ML Theory — Homework 1

your NetID here

Version 1

Instructions. (Different from homework 0.)

- Everyone must submit an individual write-up.
- You may discuss with up to 3 other people. State their NetIDs clearly on the first page. Outside of office hours, you should not discuss with anyone but these three.
- Homework is due **Wednesday, October 10, at 3:00pm**; no late homework accepted.
- Please consider using the provided LaTeX file as a template.

1. **(Miscellaneous short questions.)**

   Provide complete proofs, but try to find short solutions.

   (a) **(Justifying uniform norm: upper bound.)** Suppose ℓ is L-lipschitz, and μ is a probability measure supported on $[0,1]^d \times \{-1,+1\}$. Show

   $$\mathcal{R}_\ell(f) - \mathcal{R}_\ell(g) = \int \ell(-yf(x))\,\mathrm{d}\mu(x,y) - \int \ell(-yg(x))\,\mathrm{d}\mu(x,y) \leq L\|f-g\|_{\mathrm{u}}.$$

   (b) **(Justifying uniform norm: lower bound.)** Given any two continuous functions f and g, construct an L-lipschitz loss ℓ and a probability measure μ so that the previous part is tight: that is,

   $$\mathcal{R}_\ell(f) - \mathcal{R}_\ell(g) = L\|f-g\|_{\mathrm{u}}.$$

   Remark: together, we've shown why we aim for uniform approximation (it implies bounds for all measures).

   (c) **(Stone-Weierstrass with** cos**.)** Use the Stone-Weierstrass theorem, as stated in lecture 5 (do not use another source), to prove that for every continuous function $f : \mathbb{R}^d \to \mathbb{R}$ and $\epsilon > 0$, there exists $g \in \mathrm{span}(\mathcal{H}_{\cos})$ with $\|f-g\|_{\mathrm{u}} \leq \epsilon$. (**Hint:** refresh yourself on some trig identities.)

   (d) **(Deep, narrow networks.)** Let $\sigma_{\mathrm{r}}(z) := \max\{0,z\}$ denote the ReLU, and for convenience let $\vec{\sigma}_{\mathrm{r}}$ denote the coordinate-wise version of appropriate dimension (i.e., $\vec{\sigma}_{\mathrm{r}}(v)$ outputs a vector of the same dimension as v, whatever it happens to be).

   Suppose $f : [0,1]^d \to \mathbb{R}$ can be written as a network with a single ReLU layer, specifically $f(x) = A_2\vec{\sigma}_{\mathrm{r}}(A_1x + b_1)$ where $A_1 \in \mathbb{R}^{w\times d}$ and $A_2 \in \mathbb{R}^{1\times w}$. Construct a network with w ReLU layers and width $d+3$ which also (exactly) computes f.

   Remark: this reveals some convenient properties of ReLUs.

   (e) **(Uniform approximation with ReLU.)** Again define $\sigma_{\mathrm{r}}(z) := \max\{0,z\}$. Construct $\phi \in \mathrm{span}(\mathcal{H}_{\sigma_{\mathrm{r}}})$ which satisfies the conditions of Theorem 1.9 from Lecture 5 (and provide explicit verification of these conditions).

   Remark: consequently, Theorem 1.9 may be applied, and thus shallow ReLU networks fit continuous functions.

Solution.

(Your solution here.)

2. **(More headaches from Minsky-Papert.)**

Recall again the Minsky-Papert XOR problem, as appeared in Lecture 2. Let $\mathcal{S} := \{x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}\}$ denote the four points in that problem.

(a) Find a small axis-aligned decision tree which predicts perfectly on $\mathcal{S}$.

(b) Recall the class of *boosted decision stumps*: these are linear combinations of axis aligned decision trees with only one internal node. That is to say, a boosted decision stump is a function of the form

$$x \mapsto \sum_{i=1}^{N} \alpha_i \mathbb{1}[x_{j_i} \geq b_i]$$

where $(\alpha_1, \ldots, \alpha_N)$ are scalar, each $j_i \in \{1, \ldots, d\}$ indexes a single coordinate, and $(b_1, \ldots, b_N)$ are scalars.

Problem. Let $\epsilon > 0$ be given and construct (a) an axis-aligned decision tree g which predicts perfectly, (b) a boosted decision stump f which is incorrect on half of $\mathcal{S}$, but

$$\rho(f, g) := \int_{[-1,+1]^2} |f(x) - g(x)| \, \mathrm{d}x \leq \epsilon.$$

Note. Rather than the usual $\|f - g\|_1$ which integrates over $[0, 1]^d$, we are integrating over $[-1, +1]^2$.

(c) Prove that there can not exist a perfect boosted decision stump for the Minksy-Papert instance above.

Remark. First of all, this tells us some limitations of $\|\cdot\|_1$ approximation. Second of all, it tells us that boosted decision stumps, which were popular during roughly 1995-2005, are not so good.

Solution.

(Your solution here.)

3. **(Branching programs and decision trees.)**

Recall the discussion from the end of Lecture 5, regarding the size of $f(x) := \frac{1}{d}\sum_{i=1}^{d} x_i$ with $x \in \{0,1\}^d$ when represented as a decision tree and a branching program. A branching program of size $\mathcal{O}(d^2)$ was provided.

This question will prove that any decision tree needs size at least 2^d. In this question, the predicates computed by internal nodes are decision stumps, meaning they have the form $\mathbb{1}[x_i \geq b]$ where $i \in \{1, \ldots, d\}$ and $b \in \mathbb{R}$.

(a) As discussed in class, the leaves of the tree form a partition of the input space (in this case $\{0,1\}^d$). Each leaf can therefore be associate with a string s of length d, where $s_i \in \{\emptyset, -1, +1, \star\}$ means that inputs reaching this node respectively have nothing, -1, +1, or ± 1 in coordinate i.

Prove that given any leaf, its associated string has at least $d - p$ entries equal to $\star$, where p is the number of internal nodes (predicates) along the root-to-leaf path for this leaf.

(b) Use the preceding part to prove that any decision tree with strictly less than $2^d - 1$ internal nodes must fail to represent f (that is, it is incorrect on at least one input string $x \in \{0,1\}^d$).

Solution.

(Your solution here.)

4. **(2-layer networks fit continuous functions.)**

Recall from class the definition

$$\mathcal{H}_\sigma := \left\{ x \mapsto \sigma(\langle a, x\rangle + b) : a \in \mathbb{R}^d, b \in \mathbb{R} \right\}.$$

Using Stone-Weierstrass, we proved we can approximate continuous functions with $\mathrm{span}(\mathcal{H}_{\exp})$. It was then claimed that the rest of the proof is "essentially univariate"; this exercise completes that proof.

One more piece of notation is needed. Say that $\sigma : \mathbb{R} \to \mathbb{R}$ is *sigmoidal* when it is nondecreasing, continuous, and

$$\lim_{z \to -\infty} \sigma(z) = 0, \qquad \lim_{z \to \infty} \sigma(z) = 1.$$

(a) The first missing piece is to assert that we really are left with a univariate problem. Namely, prove the following.

> Let $\sigma : \mathbb{R} \to \mathbb{R}$ and $\phi : \mathbb{R} \to \mathbb{R}$ be given. Suppose that for any interval $[r, s]$ and any $\tau > 0$, we can always find $h \in \mathrm{span}(\mathcal{H}_\sigma)$ so that
>
> $$\sup\left\{ |h(x) - \phi(x)| : x \in [r, s]| \right\} \leq \tau.$$
>
> (In words: we have a way to approximate ϕ along $[r, s]$ with $\mathrm{span}(\mathcal{H}_\sigma)$.) Then for any $\epsilon > 0$ and any $g \in \mathcal{H}_\phi$ but now $g : \mathbb{R}^d \to \mathbb{R}$, we can still choose $f \in \mathrm{span}(\mathcal{H}_\sigma)$ with $\|f - g\|_{\mathrm{u}} \leq \epsilon$.

(b) Let sigmoidal $\sigma : \mathbb{R} \to \mathbb{R}$, target error $\tau > 0$, interval $[r, s]$, and a function $\psi : \mathbb{R} \to \mathbb{R}$ be given with ψ Lipschitz continuous along $[r, s]$. Show that there exists $h \in \mathrm{span}(\mathcal{H}_\sigma)$ satisfying

$$\sup\left\{ |h(x) - \psi(x)| : x \in [r, s]| \right\} \leq \tau.$$

> **Hints.** (a) Note that for large M, $\sigma(Mx) \approx \mathbb{1}[x \geq 0]$; (b) consider drawing a picture for the simpler case of nonincreasing ψ, with special attention to the meaning of Lipschitz continuity.

(c) Prove that exp and cos are Lipschitz continuous along any bounded interval. (Yup, that's really all you need to do for this part.)

Though you don't need to write anything about it here, I urge you to verify that the preceding steps can be combined with the material in lecture to complete the proof.

Solution.

(Your solution here.)

5. **(A nuisance.)**

Recall that the lectures on approximation of continuous functions by 2- and 3-layer networks did not include a nonlinearity on the final output. This exercise points out that we can use those as a lemma to establish that networks *with* final nonlinearities can also approximate continuous functions (albeit with restrictions on the range).

Throughout this exercise, suppose a function class $\mathcal{F}$ is given which fits continuous functions in our usual sense: for any continuous $g : \mathbb{R}^d \to \mathbb{R}$ and any $\tau > 0$, there exists $f \in \mathcal{F}$ with $\|f - g\|_{\mathrm{u}} \leq \tau$.

The following notation will be handy. Namely, given univariate $\sigma : \mathbb{R} \to [0, 1]$, define the function class

$$\mathcal{F}_\sigma := \{\sigma \circ f : f \in \mathcal{F}\}.$$

(a) Suppose $\sigma : \mathbb{R} \to [0, 1]$ is sigmoidal (as in the previous exercise), Lipschitz, and has a continuous inverse.

Show that for any continuous $g : \mathbb{R}^d \to (0, 1)$ and any $\epsilon > 0$, there exists $h \in \mathcal{F}_\sigma$ with $\|h - g\|_1 \leq \epsilon$.

Hint. Find a way to bake the inverse into the problem.

(b) **(Optional; *hard mode.*)** Suppose $\sigma : \mathbb{R} \to [0, 1]$ is sigmoidal.

Show that for any continuous $g : \mathbb{R}^d \to [0, 1]$ and any $\epsilon > 0$, there exists $h \in \mathcal{F}_\sigma$ with $\|h - g\|_{\mathrm{u}} \leq \epsilon$.

Note. If you do this part, you must *still* provide a complete independent solution to the previous part. Be nice to the TA...

Solution.

(Your solution here.)

6. **(Monomials and uniform approximation via derivatives.)**

This problem will provide an approach to uniform approximation that avoids Stone-Weierstrass; **do not** use Stone-Weierstrass or Weierstrass or anything similar in any step of the proof!

The problem will consider only the univariate case, but essentially the same proof works in the multivariate case (as discussed at the end).

For convenience, for any activation σ, define $\mathcal{G}_\sigma := \text{span}(\mathcal{H}_\sigma)$. Here are some useful analysis facts for this problem:

- Continuous functions are uniformly continuous on compact sets.
- To say a function f is C^∞ means all derivatives exist (and are continuous). If σ is C^∞, then so is every $f \in \mathcal{G}_\sigma$.

Throughout this problem, suppose σ is C^∞ and $\sigma^{(n)} \neq 0$, meaning the $n^{(\text{th})}$ derivative is not identically the zero function for every nonnegative integer n.

(a) (Closed under a single derivative.) Let $f \in \mathcal{G}_\sigma$ and any $w \in \mathbb{R}$ and any $\epsilon > 0$ be given, and define $h(x) := xf'(wx)$ (the mapping $x \mapsto \partial/\partial r f(rx)|_{r=w}$). Prove that there exists $g \in \mathcal{G}_\sigma$ so that $\|h - g\|_{\text{u}} \leq \epsilon$.

Hint. Consider the definition of $\partial/\partial r f(rx)|_{r=w}$ in terms of limits, and see how it interacts with an exact (integral remainder) Taylor expansion. Via the analysis facts above, you can conveniently bound the remainder term. Use this to construct an appropriate $g \in \mathcal{G}_\sigma$, and prove that it works.

(b) (Closed under derivatives.) For every real $w \in \mathbb{R}$ and positive integer n, define

$$h_{n,w}(x) := x^n \sigma^{(n)}(wx) = \partial^n/\partial r^n \sigma(rx)|_{r=w}.$$

Show that for any (w, ϵ, n), there exists $g \in \mathcal{G}_\sigma$ with $\|g - h_{n,w}\| \leq \epsilon$.

Hint. Combine the previous part with an induction on n and some careful reasoning about approximations. Be wary of circularity...

(c) (Monomials.) Prove that for any positive integer n and real $\epsilon > 0$, there exists $g \in \mathcal{G}_\sigma$ so that $\|g - p_n\|_{\text{u}} \leq \epsilon$ where $p_n(x) = x^n$.

Hint. Use the previous part, and double check the conditions on σ...

Now that we have monomials, we can use the Weierstrass Theorem (which has a simple constructive proof). Also, the proof above goes through no problem in the multivariate case (now use $x \mapsto \sigma(\langle w, x \rangle)$, and take different partial derivatives to get various monomials).

Solution.

(Your solution here.)

7. **(Why?)**

You receive full credit for this question so long as you write at least one sentence for each answer. Please be honest and feel free to be critical.

(a) Why are you taking this class? What do you expect from it?

(b) What do you expect to gain (e.g., in research, work, life) by knowing ML Theory?

(c) Do you have any feedback about the class, lectures, or instructor?

Solution.

(Your solution here.)