CS 446 MJT — Homework 4

junzhew3
haoxuansun8

2019/04/02

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: $\operatorname{sgn}(x) = 1$ if $x \geq 0$ while $\operatorname{sgn}(x) = -1$ 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

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

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) Assume all the data examples are on the unit sphere $\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| = 1\}$.The number of data examples are $\boldsymbol{n}$ ($\boldsymbol{n}$ can be any integer which greater than 1). The $\boldsymbol{m}$ data examples are labeled as +1, the rest $(\boldsymbol{n} - \boldsymbol{m})$ data examples are labeled as -1.

If $\boldsymbol{m} = 0$: we can set $\boldsymbol{C} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| \leq 0.5\}$, so all the points with -1 label are outside the $\boldsymbol{C}$.

if $\boldsymbol{m} = 1$: we can find a subspace $\boldsymbol{C}$ which is tangent to the unit sphere $\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| = 1\}$ at the point which is labeles as +1.

If $2 \leq \boldsymbol{m} \leq n$: Let us connect all the points with +1 label using subset of $\mathbb{R}^{d-1}$ one by one to form a subset of $\mathbb{R}^d$. Since this subset is a Polyhedra, this subset is a convex subset.

So $\boldsymbol{C}$ is equal to this subset. So all the points with +1 label are in the $\boldsymbol{C}$, all the points with -1 label are outside the $\boldsymbol{C}$.

So $\mathsf{VC}(\mathcal{F}) \geq n$ for all n.

SO $\mathsf{VC}(\mathcal{F}) = \infty$.

(b) So I will make all data examples $((\boldsymbol{X}_i, \boldsymbol{Y}_i))_{i=1}^n$ be on a line. The distance between data examples is large enough to make sure that the RBF kernel SVM is close to the 1-nearest neighbor predictor. So in this situation, every data examples are support vectors.

1. Set m is equal to the number of support vectors, so m = n.
2. Set $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is equal to the support vectors, so $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = (\boldsymbol{X}_i)_{i=1}^n$
3. Set $(\alpha_i)_{i=1}^n$ is equal to $(\boldsymbol{Y}_i)_{i=1}^n$.

For each point $\boldsymbol{X}_i$, let $\rho := \min_{j\in n} \|\boldsymbol{X}_i - \boldsymbol{x}_j\|_2$, $T := \{j \in n : \|\boldsymbol{X}_i - \boldsymbol{x}_j\| = \rho\}$

$$\operatorname{sgn}\left(\sum_{j=1}^{n} \alpha_i \exp\left(-\|\boldsymbol{X}_i - \boldsymbol{x}_j\|^2/(2\sigma^2)\right)\right) = \operatorname{sgn}\left(\sum_{j\in T} \alpha_j \exp\left(-\rho^2/2\sigma^2\right)\right). \tag{1}$$

So for each point $\boldsymbol{X}_i$, $\rho := 0$ when $\boldsymbol{X}_i = \boldsymbol{x}_i$, so:

$$\operatorname{sgn}\left(\sum_{j\in T} \hat{\alpha}_j \exp\left(-\rho^2/2\sigma^2\right)\right) = \operatorname{sgn}\left(\alpha_i\right) = \operatorname{sgn}(\boldsymbol{Y}_i). \tag{2}$$

So $\boldsymbol{Y}_i(predicted) = \boldsymbol{Y}_i(True)$. Each data example is correctly labled.

So $\mathsf{VC}(\mathcal{G}_\sigma) \geq n$ for all n.

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

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 $\mathrm{Rad}(\mathcal{F}) \le {}^{RW}/\sqrt{n}$.

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

Solution. *(Your solution here.)*

(a) Using Cauchy-Schwarz inequality:

$$\frac{1}{n}\sum_{i=1}^n \epsilon_i f(\boldsymbol{x}_i) = \frac{1}{n}\sum_{i=1}^n \epsilon_i \boldsymbol{w}^\intercal \boldsymbol{x}_i = \boldsymbol{w}^\intercal \left(\frac{1}{n}\sum_{i=1}^n \epsilon_i \boldsymbol{x}_i\right) \le \|\boldsymbol{w}\|\|\boldsymbol{x}_\varepsilon\| \le W\|\boldsymbol{x}_\varepsilon\| \tag{3}$$

$$So: \quad \max_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(\boldsymbol{x}_i \le W\|\boldsymbol{x}_\varepsilon\| \tag{4}$$

(b) Since each value of ϵ_i has a 1/2 probability of being -1 and 1/2 probability of being +1.

So $\mathbb{E}(\epsilon_i) = 0,\ D(\epsilon_i) = \mathbb{E}(\epsilon_i^2) - \mathbb{E}(\epsilon_i) = 1,$

for any $\boldsymbol{a}, \boldsymbol{b} \in [1, n], \boldsymbol{a} \ne \boldsymbol{b},\ Cov(\epsilon_a, \epsilon_b) = \mathbb{E}[\epsilon_a - \mathbb{E}(\epsilon_a)][\epsilon_b - \mathbb{E}(\epsilon_b)] = \mathbb{E}(\epsilon_a\epsilon_b)$

Since ϵ_a and ϵ_b are independent, so $\mathbb{E}(\epsilon_a\epsilon_b) = \mathbb{E}(\epsilon_a)\mathbb{E}(\epsilon_b) = 0$, so $Cov(\epsilon_a, \epsilon_b) = 0$

So assume $\boldsymbol{x}_i \in \mathbb{R}^r$:

$$\mathbb{E}_\varepsilon\|\boldsymbol{x}_\varepsilon\|^2 = \mathbb{E}_\varepsilon(x_{\varepsilon 1}^2 + \cdots + x_{\varepsilon r}^2) = \mathbb{E}_\varepsilon[(x_{11}\epsilon_1 + \cdots + x_{n1}\epsilon_n)^2 + \cdots + (x_{1r}\epsilon_1 + +x_{nr}\epsilon_n)^2]/n^2 \tag{5}$$

$$= (\mathbb{E}_\varepsilon[(x_{11}\epsilon_1 + \cdots + x_{n1}\epsilon_n)^2] + \cdots + [\mathbb{E}_\varepsilon(x_{1r}\epsilon_1 + \cdots + x_{nr}\epsilon_n)^2])/n^2 \tag{6}$$

$$\mathbb{E}_\varepsilon[(x_{11}\epsilon_1 + \cdots + x_{n1}\epsilon_n)^2] = \mathbb{E}_\varepsilon(x_{11}\epsilon_1 + \cdots + x_{n1}\epsilon_n)^2 + D(x_{11}\epsilon_1 + \cdots + x_{n1}\epsilon_n) \tag{7}$$

$$= (x_{11}\mathbb{E}_\varepsilon(\epsilon_1) + \cdots + x_{n1}\mathbb{E}_\varepsilon(\epsilon_n))^2 + D(x_{11}\epsilon_1 + \cdots + x_{n1}\epsilon_n) = D(x_{11}\epsilon_1 + \cdots + x_{n1}\epsilon_n) \tag{8}$$

$$= D(x_{11}\epsilon_1) + \cdots + D(x_{n1}\epsilon_n) + Cov(x_{11}\epsilon_1, x_{21}\epsilon_2) + \cdots + Cov(x_{11}\epsilon_1, x_{n1}\epsilon_n) + Cov(x_{21}\epsilon_2, x_{31}\epsilon_3) \tag{9}$$

$$+ \cdots + Cov(x_{21}\epsilon_2, x_{n1}\epsilon_n) + \cdots\cdots + Cov(x_{(n-1)1}\epsilon_{n-1}, x_{n1}\epsilon_n) \tag{10}$$

$$= D(x_{11}\epsilon_1) + \cdots + D(x_{n1}\epsilon_n) = x_{11}^2 D(\epsilon_1) + \cdots + x_{n1}^2 D(\epsilon_n) = x_{11}^2 + \cdots + x_{n1}^2 \tag{11}$$

So in the same way:

$\mathbb{E}_\varepsilon[(x_{12}\epsilon_1 + \cdots + x_{n2}\epsilon_n)^2] = x_{12}^2 + \cdots + x_{n2}^2, \cdots, \mathbb{E}_\varepsilon[(x_{1r}\epsilon_1 + \cdots + x_{nr}\epsilon_n)^2] = x_{1r}^2 + \cdots + x_{nr}^2$

So: $\mathbb{E}_\varepsilon\|\boldsymbol{x}_\varepsilon\|^2 = [(x_{11}^2 + \cdots + x_{1r}^2) + (x_{21}^2 + \cdots + x_{2r}^2) + \cdots\cdots + (x_{n1}^2 + \cdots + x_{nr}^2)]/n^2$

$= (\|\boldsymbol{x}_1\|^2 + \|\boldsymbol{x}_2\|^2 + \cdots + \|\boldsymbol{x}_n\|^2)/n^2$

Since $\|\boldsymbol{x}_i\| \leq R$, so $\mathbb{E}_\varepsilon\|\boldsymbol{x}_\varepsilon\|^2 \leq nR^2/n^2 = R^2/n$, so it is proved.

(c) $\mathrm{Rad}(\mathcal{F}) = \mathbb{E}_\varepsilon\left(\max_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(\boldsymbol{x}_i)\right)$.

Using the conclusion in (a), we can get: $\mathrm{Rad}(\mathcal{F}) \leq W\mathbb{E}_\varepsilon\|\boldsymbol{x}_\varepsilon\|$

Since $\mathbb{E}_\varepsilon\|\boldsymbol{x}_\varepsilon\|^2 \leq R^2/n$, so $(\mathbb{E}_\varepsilon\|\boldsymbol{x}_\varepsilon\|)^2 \leq \mathbb{E}_\varepsilon\|\boldsymbol{x}_\varepsilon\|^2 \leq R^2/n$, so $\mathbb{E}_\varepsilon\|\boldsymbol{x}_\varepsilon\| \leq R/\sqrt{n}$

So $\mathrm{Rad}(\mathcal{F}) \leq RW/\sqrt{n}$

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, $\|\boldsymbol{x}\| \leq R$ and $y \in \{-1, +1\}$.

Consider the following version of the soft-margin SVM:

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

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

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

Hint: use the fact from slide 5/61 of the first ML Theory lecture that $\|\hat{\boldsymbol{w}}\| \leq \sqrt{2/\lambda}$, 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. Since $\|\boldsymbol{w}\| \leq \sqrt{2/\lambda}$, using the conclusion from the Problem 2, we can get:

$$\text{Rad}(\mathcal{F}) \leq R\sqrt{2/n\lambda} \tag{12}$$

2. There exists $\rho \geq 0$ so that for any $f, g \in \mathcal{F}$

$$|\ell(f(\boldsymbol{x}), y) - \ell(g(\boldsymbol{x}), y)| \leq \rho|f(\boldsymbol{x}) - g(\boldsymbol{x})| \tag{13}$$

Here, $\ell(f(\boldsymbol{x}), y) = \left[1 - \boldsymbol{w}_1^\top \boldsymbol{x} y\right]_+$, $\ell(g(\boldsymbol{x}), y) = \left[1 - \boldsymbol{w}_2^\top \boldsymbol{x} y\right]_+$.

(1) If $(1 - \boldsymbol{w}_1^\top \boldsymbol{x} y) > 0, (1 - \boldsymbol{w}_2^\top \boldsymbol{x} y) > 0$:

$$|\ell(f(\boldsymbol{x}), y) - \ell(g(\boldsymbol{x}), y)| = |\boldsymbol{w}_2^\top \boldsymbol{x} y - \boldsymbol{w}_1^\top \boldsymbol{x} y| = |\boldsymbol{w}_2^\top \boldsymbol{x} - \boldsymbol{w}_1^\top \boldsymbol{x}| = |f(\boldsymbol{x}) - g(\boldsymbol{x})| \tag{14}$$

(2) If $(1 - \boldsymbol{w}_1^\top \boldsymbol{x} y) \leq 0, (1 - \boldsymbol{w}_2^\top \boldsymbol{x} y) \leq 0$:

$$|\ell(f(\boldsymbol{x}), y) - \ell(g(\boldsymbol{x}), y)| = 0 \leq |f(\boldsymbol{x}) - g(\boldsymbol{x})| \tag{15}$$

(3) If $(1 - \boldsymbol{w}_1^\top \boldsymbol{x} y) \leq 0, (1 - \boldsymbol{w}_2^\top \boldsymbol{x} y) > 0$, $\boldsymbol{w}_1^\top \boldsymbol{x} y \geq 1$, $\boldsymbol{w}_2^\top \boldsymbol{x} y < 1$:

If $y = 1$: $\boldsymbol{w}_1^\top \boldsymbol{x} \geq 1$, $\boldsymbol{w}_2^\top \boldsymbol{x} < 1$

$$|\ell(f(\boldsymbol{x}), y) - \ell(g(\boldsymbol{x}), y)| = |1 - \boldsymbol{w}_2^\top \boldsymbol{x} y| = 1 - \boldsymbol{w}_2^\top \boldsymbol{x} \leq \boldsymbol{w}_1^\top \boldsymbol{x} - \boldsymbol{w}_2^\top \boldsymbol{x} = |f(\boldsymbol{x}) - g(\boldsymbol{x})| \tag{16}$$

If $y = -1$: $\boldsymbol{w}_1^\top \boldsymbol{x} \leq -1$, $\boldsymbol{w}_2^\top \boldsymbol{x} > -1$

$$|\ell(f(\boldsymbol{x}), y) - \ell(g(\boldsymbol{x}), y)| = |1 - \boldsymbol{w}_2^\top \boldsymbol{x} y| = 1 + \boldsymbol{w}_2^\top \boldsymbol{x} \leq \boldsymbol{w}_2^\top \boldsymbol{x} - \boldsymbol{w}_1^\top \boldsymbol{x} = |f(\boldsymbol{x}) - g(\boldsymbol{x})| \tag{17}$$

(4) If $(1 - \boldsymbol{w}_1^\top \boldsymbol{x} y) > 0, (1 - \boldsymbol{w}_2^\top \boldsymbol{x} y) \leq 0$, $\boldsymbol{w}_1^\top \boldsymbol{x} y < 1$, $\boldsymbol{w}_2^\top \boldsymbol{x} y \geq 1$:

If $y = 1$: $\boldsymbol{w}_1^\top \boldsymbol{x} < 1$, $\boldsymbol{w}_2^\top \boldsymbol{x} \geq 1$

$$|\ell(f(\boldsymbol{x}), y) - \ell(g(\boldsymbol{x}), y)| = |1 - \boldsymbol{w}_1^\top \boldsymbol{x} y| = 1 - \boldsymbol{w}_1^\top \boldsymbol{x} \leq \boldsymbol{w}_2^\top \boldsymbol{x} - \boldsymbol{w}_1^\top \boldsymbol{x} = |f(\boldsymbol{x}) - g(\boldsymbol{x})| \tag{18}$$

If $y = -1$: $\boldsymbol{w}_1^\top \boldsymbol{x} > -1$, $\boldsymbol{w}_2^\top \boldsymbol{x} \leq -1$

$$|\ell(f(\boldsymbol{x}), y) - \ell(g(\boldsymbol{x}), y)| = |1 - \boldsymbol{w}_1^\top \boldsymbol{x} y| = 1 + \boldsymbol{w}_1^\top \boldsymbol{x} \leq \boldsymbol{w}_1^\top \boldsymbol{x} - \boldsymbol{w}_2^\top \boldsymbol{x} = |f(\boldsymbol{x}) - g(\boldsymbol{x})| \tag{19}$$

So $\rho = 1$ here.

3. There exists $[a, b]$ so that $\ell(f(\boldsymbol{x}), y) \in [a, b]$ for any $f \in \mathcal{F}$.

Here $\ell(f(\boldsymbol{x}), y) = \left[1 - \boldsymbol{w}^\top \boldsymbol{x} y\right]_+ \geq 0$, so a=0 here.

In order to get the maximun of the $\ell(f(\boldsymbol{x}), y)$, $(1 - \boldsymbol{w}^\top \boldsymbol{x} y > 0)$, so now:

$$\ell(f(\boldsymbol{x}), y) = \left[1 - \boldsymbol{w}^\top \boldsymbol{x} y\right]_+ = 1 - \boldsymbol{w}^\top \boldsymbol{x} y = 1 - \|\boldsymbol{x}\|\|\boldsymbol{w}\| cos(\theta) y \leq 1 + \|\boldsymbol{x}\|\|\boldsymbol{w}\| = 1 + R\sqrt{2/\lambda} \tag{20}$$

So $b = \left(1 + R\sqrt{2/\lambda}\right)$ when $ycos(\theta) = -1, \|\boldsymbol{x}\| = R, |\boldsymbol{w}\| = \sqrt{2/\lambda}$.

So $b - a = \left(1 + R\sqrt{2/\lambda}\right)$

4. With probability $\geq 1 - \delta$, every $f \in \mathcal{F}$ satisfies:

$$\mathcal{R}_\ell(f) \leq \widehat{\mathcal{R}}_\ell(f) + 2\rho \operatorname{Rad}(\mathcal{F}) + 3(b - a)\sqrt{\frac{\ln(2/\delta)}{2n}} \tag{21}$$

Here $\rho = 1$, $b - a = \left(1 + R\sqrt{2/\lambda}\right)$ and $\operatorname{Rad}(\mathcal{F}) \leq R\sqrt{2/n\lambda}$.

$$So: \quad \mathcal{R}(\hat{f}) \leq \widehat{\mathcal{R}}(\hat{f}) + R\sqrt{\frac{8}{\lambda n}} + 3\left(1 + R\sqrt{2/\lambda}\right)\sqrt{\frac{\ln(2/\delta)}{2n}}. \tag{22}$$