DiffPDF • hw1.pdf vs. hw1v2.pdf • 2018-10-03

ML Theory — Homework 1

your NetID here

Version 0

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.

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.

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 $\mathbf{1}[x_i \geq b]$ 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 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.)

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 $\mathbf{1}[x_i \geq b]$ 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.)

DiffPDF • hw1.pdf vs. hw1v2.pdf • 2018-10-03

6. **(New problem appearing before October 1.)**

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 $\|h - g\|_u \leq \epsilon$.
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 $\|g - p_n\|_u \leq \epsilon$ 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. **(New problem appearing before October 1.)**

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.)