2023-2024

CRITICAL ANALYSIS OF A RESEARCH ARTICLE

Alex Pierron & Matisse Roche

 https://github.com/AlexPierron/Theoretical-Principal-of-Deep-Learning

Theoretical Principles of Deep Learning

M2 Mathematics and Artificial Intelligence

Paris-Saclay University/ CentraleSupélec / Mathematical Institute of Orsay

February 16, 2024

Table of Contents

1   Introduction and Context

Supervised learning consists of developing predictive models from labeled examples. This approach involves the use of a training set to learn a prediction function. The ultimate goal is to generalize effectively to new data, i.e. to make accurate predictions on data not present in the training set. In order to obtain guarantees on this generalization capacity, we want to establish theoretical bounds. These bounds aim to ensure, with high probability, that the accuracy rate on test data is well estimated based on the accuracy rate on training data.

Historically, theoretical generalization bounds relied mainly on the VC (Vapnik-Chervonenkis) dimension of the functions used for prediction. This dimension poses practical challenges as it can be sensitive to small perturbations. The article we have chosen to study: "The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network" [1] uses the FS (fat-shattering) dimension as an alternative indicator. This new approach shows that under certain conditions, generalization in a neural network depends less on the number of parameters and more on the weight of those parameters.

2   Formalism and Mathematical Results

2.1   Definitions

We will use the following notions:

- **Threshold function:** $\mathrm{sgn}(\alpha) = \begin{cases} -1 & \text{if } \alpha < 0 \\ 1 & \text{if } \alpha \geq 0 \end{cases}$

- **Probability of misclassification:** $\mathrm{er}_P(h) = P\{\mathrm{sgn}(h(x)) \neq y\}$

- **Training data:** $z = \{(x_i, y_i) \mid (x_i, y_i) \in X \times \{-1, 1\}, \ 1 \leq i \leq m\}$

- **Error estimation:**
$$\hat{\mathrm{er}}_z^\gamma(h) = \frac{1}{m} \left|\{i : y_i h(x_i) < \gamma\}\right|$$

- **FS Dimension**

  Let H be a class of real functions defined on X.
  For $\gamma > 0$, a sequence $(x_1, \ldots, x_m) \in X^m$ is said to be γ-**shattered by** H if:

  $$\exists \, s = (s_1, \ldots, s_m) \in \mathbb{R}^m, \forall \, b = (b_1, \ldots, b_m) \in \{-1, 1\}^m, \exists \, h \in H \ textsuchthat \, (h(x_i) - s_i) \, b_i \geq \gamma$$

  The **fat-shattering dimension** of H is defined as:

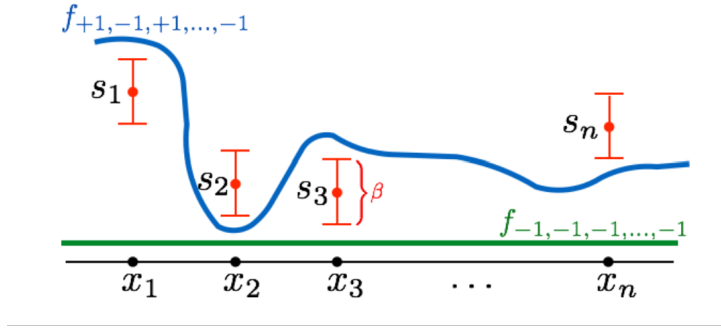 $$\mathrm{fat}_H(\gamma) = \max\{m : H \ \gamma\text{-shatters some } x \in X^m\}.$$

Figure 1: Illustration of fat-shattering

- Let (S, ρ) be a pseudometric space. For $A \subseteq S$, a set $T \subseteq S$ is an ϵ-covering of A with respect to ρ if for every a in A, there exists a t in T such that $\rho(t, a) < \epsilon$. We define $\mathcal{N}(A, \epsilon, \rho)$ as the size of the smallest ϵ-covering of A.

  For a class F of functions defined on a set X and a sequence $x = (x_1, \ldots, x_m) \in X^m$, we define the pseudometric $d_{\ell_\infty(x)}$ by

  $$d_{\ell_\infty(x)}(f, g) = \max_i |f(x_i) - g(x_i)|.$$

  We denote $\mathcal{N}_\infty(A, \epsilon, m) = \max_{x \in X^m} \mathcal{N}\left(A, \epsilon, d_{\ell_\infty(x)}\right).$

- We denote $\mathcal{M}_\infty(F, \alpha, m)$ the maximum over all $x \in X^m$ of the size of the largest subset of F for which all pairs of elements are α-separated with respect to $d_{\ell_\infty(x)}$.

2.2 Main Results

The two main results of the article are Theorems 2 and 28. We use Lemma 4 and Theorem 5 in the proof of Theorem 2.

Theorem 2 Suppose $z = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ is chosen by m independent draws according to P. Then, with probability at least $1 - \delta$, for each h in H, we have

$$\mathrm{er}_P(h) < \hat{\mathrm{er}}_z^\gamma(h) + \sqrt{\frac{2}{m}\left(d\ln(34em/d)\log_2(578) + \ln(4/\delta)\right)}, \qquad \text{(Theorem 2)}$$

Lemma 4 Suppose $\gamma > 0$, $0 < \delta < 1/2$, P is a probability distribution over $X \times \{-1, 1\}$, and $z = ((x_1, y_1), \ldots, (x_m, y_m))$ is chosen by m draws independently according to P. Then with probability at least $1 - \delta$, each h in H has

$$\mathrm{er}_P(h) < \hat{\mathrm{er}}_z^\gamma(h) + \sqrt{\frac{2}{m}\ln\left(\frac{2\mathcal{N}_\infty\left(\pi_\gamma(H), \gamma/2, 2m\right)}{\delta}\right)}$$

Theorem 5

Let F be a class of functions from $\{1, \ldots, n\}$ to $\{1, \ldots, b\}$ with $\mathrm{fat}_F(1) \leq d$. Then

$$\log_2 \mathcal{N}_\infty(F, 2, n) < 1 + \log_2\left(nb^2\right)\log_2\left(\sum_{i=0}^d \binom{n}{i} b^i\right),$$

4

provided that

$$n \geq 1 + \log_2 \left(\sum_{i=0}^{d} \binom{n}{i} b^i \right).$$

__Theorem 28__ (Part 1) Let $0 < \gamma \leq 1$ and $0 < \delta < \dfrac{1}{2}$. Let $\sigma : \mathbb{R} \to [-1, 1]$ be a non-decreasing function. Let the class F of functions on $\mathbb{R}^n$ be given by $F = \{x \mapsto \sigma(w \cdot x + w_0) : w \in \mathbb{R}^n, w_0 \in \mathbb{R}\}$.

$$\text{Let } H = \left\{ \sum_{i=1}^{N} \alpha_i f_i : N \in \mathbb{N}, f_i \in F, \sum_{i=1}^{N} |\alpha_i| \leq A \right\}, \text{for } A \geq 1.$$

Then, with probability at least $1 - \delta$, for each h in H we have

$$\mathrm{er}_P(h) < \mathrm{\hat{e}r}_z^{\gamma}(h) + \sqrt{\frac{64 \log_2(6)}{m} \left(\frac{A^2 n}{\gamma^2} \log\left(\frac{A}{\gamma}\right) \log^2 m + \log\left(\frac{1}{\delta}\right) \right)} \qquad \text{(Theorem 28)}$$

3   Proof of Theorem 2

Let $Q_\alpha : \mathbb{R} \to \mathbb{R}$ be the quantization function defined by :

$$Q_\alpha(x) = x \left\lceil \frac{x - \frac{\alpha}{2}}{\alpha} \right\rceil$$

α. $Q_\alpha(x)$ is the nearest multiple of α to x. Thus, for any $x \in \mathbb{R}$, we have $|Q_\alpha(x) - x| \leqslant \dfrac{\alpha}{2}$.

Let $F = Q_{\frac{\gamma}{8}}(\pi_\gamma(H))$.

We want to show that $\mathrm{fat}_F(\gamma/8) \leq \mathrm{fat}_{\pi_\gamma(H)}(\gamma/16)$.

To do this, let $x = (x_1, \ldots, x_n) \in X^n$ which is $\dfrac{\gamma}{8}$-shattered by F. We will show that x is $\dfrac{\gamma}{16}$-shattered by $\pi_\gamma(H)$.

$$\exists (r_1, \ldots, r_n) \in X^n, \forall b = (b_1, \ldots, b_n) \in \{-1; 1\}^n, \exists f \in \pi_\gamma(H),$$

$$\left(Q_{\frac{\gamma}{8}}(f(x_i)) - r_i \right) b_i \geqslant \frac{\gamma}{8}$$

We have :

$$\left(Q_{\frac{\gamma}{8}}(f(x_i)) - r_i \right) b_i$$

$$= \left(Q_{\frac{\gamma}{8}}(f(x_i)) - f(x_i) \right) b_i + (f(x_i) - r_i) b_i$$

Now, $\left| Q_{\frac{\gamma}{8}}(f(x_i)) - f(x_i) \right| \leqslant \dfrac{\gamma}{16}$. Thus,

$$(f(x_i) - r_i) b_i \geqslant \left(Q_{\frac{\gamma}{8}}(f(x_i)) - r_i \right) b_i - \frac{\gamma}{16} \geqslant \frac{\gamma}{8} - \frac{\gamma}{16} = \frac{\gamma}{16}.$$

So x is indeed $\dfrac{\gamma}{16}$-shattered by $\pi_\gamma(H)$. Hence $\mathrm{fat}_F(\gamma/8) \leq \mathrm{fat}_{\pi_\gamma(H)}(\gamma/16)$.

On one hand, we have $\mathcal{M}_\infty(\pi_\gamma(H), \gamma/2, 2m) \leq \mathcal{M}_\infty(F, \gamma/2, 2m)$, and $\mathcal{N}_\infty(\pi_\gamma(H), \gamma/2, 2m) \leq \mathcal{M}_\infty(\pi_\gamma(H), \gamma/2, 2m)$. Finally, $\mathcal{M}_\infty(F, \gamma/2, 2m) \leq \mathcal{N}_\infty(F, \gamma/4, 2m)$ gives us

$$\mathcal{N}_\infty(\pi_\gamma(H), \gamma/2, 2m) \leq \mathcal{N}_\infty(F, \gamma/4, 2m).$$

By applying Theorem 5 with $n = 2m$ and $b = 17$, we obtain :

$$\log_2 \mathcal{N}_\infty(\pi_\gamma(H), \gamma/2, 2m) < 1 + d \log_2(34em/d) \log_2(578m)$$

Substituting this inequality into Lemma 4 and using $\mathrm{fat}_{\pi_\gamma(H)}(\gamma/16) \leq \mathrm{fat}_H(\gamma/16)$, we obtain the desired result.

4 Numerical Simulation and Experimentation

4.1 Principle of the Experiment

We conduct experiments with neural networks having a single hidden layer, following the conditions stated in Theorem 28. These models, single-hidden-layer MLPs with size h_{size}, are initialized with weights following a centered normal distribution, where the standard deviation (**std**) is chosen to obtain weights of varying magnitudes. By fixing the same hyperparameters such as the learning rate and the number of epochs, we utilize the stochastic gradient descent optimizer (`torch.optim.SGD`).

We measure the accuracy on the training and test sets, recorded respectively as `train_acc` and `test_acc`. Additionally, we compute the bound proposed by the article for $\gamma = 0.5$ and $\delta = 0.25$.

These experiments are conducted on the MNIST dataset, distinguishing even and odd digits, as well as on internally generated synthetic data. We carry out these experiments for single cases and for a unified procedure. The distinction lies in the automation of execution as well as the returned metrics. The code is available at this link github.

4.2 Results

In this section, we present the results obtained for our experiments, first conducted on single cases. The results obtained for the unified procedure are available on our github.

The following table compiles the results obtained for our approach on MNIST and with our synthetic data for single cases.

The table below summarizes the results of our experiments on MNIST and synthetic data for various network size configurations. The "bound" column corresponds to the theoretical bound obtained according to Theorem 28. However, we observe that these bounds, being greater than 1, do not provide useful indications regarding the probability of error on the test data.

Weight initialization in neural networks is crucial, with models initialized with higher weights often showing poorer final performances at equal size. For example, models A and B for MNIST, and B and C for synthetic data. Sometimes, smaller models outperform larger ones due to differences in initial weight values, like model A compared to model E on MNIST.

However, there are inconsistencies between observed performances and theoretical bounds. For instance, although the bound of A is higher than that of B, experimental results do not always follow this trend. These observations can be partly explained by the use of stochastic gradient descent and the dependence on the learning rate, which introduce random components difficult to control.

	Model	$h_{\textbf{size}}$	std	train_acc	test_acc	bound
MNIST (150 epochs)	A	50	0.00001	98.15	97.64	278.13
	B	50	100	77.86	78.27	112.32
	C	200	0.00001	98.05	97.52	357.42
	D	200	100	69.19	77.00	7701.10
	E	200	1	93.04	92.50	162.13
	F	50	1	91.46	91.31	54.69
Synthetic Data (65 epochs)	A	50	0.01	99.53	99.45	10.76
	B	200	0.1	95.64	95.62	3.59
	C	200	0.01	99.79	99.70	23.92

Table 1: Results on MNIST and synthetic data for various network size configurations. The "bound" column corresponds to the complete bound obtainable by Theorem 28.

Regarding the unified procedure on synthetic data, the results are similar to those of the previous table, but with bounds closer to 1. This suggests potential utility of these bounds in the case of massive data and low dimensionality. However, the strong causality break between the value of the final bound (and thus the final weights) and performance persists, due to the numerous random steps and the complex influence of the learning rate and gradient computations on the weights, especially for smaller models.

4.3   Conclusion and Outlook

In conclusion, the theoretical bounds presented in this article are interesting as they highlight the importance of network weights rather than the number of neurons concerning generalization ability. Additionally, they provide a result of asymptotic convergence. However, since the article dates back to 1997, it is not accompanied by illustrative numerical simulations. Our own simulations have revealed that it was not possible to obtain a usable bound (with a second member less than 1) without having enormous amounts of training data.

A   References

[1] Peter Bartlett , *The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network*, IEEE Transactions on Information Theory, 44(2), pp. 525-536, (1998).