, you can use the elementary fact that for  $x \in (0, \frac{1}{2})$ ,  $e^{-x} \geq 1 - x \geq e^{-2x}$ .

1. Consider  $D_{-1}$  and  $D_{+1}$  as follows: for every  $i$  in  $\{\pm 1\}$ ,

$$D_i(x, y) = \begin{cases} 1 - 2\epsilon, & (x, y) = (z_0, -1), \\ 2\epsilon, & (x, y) = (z_1, i), \\ 0 & \text{otherwise.} \end{cases}$$

Note that  $\min_{h \in \mathcal{H}} \text{err}(h', D_i) = 0$  for both  $i \in \{\pm 1\}$ . For every  $j$ , denote by  $P_j((x_i, y_i)_{i=1}^m) = \prod_{i=1}^m D_j(x_i, y_i)$  the distribution over training sets (observations). Use Le Cam's method to show that for any hypothesis tester  $f$ , there exists an  $i$  in  $\{\pm 1\}$ , such that

$$\mathbb{P}_i(f(S) \neq I) > \frac{1}{2}(1 - 4\epsilon)^m.$$

2. Conclude that for any learning algorithm  $\mathcal{A}$ , if sample size  $m \leq \frac{1}{4\epsilon} \ln \frac{1}{4\delta}$ , then there exists an  $i$  in  $\{\pm 1\}$ ,

$$\mathbb{P}_i(\text{err}(\hat{h}, D_i) > \epsilon) > \delta.$$

3. For every  $\tau \in \{\pm 1\}^{d-1}$ , consider  $D_\tau$  as follows:

$$D_\tau(x, y) = \begin{cases} 1 - 4\epsilon, & (x, y) = (z_0, -1), \\ \frac{4\epsilon}{d-1}, & (x, y) = (z_i, \tau_i) \text{ for some } i \in \{1, \dots, d-1\}, \\ 0 & \text{otherwise.} \end{cases}$$

Note that  $\min_{h \in \mathcal{H}} \text{err}(h', D_\tau) = 0$  for all  $\tau \in \{\pm 1\}^{d-1}$ . For every  $\tau$ , denote by  $P_\tau((x_i, y_i)_{i=1}^m) = \prod_{i=1}^m D_\tau(x_i, y_i)$  the distribution over training sets (observations).

Use Assouad's method to show that for any hypothesis tester  $f_1, \dots, f_{d-1}$ , there exists  $\tau \in \{\pm 1\}^{d-1}$ ,

$$\mathbb{E}_\tau \left[ \sum_{j=1}^{d-1} \mathbf{1}(f_j(S) \neq \tau_j) \right] > \frac{d-1}{2} \left( 1 - \frac{4\epsilon}{d-1} \right)^m.$$

4. Conclude that for any learning algorithm  $\mathcal{A}$ , suppose that sample size  $m \leq \frac{d-1}{128\epsilon}$ , then there exists a  $\tau \in \{\pm 1\}^d$ , such that

$$\mathbb{P}_\tau(\text{err}(\hat{h}, D_\tau) > \epsilon) > \frac{1}{4}.$$

## Hints

2.1 Use the elementary fact that  $\ln(z) \leq z - 1$  for  $z = \frac{x}{2a}$ .

2.2 Use the elementary fact that  $\binom{n}{i} \leq n^i$ .

- 2.3 (1) consider  $S$  of size  $n$  shattered by  $\mathcal{H}$ . We know that  $|\Pi_{\mathcal{H}}(S)| = 2^n$ . Use Sauer's Lemma to obtain an upper bound on  $|\Pi_{\mathcal{H}}(S)|$  in terms of  $v$ . (2) consider using the contrapositive of item 1.
- 2.4 Write  $\mathcal{H}$  as a union of  $\binom{d}{k}$  hypothesis classes, each of which has VC dimension  $k$ , then apply item 3.)
- 3.1 Use Chernoff bound for Bernoulli distributions (the version with exponent  $-\frac{m p \mu^2}{4}$ ).
- 3.4 Consider three cases: (1) there exists some  $i$ ,  $(f(Z_i), f(Z'_i)) = (1, 1)$ ; (2)  $\sum_{i=1}^m f(Z_i) + f(Z'_i) < \frac{m\epsilon}{2}$ ; (3) both (1) and (2) are not satisfied. Observe that in the first two cases, the probability is identically zero.
- 4.1 Consider observation  $S = ((z_0, -1), \dots, (z_0, -1))$ . Show that  $\mathbb{P}_{-1}(S) = \mathbb{P}_{+1}(S)$ .
- 4.2 Define an appropriate hypothesis tester  $f$  that depends on  $\mathcal{A}$ .
- 4.3 Define  $A_j = \{S = (x_i, y_i)_{i=1}^m : x_i \neq z_j \text{ for all } i\}$ . Show that for every  $\tau \stackrel{j}{\sim} \tau'$ ,  $\mathbb{P}_{\tau}(S) = \mathbb{P}_{\tau'}(S)$  for all  $S$  in  $A_j$ . In addition, for  $\sigma \in \{\tau, \tau'\}$ ,  $\mathbb{P}_{\sigma}(S \in A_j) > \frac{7}{8}$ . Intuitively, seeing only examples other than  $z_j$  does not help determining the optimal classifier's labeling on  $z_j$ .
- 4.4 First show that  $\sum_{j=1}^{d-1} \mathbf{1}(f_j(S) \neq \tau_j) > \frac{d-1}{2}$  with probability  $> \frac{1}{4}$ . Then define an appropriate hypothesis tester  $f = (f_1, \dots, f_{d-1})$  that depends on  $\mathcal{A}$ .

## Problem 5 (No need to submit)

In this problem, we develop an alternative proof of Sauer's Lemma: any hypothesis class  $\mathcal{H}$  with VC dimension  $d$  can have at most  $\binom{n}{\leq d}$  labelings on any dataset  $S = \{z_1, \dots, z_n\}$ . Throughout, we will be using the notation that

$$\binom{\{1, \dots, n\}}{d+1} \triangleq \{(i_1, \dots, i_{d+1}) : 1 \leq i_1 < \dots < i_{d+1} \leq n\}$$

to denote the set of  $(d+1)$ -tuples whose entries are distinct. Note that  $\left| \binom{\{1, \dots, n\}}{d+1} \right| = \binom{n}{d+1}$ .

1. Show that for any indices  $I = (i_1, \dots, i_{d+1}) \in \binom{\{1, \dots, n\}}{d+1}$ , there exists a string  $s_I \in \{\pm 1\}^{d+1}$ , such that none of the labelings in

$$L_I = \{b \in \{\pm 1\}^n : (b_{i_1}, \dots, b_{i_{d+1}}) = s_I\}$$

are achievable by classifiers in  $\mathcal{H}$ .

2. Show the following basic facts:

- (a) For a finite set  $A$  and a function  $f$ , denote by  $f(A) = \{f(a) : a \in A\}$ . Then  $|f(A)| \leq |A|$ , where  $|B|$  denotes the cardinality of set  $B$ .
- (b) Suppose  $\mathcal{I}$  is a set of indices. Given a collection of sets  $\{L_I\}_{I \in \mathcal{I}}$  and a function  $f$ ,

$$\left| \bigcup_{I \in \mathcal{I}} f(L_I) \right| \leq \left| \bigcup_{I \in \mathcal{I}} L_I \right|. \quad (1)$$

3. Use the above two facts to conclude that

$$\left| \bigcup_{I \in \binom{\{1, \dots, n\}}{d+1}} L_I \right| \geq \sum_{i=d+1}^n \binom{n}{i}.$$

(Hint: consider functions  $f_1, \dots, f_n$ , where  $f_i(s_1, \dots, s_n) = (s_1, \dots, s_{i-1}, -1, s_{i+1}, \dots, s_n)$  is the function that sets a length  $n$  string's  $i$ -th entry to  $-1$ . Iteratively applying Equation (1) for  $f_1, \dots, f_n$ , what do you get?)

4. Use item 3 to conclude that  $|\Pi_{\mathcal{H}}(S)| \leq \binom{n}{\leq d}$ .