CS 446 MJT — Homework 4

haoyuan9

Version 2

Instructions.

- Homework is due **Tuesday, April 2, at 11:59pm**; no late homework accepted.
- Everyone must submit individually at gradescope under `hw4`. (There is no `hw4code`!)
- The "written" submission at `hw4` **must be typed**, and submitted in any format gradescope accepts (to be safe, submit a PDF). You may use LaTeX, markdown, google docs, MS word, whatever you like; but it must be typed!
- When submitting at `hw4`, gradescope will ask you to mark out boxes around each of your answers; please do this precisely!
- Please make sure your NetID is clear and large on the first page of the homework.
- Your solution **must** be written in your own words. Please see the course webpage for full academic integrity information. Briefly, you may have high-level discussions with at most 3 classmates, whose NetIDs you should place on the first page of your solutions, and you should cite any external reference you use; despite all this, your solution must be written in your own words.

1. **VC dimension.**

This problem will show that two different classes of predictors have infinite VC dimension.

Hint: to prove infinite $\mathsf{VC}(\mathcal{H}) = \infty$, it is usually most convenient to show $\mathsf{VC}(\mathcal{H}) \geq n$ for all n.

(a) Let $\mathcal{F} := \{\boldsymbol{x} \mapsto 2 \cdot \mathbb{1}[\boldsymbol{x} \in C] - 1 : C \subseteq \mathbb{R}^d \text{ is convex}\}$ denote the set of all classifiers whose decision boundary is a convex subset of $\mathbb{R}^d$ for $d \geq 2$. Prove $\mathsf{VC}(\mathcal{F}) = \infty$.

Hint: Consider data examples on the unit sphere $\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| = 1\}$.

(b) Given $x \in \mathbb{R}$, let sgn denote the sign of x: $\text{sgn}(x) = 1$ if $x \geq 0$ while $\text{sgn}(x) = -1$ if $x < 0$.

Let $\sigma > 0$ be given, and define $\mathcal{G}_\sigma$ to be the set of (sign of) all RBF classifiers with bandwidth σ, meaning

$$\mathcal{G}_\sigma := \left\{ \boldsymbol{x} \mapsto \text{sgn}\left(\sum_{i=1}^{m} \alpha_i \exp\left(-\|\boldsymbol{x} - \boldsymbol{x}_i\|^2/(2\sigma^2)\right) \right) \ : \ m \in \mathbb{Z}_{\geq 0},\ \boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \mathbb{R}^d,\ \boldsymbol{\alpha} \in \mathbb{R}^m \right\}.$$

Prove $\mathsf{VC}(\mathcal{G}_\sigma) = \infty$.

Remark: the sign of 0 is not important: you have the freedom to choose some nice data examples and avoid this case.

Hint: remember in `hw3` it is proved that if σ is small enough, the RBF kernel SVM is close to the 1-nearest neighbor predictor. In this problem, σ is fixed, but you have the freedom to choose the data examples. If the distance between data examples is large enough, the RBF kernel SVM could still be close to the 1-nearest neighbor predictor. Make sure to have an explicit construction of such a dataset.

Solution. *(Your solution here.)*

(a) For any d, choose n examples on the unit sphere $\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| = 1\}$. Then for any labelings, we can always choose one set of data samples A and define C as the minimum convex hull of A. So $\mathcal{F}$ can shatter any n samples on the unit sphere. Hence, $\mathsf{VC}(\mathcal{F}) \geq n$ for all n. So we have $\mathsf{VC}(\mathcal{F}) = \infty$.

(b) Choose n data samples, where $\|\boldsymbol{x}_{1i} - \boldsymbol{x}_{1j}\|^2 = \infty$ for any pair of i, j where $i \neq j$. Then construct $\mathcal{G}_\sigma$ by making $m = n$, and choosing n $\boldsymbol{x}_i$ where $\|\boldsymbol{x}_{2i} - \boldsymbol{x}_{1i}\|^2 = c$. c is a constant.

Just as HW3, we can devided the indicator function by $\exp\left(-\rho^2/2\sigma^2\right)$ without changing the sign, where $\rho = \|\boldsymbol{x}_{2i} - \boldsymbol{x}_{1i}\|^2$. Then the classifier becomes

$$\mathcal{G}_\sigma := \left\{ \boldsymbol{x} \mapsto \text{sgn}\left(\frac{\sum_{i=1}^{m} \alpha_i \exp\left(-\|\boldsymbol{x} - \boldsymbol{x}_{2i}\|^2/(2\sigma^2)\right)}{\exp\left(-\rho^2/2\sigma^2\right)} \right) \ : \ m = n,\ \boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \mathbb{R}^d,\ \boldsymbol{\alpha} \in \mathbb{R}^m \right\}.$$

Since $\|\boldsymbol{x}_{1i} - \boldsymbol{x}_{1j}\|^2 = \infty$, we have $\|\boldsymbol{x}_{2i} - \boldsymbol{x}_{1j}\|^2 = \infty$. Hence

$$\boldsymbol{x}_{1i} \mapsto \text{sgn}\left(\alpha_i\right)$$

Hence, for any labelings for a data sample of size n, we can choose α_i the same sign as x_{1i}. So $\mathsf{VC}(\mathcal{G}_\sigma) \geq n$ and $\mathsf{VC}(\mathcal{G}_\sigma) = \infty$.

2. **Rademacher complexity of linear predictors.**

   Let examples $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ be given with $\|\boldsymbol{x}_i\| \le R$, along with linear functions $\{\boldsymbol{x} \mapsto \boldsymbol{w}^\intercal \boldsymbol{x} : \|\boldsymbol{w}\| \le W\}$. The goal in this problem is to show $\text{Rad}(\mathcal{F}) \le {}^{RW}/\sqrt{n}$.

   (a) For a fixed sign vector $\varepsilon \in \{-1, +1\}^n$, define $\boldsymbol{x}_\varepsilon := \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \epsilon_i$. Show

   $$\max_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(\boldsymbol{x}_i) \le W\|\boldsymbol{x}_\varepsilon\|.$$

   Hint: Cauchy-Schwarz!

   (b) Show $\mathbb{E}_\varepsilon \|\boldsymbol{x}_\varepsilon\|^2 \le {}^{R^2}/n$.

   (c) Now combine the pieces to show $\text{Rad}(\mathcal{F}) \le {}^{RW}/\sqrt{n}$.
   Hint: one missing piece is to write $\|\cdot\| = \sqrt{\|\cdot\|^2}$ and use Jensen's inequality.

Solution. *(Your solution here.)*

(a)

$$\begin{aligned}
\max_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(\boldsymbol{x}_i) &= \boldsymbol{w}_{optim} \frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i \epsilon_i \\
&= \boldsymbol{w}_{optim} \boldsymbol{x}_\epsilon \\
&\le \|\boldsymbol{w}_{optim}\| \; \|\boldsymbol{x}_\epsilon\| \\
&\le W \|\boldsymbol{x}_\epsilon\|
\end{aligned}$$

(b)

$$\begin{aligned}
\mathbb{E}_\varepsilon \|\boldsymbol{x}_\varepsilon\|^2 &= \mathbb{E}_\varepsilon \sum_j \left(\frac{1}{n} \sum_{i=1}^n \; \boldsymbol{x}_{ij} \epsilon_i \right)^2 \\
&= \frac{1}{n^2} \sum_j \mathbb{E}_\varepsilon \left(\sum_{i=1}^n \; \boldsymbol{x}_{ij} \epsilon_i \right)^2 \\
&= \frac{1}{n^2} \sum_j \left(Var(\sum_{i=1}^n \; \boldsymbol{x}_{ij} \epsilon_i) + \mathbb{E}_\varepsilon^2(\sum_{i=1}^n \; \boldsymbol{x}_{ij} \epsilon_i) \right) \\
&= \frac{1}{n^2} \sum_{i=1}^n \sum_j x_{ij}^2 Var(\epsilon_i) \\
&= \frac{1}{n^2} \sum_{i=1}^n \|\boldsymbol{x}_i\|^2 \\
&\le \frac{1}{n^2} n R^2 \\
&= \frac{R^2}{n}
\end{aligned}$$

(c)

$$
\begin{aligned}
\mathrm{Rad}(\mathcal{F}) &= \mathbb{E}_\varepsilon \max_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(\boldsymbol{x}_i) \\
&\leq \mathbb{E}_\varepsilon W \|\boldsymbol{x}_\varepsilon\| \\
&= W \sqrt{(\mathbb{E}_\varepsilon \|\boldsymbol{x}_\varepsilon\|)^2} \\
&\leq W \sqrt{\mathbb{E}_\varepsilon (\|\boldsymbol{x}_\varepsilon\|^2)} \\
&= W \sqrt{\frac{R^2}{n}} \\
&= \frac{WR}{\sqrt{n}}
\end{aligned}
$$

3. **Generalization bounds for a few linear predictors.**

   In this problem, it is always assumed that for any $(\boldsymbol{x}, y)$ sampled from the distribution, $\|\boldsymbol{x}\| \le R$ and $y \in \{-1, +1\}$.

   Consider the following version of the soft-margin SVM:

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \quad \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \frac{1}{n}\sum_{i=1}^{n}\Big[1 - \boldsymbol{w}^\top \boldsymbol{x}_i y_i\Big]_+ = \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \widehat{\mathcal{R}}_{\text{hinge}}(\boldsymbol{w}).$$

   Let $\hat{\boldsymbol{w}}$ denote the (unique!) optimal solution, and $\hat{f}(\boldsymbol{x}) = \hat{\boldsymbol{w}}^\top \boldsymbol{x}$.

   Prove that for any regularization level $\lambda > 0$, with probability at least $1 - \delta$, it holds that

$$\mathcal{R}(\hat{f}) \le \widehat{\mathcal{R}}(\hat{f}) + R\sqrt{\frac{8}{\lambda n}} + 3\left(1 + R\sqrt{2/\lambda}\right)\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

   Hint: use the fact from slide 5/61 of the first ML Theory lecture that $\|\hat{\boldsymbol{w}}\| \le \sqrt{2/\lambda}$, the linear predictor Rademacher complexity bound from the previous problem, and the Rademacher generalization theorem on slide 57 of the final theory lecture.

   Solution. *(Your solution here.)*

1. First, let's proof Hinge Loss is 1-Lipschitz.

   For any $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$, if $\boldsymbol{w}_1 > 1$, then

$$\begin{aligned}
|[1 - \boldsymbol{w}_1^\intercal xy]_+ - [1 - \boldsymbol{w}_2^\intercal xy]_+| &= |[1 - \boldsymbol{w}_2^\intercal xy]_+| \\
&\le |[\boldsymbol{w}_1^\intercal xy - \boldsymbol{w}_2^\intercal xy]_+| \\
&\le |\boldsymbol{w}_1^\intercal xy - \boldsymbol{w}_2^\intercal xy|
\end{aligned}$$

   Similarly, we can prove that when $\boldsymbol{w}_2 > 1$, $|[1 - \boldsymbol{w}_1^\intercal xy]_+ - [1 - \boldsymbol{w}_2^\intercal xy]_+| \le |\boldsymbol{w}_1^\intercal xy - \boldsymbol{w}_2^\intercal xy|$

   When $\boldsymbol{w}_1 \le 1, \boldsymbol{w}_2 \le 1$

$$\begin{aligned}
|[1 - \boldsymbol{w}_1^\intercal xy]_+ - [1 - \boldsymbol{w}_2^\intercal xy]_+| &= |1 - \boldsymbol{w}_1^\intercal xy - 1 + \boldsymbol{w}_2^\intercal xy| \\
&= |\boldsymbol{w}_1^\intercal xy - \boldsymbol{w}_2^\intercal xy|
\end{aligned}$$

   Hence, Hinge Loss using affine $\boldsymbol{f}$ is 1-Lipschitz.

   In addition, since $\|\boldsymbol{x}\| \le R$ and $\|\boldsymbol{w}\| \le \sqrt{\frac{2}{\lambda}}$, we have $0 \le [1 - \boldsymbol{w}^\intercal xy]_+ \le 1 + R\sqrt{\frac{2}{\lambda}}$

   Hence,

$$\begin{aligned}
\mathcal{R}(\hat{f}) &\le \widehat{\mathcal{R}}(\hat{f}) + \text{Rad}(\mathcal{F}) + (b - a)\sqrt{\frac{\ln(1/\delta)}{n}} \\
&\le \widehat{\mathcal{R}}(\hat{f}) + R\sqrt{\frac{2}{\lambda n}} + (1 + R\sqrt{\frac{2}{\lambda}})\sqrt{\frac{\ln(1/\delta)}{n}} \\
&\le \widehat{\mathcal{R}}(\hat{f}) + R\sqrt{\frac{8}{\lambda n}} + 3\left(1 + R\sqrt{2/\lambda}\right)\sqrt{\frac{\ln(2/\delta)}{2n}}.
\end{aligned}$$