CSC 665: Midterm

Chicheng Zhang

November 10, 2019

Please complete the following set of problems. You must do the exercises completely on your own (no collaboration allowed this time). The exam is due **on Oct 24, 12:30pm, on Gradescope**. You are free to cite existing theorems from the textbooks and course notes.

Problem 1

Define $\mathcal{H} = \left\{ \text{sign}(p(x)) : p \text{ is a polynomial of } x \text{ of degree} \leq n \right\}$ (where $x \in \mathbb{R}$). Here $\text{sign}(z) = 2\mathbf{1}(z > 0) - 1$. What is the VC dimension of $\mathcal{H}$?

Solution

We show $\text{VC}(\mathcal{H}) = n + 1$.

1. We first show $\text{VC}(\mathcal{H}) \geq n+1$, i.e. there exists $n+1$ points that are shattered by $\mathcal{H}$. Pick $n+1$ distinct numbers $x_1, \ldots, x_{n+1}$ in $\mathbb{R}$. By a standard fact in analysis, for any values $y_1, \ldots, y_{n+1}$ in $\mathbb{R}$, there exists a unique degree-at-most-n polynomial p that passes all points $(x_i, y_i)_{i=1}^{n+1}$ (this can be shown using Cramer's rule or Lagrange polynomials).

   Specifically, for any $y_1, \ldots, y_{n+1}$ in $\{\pm 1\}^{n+1}$, there exists p of degree of at most n such that $p(x_i) = y_i$; as a consequence, $\text{sign}(p)$ in $\mathcal{H}$ also achieves the labeling of $(y_1, \ldots, y_{n+1})$ on $(x_1, \ldots, x_{n+1})$. Therefore, $\mathcal{H}$ shatters $x_1, \ldots, x_{n+1}$, a set of size $n + 1$.

2. We now show $\text{VC}(\mathcal{H}) \leq n+1$. Note that an alternative way of writing $\mathcal{H}$ is $\left\{ \text{sign}(\langle a, \phi(x) \rangle) : a \in \mathbb{R}^{n+1} \right\}$, where $\phi(x) = (1, x, \ldots, x^n)$. If there are $n+2$ points $x_1, \ldots, x_{n+2}$ shattered by $\mathcal{H}$, this also means that $\phi(x_1), \ldots, \phi(x_{n+2})$ is shattered by $\mathcal{F} = \left\{ \text{sign}(\langle a, x \rangle) : a \in \mathbb{R}^{n+1} \right\}$. But in class, we have seen that $\mathcal{F}$ has VC dimension $n + 1$, contradiction.

Problem 2

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

Now, given $\mathcal{B}$ and the ability of drawing fresh training examples, how can you design an algorithm $\mathcal{A}$ that (ϵ, δ)-PAC learns $\mathcal{H}$ for any ϵ, δ? What is its sample complexity? (You may want to run $\mathcal{B}$ multiple times.)

Solution

Consider the following algorithm $\mathcal{A}(\epsilon, \delta)$:

1. Repeated $\mathcal{B}(\epsilon/2)$ for k times, getting classifiers $h_1 \ldots, h_k$, where $k = \lceil \log_2 \frac{2}{\delta} \rceil$.

2. $S \leftarrow$ Sample $\frac{32}{\epsilon^2} \ln \frac{4k}{\delta}$ iid examples from D.

3. Select $\hat{h} = \arg\min_{h \in \mathcal{F}} \mathrm{err}(h, S)$, where $\mathcal{F} = \{h_1, \ldots, h_k\}$.

We show that $\mathcal{A}$ outputs a classifier with error ϵ with probability $1-\delta$. Define event $E_1 = \left\{\exists i, \mathrm{err}(h_i, D) \le \epsilon/2\right\}$, and event $E_2 = \left\{\forall h \in \mathcal{F}, |\mathrm{err}(h, S) - \mathrm{err}(h, D)| \le \epsilon/4\right\}$.

First,

$$\mathbb{P}(E_1) = 1 - \prod_{i=1}^{k} \mathbb{P}(\mathrm{err}(h_i, D) \le \epsilon/2) \ge 1 - \frac{1}{2^k} \ge 1 - \delta/2.$$

Second, by Hoeffding's inequality, given $h_1, \ldots, h_k$, as S is a fresh set of examples independent of $h_1, \ldots, h_k$,

$$\mathbb{P}(E_2 | h_1, \ldots, h_k) \ge 1 - \sum_{i=1}^{k} \mathbb{P}(|\mathrm{err}(h, S) - \mathrm{err}(h, D)| > \epsilon/4) = 1 - 2k \cdot e^{-2m \cdot \frac{\epsilon^2}{16}} \ge 1 - \delta/2.$$

Therefore, $\mathbb{P}(E_2) \ge 1 - \delta/2$, and consequently, by union bound, $\mathbb{P}(E_1 \cap E_2) \ge 1 - \delta$.

Now on event E_2, we have that

$$\mathrm{err}(\hat{h}, D) \le \mathrm{err}(\hat{h}, S) + \epsilon/4 = \min_{i=1}^{k} \mathrm{err}(h_i, S) + \epsilon/4 \le \min_{i=1}^{k} \mathrm{err}(h_i, D) + \epsilon/2.$$

Therefore, on event $E_1 \cap E_2$,

$$\mathrm{err}(\hat{h}, D) \le \min_{i=1}^{k} \mathrm{err}(h_i, D) + \epsilon/2 \le \epsilon/2 + \epsilon/2 = \epsilon.$$

As can been seen from the description of $\mathcal{A}$, it has a sample complexity of

$$k m(\epsilon/2) + \frac{32}{\epsilon^2} \ln \frac{4k}{\delta} = \lceil \log_2 \frac{2}{\delta} \rceil \cdot m(\epsilon/2) + \frac{32}{\epsilon^2} \ln \frac{4k}{\delta}.$$

Problem 3

Suppose $X_1, \ldots, X_n$ is a sequence of n iid random variables, and let $\sigma^2 = \mathrm{var}(X_i)$ and $\mu = \mathbb{E}(X_i)$. Suppose $n = mk$ for some integer m and odd integer $k \ge 20 \ln \frac{1}{\delta}$. Denote by

$$\hat{\mu} = \mathrm{median}(\hat{\mu}_1, \ldots, \hat{\mu}_k),$$

where $\hat{\mu}_i = \frac{1}{m} \sum_{j=(i-1)m+1}^{im} X_j$.

1. Show that for every j,

$$\mathbb{P}(|\hat{\mu}_j - \mu| \le \frac{2\sigma}{\sqrt{m}}) \ge \frac{3}{4}.$$

2. Show that

$$\mathbb{P}(|\hat{\mu} - \mu| \le \frac{2\sigma}{\sqrt{m}}) \ge 1 - \delta.$$

Solution

1. $\mathrm{var}(\hat{\mu}_j) = \frac{1}{m^2}\sum_{j=(i-1)m+1}^{im} \mathrm{var}(X_j) = \frac{1}{m^2} \cdot m \cdot \sigma^2 = \frac{\sigma^2}{m}$.
   By Chebyshev's Inequality,
   $$\mathbb{P}(|\hat{\mu}_j - \mu| > \frac{2\sigma}{\sqrt{m}}) \leq \frac{\delta(\hat{\mu}_j)}{(\frac{2\sigma}{\sqrt{m}})^2} \leq \frac{1}{4}.$$
   The stated result follows by taking the complement of the event considered.

2. Denote by
   $$Z_j = \mathbf{1}(|\hat{\mu}_j - \mu| \leq \frac{2\sigma}{\sqrt{m}}).$$
   Note that $Z_1, \ldots, Z_k$ are independent, and has mean p at least $\frac{3}{4}$. By Hoeffding's inequality, when $k \geq 20 \ln \frac{1}{\delta}$,
   $$\mathbb{P}(\sum_{j=1}^{k} Z_j \geq \frac{k}{2}) \geq 1 - \exp\left(2k(p - \frac{1}{2})^2\right) \geq 1 - \delta.$$
   Now, consider the event $E = \left\{\sum_{j=1}^{k} Z_j \geq \frac{k}{2}\right\}$. We claim that under E, $\hat{\mu}$ would be inside interval $I = [\mu - \frac{2\sigma}{\sqrt{m}}, \mu + \frac{2\sigma}{\sqrt{m}}]$. Indeed, if $\hat{\mu}$ is outside I, then two cases would happen:
   - $\hat{\mu} < \mu - \frac{2\sigma}{\sqrt{m}}$. As $\hat{\mu}$ is the median of μ_i's, at least half of μ_i's would also be smaller than $\mu - \frac{2\sigma}{\sqrt{m}}$, contradiction to the fact that E happens.
   - $\hat{\mu} > \mu - \frac{2\sigma}{\sqrt{m}}$. Symmetrically, this would also contradict with the fact that E happens.

   In summary, in event E, which happens with probability $1 - \delta$, $\hat{\mu} \in [\mu - \frac{2\sigma}{\sqrt{m}}, \mu + \frac{2\sigma}{\sqrt{m}}]$.

Remark. You may wonder: *what is the motivation of this question?* The estimator $\hat{\mu}$ is interesting in that it gives a better mean estimator than naive sample mean for heavy-tailed random variables. Sample mean can sometimes have bad concentration properties; see [1, 2] for discussions.

References

[1] Jean-Yves Audibert, Olivier Catoni, et al. Robust linear least squares regression. *The Annals of Statistics*, 39(5):2766–2794, 2011.

[2] Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research*, 17(1):543–582, 2016.