

CSC 588: Homework 1

Chicheng Zhang

February 3, 2021

- This homework is due on Feb 18 on gradescope.
- 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.
- Feel free to use existing theorems from the course notes / the textbook.

Problem 1

1. Show that in \mathbb{R}^d , we can find at most d vectors that are pairwise orthogonal.
2. Next we will use Hoeffding's inequality to show that, in sharp contrast to above, it is possible to find exponentially many ($n = e^{\Omega(d)}$) vectors that are *almost* orthogonal; that is, there exist x_1, \dots, x_n in \mathbb{R}^d , such that for every pair (i, j) ($1 \leq i < j \leq n$), the angle between x_i and x_j is between 89° and 91° . To this end, consider the following randomized construction:

Draw n random vectors X_1, X_2, \dots, X_n in \mathbb{R}^d , where for each i , $X_i = \frac{1}{\sqrt{d}}(Z_{i,1}, \dots, Z_{i,d})$. Here $\{Z_{i,j}\}_{i \in \{1, \dots, n\}, j \in \{1, \dots, d\}}$'s are iid, and $Z_{i,j}$ takes value 1 with probability $1/2$, and takes value -1 with probability $1/2$.

- (a) Check that all X_i 's have unit length, i.e. $\|X_i\|_2 = 1$.
- (b) Use Hoeffding's Inequality to show that for any fixed pair $i, j \in \{1, \dots, n\}$, $i < j$,

$$\mathbb{P}(|\langle X_i, X_j \rangle| \geq \sin(1^\circ)) \leq 2 \exp\{-0.00014d\}.$$

- (c) Suppose $n = \exp\{0.00005d\}$. Use the union bound to show that

$$\mathbb{P}(\forall i < j, \text{ the angle between } X_i \text{ and } X_j \text{ is in } [89^\circ, 91^\circ]) > 0.$$

(Note that this proves the claim at the beginning of item 2.)

Problem 2

Suppose we have an algorithm \mathcal{B} that learns hypothesis class \mathcal{H} in the following sense. There exists a function $f : (0, 1) \rightarrow \mathbb{N}$, such that for any distribution D realizable by \mathcal{H} , for any $\epsilon > 0$, if \mathcal{B} draws $m \geq f(\epsilon)$ iid training examples from D , then with probability at least $\frac{1}{2}$, \mathcal{B} returns a classifier whose generalization error on D is at most ϵ .

Now, given \mathcal{B} , and the ability to draw fresh training examples, how can you design an algorithm \mathcal{A} that (ϵ, δ) -PAC learns \mathcal{H} for any ϵ and δ ? What is \mathcal{A} 's sample complexity?

Problem 3

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

$$\mathcal{H} = \left\{ h_{w,b}(x) = 2I(\langle w, x \rangle + b > 0) - 1 : w \in \mathbb{R}^d, b \in \mathbb{R} \right\}$$

has VC dimension $d + 1$.

2. Define the class of polynomial threshold functions

$$\mathcal{H}_n = \{ 2I(p(x) > 0) - 1 : p \text{ is a polynomial of } x \text{ of degree } \leq n \}$$

(where $x \in \mathbb{R}$). What is the VC dimension of \mathcal{H} ?

3. Suppose we have a natural number $v \geq 1$, and l hypothesis classes $\mathcal{H}_1, \mathcal{H}_2, \dots, \mathcal{H}_l$, where for every i , $VC(\mathcal{H}_i) \leq v$. Define $\mathcal{H} \triangleq \cup_{i=1}^l \mathcal{H}_i$. Show that there exists some constant $c > 0$ such that

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

Problem 4

In this exercise, we will unify the analysis of the empirical risk minimization algorithm in realizable and agnostic settings to recover the $O(\frac{1}{\epsilon})$ -type sample complexity and the $O(\frac{1}{\epsilon^2})$ -style sample complexity given in the class, using Bernstein's Inequality. Suppose \mathcal{H} is a finite hypothesis class, D is a distribution over labeled examples, and S is a training set of size m drawn iid from D . Denote by $\nu^* = \min_{h \in \mathcal{H}} \text{err}(h, D)$ as the optimal generalization error, and \hat{h} the output of the empirical risk minimization algorithm.

1. Using the Bernstein's Inequality we have seen in the class, show that with probability $1 - \delta$, for all classifiers h in \mathcal{H} ,

$$\begin{aligned} \text{err}(h, S) &\leq \text{err}(h, D) + \sqrt{\text{err}(h, D) \frac{4 \ln \frac{2|\mathcal{H}|}{\delta}}{m}} + \frac{2 \ln \frac{2|\mathcal{H}|}{\delta}}{m}, \\ \text{err}(h, D) &\leq \text{err}(h, S) + \sqrt{\text{err}(h, S) \frac{4 \ln \frac{2|\mathcal{H}|}{\delta}}{m}} + \frac{12 \ln \frac{2|\mathcal{H}|}{\delta}}{m}. \end{aligned}$$

(Hint: to get the second inequality, you can use the elementary fact that for $A, B, C > 0$, $A \leq B + C\sqrt{A}$ implies $A \leq B + C^2 + C\sqrt{B}$.)

2. Show that with probability $1 - \delta$, \hat{h} satisfies that

$$\text{err}(\hat{h}, D) \leq \nu^* + c_1 \sqrt{\frac{2 \ln \frac{2|\mathcal{H}|}{\delta}}{m} \nu^*} + c_2 \frac{\ln \frac{2|\mathcal{H}|}{\delta}}{m},$$

for some positive constants c_1 and c_2 . (Hint: you may find the following elementary facts useful: for $A, B > 0$, $\sqrt{AB} \leq A + B$, $\sqrt{A + B} \leq \sqrt{A} + \sqrt{B}$. The tightness of constants c_1 and c_2 won't be graded.)

3. Use the above item to conclude that:

- (a) There exists a function m_A such that $m_A(\epsilon, \delta) = O(\frac{\ln |\mathcal{H}| + \ln \frac{1}{\delta}}{\epsilon^2})$, when $m \geq m_A(\epsilon, \delta)$, for all distributions D , $\text{err}(\hat{h}, D) \leq \nu^* + \epsilon$ with probability $1 - \delta$.
- (b) There exists a function m_R such that $m_R(\epsilon, \delta) = O(\frac{\ln |\mathcal{H}| + \ln \frac{1}{\delta}}{\epsilon})$, when $m \geq m_R(\epsilon, \delta)$, for all distributions D such that $\nu^* = 0$, $\text{err}(\hat{h}, D) \leq \epsilon$ with probability $1 - \delta$.

Problem 5

How much time did it take you to complete this homework?