CS 446 MJT — Homework 4

your NetID here

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 $\mathcal{F} := \left\{ \boldsymbol{x} \mapsto 2 \cdot \mathbb{1}[\boldsymbol{x} \in C] - 1 : C \subseteq \mathbb{R}^d \text{ is convex} \right\}$ 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.)*

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

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

   (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

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

   Hint: Cauchy-Schwarz!

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

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

Solution. *(Your solution here.)*

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.)*