Upper bounds on the Natarajan dimensions of some function classes

Ying Jin

https://ying531.github.io

Department of Statistics, Stanford University

IEEE International Symposium on Information Theory (ISIT), June 27, 2023

Natarajan dimension and multi-class learnability

- Empirical risk minimization for multi-class classification

$$\widehat{f} = \operatorname*{argmax}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i),$$

  where $Y_i \in \{1, \ldots, d\}$ is categorical, and $\ell(\cdot, \cdot)$ is some (classification) loss

- In learning theory, the performance/learnability of $\widehat{f}$ depends on the complexity of $\mathcal{F}$

Natarajan dimension and multi-class learnability

- Natarajan dimension is a complexity measure for multi-class classification

Definition (Natarajan dimension)

Let $\mathcal{H}$ be a class of functions $h \colon \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y} = \{1, \dots, d\}$, and let $S \subseteq \mathcal{X}$. We say that $\mathcal{H}$ N-shatters S if there exists $f_1, f_2 \colon S \to \mathcal{Y}$ such that $f_1(x) \neq f_2(x)$ for all $x \in S$, and for every $T \subseteq S$, there exists some $g \in \mathcal{H}$ such that

$$\forall x \in T,\ g(x) = f_1(x), \quad \text{and} \quad \forall x \in S \setminus T,\ g(x) = f_2(x).$$

The Natarajan dimension of $\mathcal{H}$, denoted as $d_N(\mathcal{H})$, is the maximal cardinality of any set that is N-shattered by $\mathcal{H}$.

- It generalizes the Vapnik-Chervonenkis (VC) dimension from binary to multi-class classification
- An equivalent notion is the graph dimension (see paper)

This work, and related ones

- This work: upper bounds on the Natarajan dimension of popular function classes
  - Decision trees and random forests
  - Neural networks with binary, linear, and ReLU activations
- Existing upper bounds on the Natarajan dimensions
  - Generalized linear models and reduction trees [Daniely et al., 2011]
  - Multi-class support vector machines [Guermeur, 2010]
  - One-versus-all, all-pairs, error-correcting-output-codes methods [Daniely et al., 2012]
- Proof techniques for neural nets generalize the techniques in [Sontag et al., 1998]

Upper bounding Natarajan dimension by growth functions

- The high-level idea of our bounds is by noting that for any function class $\mathcal{H}$,

$$2^{d_N(\mathcal{H})} \leq G(\mathcal{H}, d_N(\mathcal{H})),$$

where we define the growth function of $\mathcal{H}$ as

$$G(\mathcal{H}, n) := \max_{x_1,\ldots,x_n \in \mathcal{X}} \left| \left\{ \left(f(x_1), f(x_2), \ldots, f(x_n)\right) : f \in \mathcal{H} \right\} \right|$$

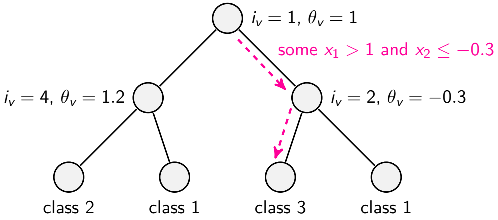
- That is, we get an upper bound $\mathcal{U}(\mathcal{H}, n)$ of $G(\mathcal{H}, n)$ in terms of n and function class parameters. And then solve the inequality (for n)

$$2^n \leq \mathcal{U}(\mathcal{H}, n)$$

to get an upper bound on $d_N(\mathcal{H})$

Decision trees and random forests

- Depth-L, d-class decision tree function class $\Pi_{L,d}^{\mathrm{dtree}}$
  - Each internal node v is associated with a feature $i_v \in \{1, \ldots, p\}$ and a threshold $\theta_v \in \mathbb{R}$
  - Each leaf node is associated with a class $k \in \{1, \ldots, d\}$
  - For input $x \in \mathbb{R}^p$, the output is obtained by traversing a path of length $L - 1$ from the root node to the leaf node. At each node, go to left child if $x_{i_v} \leq \theta_v$ and to right child otherwise



Decision trees and random forests

- Depth-L, d-class, T-tree random forests $\Pi^{\text{forest}}_{L,T,d}$
  - a classifier $F(\cdot)$ based on T depth-L d-class decision trees $f_j(\cdot)$, $j = 1, \ldots, T$
  - $F(x) = \text{argmax}_{1 \le k \le d} \sum_{j=1}^{T} \mathbf{1}\{f_j(x) = k\}$, the most-frequently predicted class among all T trees
- We derive an upper bound of the N-dim of $\Pi^{\text{forest}}_{L,T,d}$ based on an upper bound of $\Pi^{\text{dtree}}_{L,d}$

Decision trees and random forests

Theorem (Decision trees; J. 2023)

The Natarajan dimension of $\Pi_{L,d}^{dtree}$ *with inputs from* $\mathbb{R}^p$ *is no greater than* $\mathcal{O}(L2^L \log(pd))$.

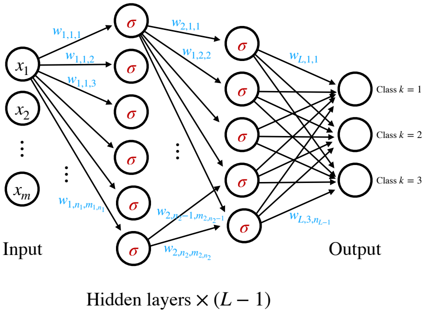
Theorem (Random forests; J. 2023)

The Natarajan dimension of $\Pi_{L,T,d}^{forest}$ *with inputs from* $\mathbb{R}^p$ *is no greater than* $\mathcal{O}(LT2^L \log(pd))$.

- Proof idea: bound # of distinct classifications over any $\{x_1, \ldots, x_n\}$ (the growth function)
- The growth function of $\Pi_{L,T,d}^{\mathrm{forest}}$ is bounded by that of $\Pi_{L,d}^{\mathrm{dtree}}$ to the power of T
- Agree with the VC-dimension upper bound in [Leboeuf et al. 2022] (a very recent result that appeared later than the arXiv version of this paper)

Multi-class neural networks with binary & linear activation

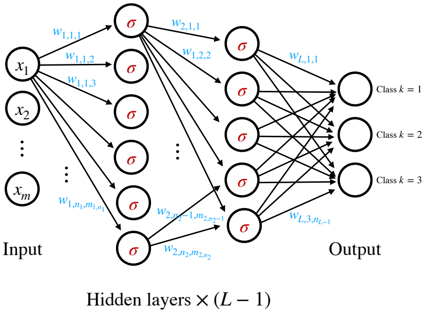
- Neural network function class $\Pi_{p,S}^{\text{bin-lin}}$ with a fixed structure S of p parameters



- There are p parameters in totol: $\{w_{\ell,j,s}\}_{1\leq\ell\leq L, 1\leq j\leq n_\ell, 1\leq s\leq m_{\ell,j}}$
- Input layer has one node for each feature
- Output for node j in hidden layer ℓ is $f_j^{(\ell)}(x) = \sigma(\sum_{s=1}^{m_{\ell,j}} w_{\ell,j,s} f_s^{(\ell-1)}(x))$, where $m_{\ell,j}$ is the number of nodes in layer $\ell - 1$ that are connected to node j in layer ℓ
- $\sigma(\cdot)$ is the activation function; in this class, either $\sigma(z) = z$ or $\sigma(z) = \mathbf{1}\{z > 0\}$
- Each class has a final output (fully connected), and the classification is given by the maximum output: $f(x; w) = \mathrm{argmax}_{1\leq k\leq d} \sum_{s=1}^{n_{L-1}} w_{L,k,s} f_s^{(L-1)}(x)$

Multi-class neural networks with binary & linear & ReLU activation

- Neural network function class $\Pi_{p,S}^{\mathrm{ReLU}}$ with a fixed structure S of p parameters



- Structure notations the same as before
- Allow for ReLU activation function; in this class, either $\sigma(z) = z$ or $\sigma(z) = \mathbf{1}\{z > 0\}$ or $\sigma(z) = z\mathbf{1}\{z > 0\}$

Multi-class neural networks

Theorem (J. 2023)

The Natarajan dimensions of $\Pi_{p,S}^{bin\text{-}lin}$ and $\Pi_{p,S}^{ReLU}$ are both upper bounded by $\mathcal{O}(d \cdot p^2)$, where d is the number of classes, and p is the number of parameters.

- Textbook result [Shalev-Shwartz and Ben-David, 2014] shows neural nets with p parameters and *only binary* activation has VC dimension $\mathcal{O}(p \log p)$, while Sontag et al. [1998] shows neural nets with p parameters and binary & linear activations has VC dimension $\mathcal{O}(p^2)$
- Results on VC dimensions suggest linear activation incurs a factor of p
- Our bound adds a factor of d for d-class classification, and agrees with Sontag et al. [1998] when reduced to binary classification
    - Our proof idea generalizes Sontag et al. [1998], which depends on an equivalent description of all possible distinct functions that can be expressed by functions in the class

Thank you!

Feel free to check arXiv: 2209.07015

Questions? reach me at ying531[at]stanford[dot]edu

My website https://ying531.github.io