MythBusters: A Deep Learning Edition

Sasha Rakhlin
MIT

Jan 18-19, 2018

Outline

Myth #1: Current theory is lacking because deep neural networks have too many parameters.

Myth #1: Current theory is lacking because deep neural networks have too many parameters.

P. Bartlett, "The Sample Complexity of Pattern Classification with Neural Networks: The Size of the Weights is More Important than the Size of the Network," 1998

Myth #1: Current theory is lacking because deep neural networks have too many parameters.

P. Bartlett, "The Sample Complexity of Pattern Classification with Neural Networks: The Size of the Weights is More Important than the Size of the Network," 1998

- Margin theory was developed to address this very problem for Boosting and NN (e.g. Koltchinskii & Panchenko '02 and references therein)
- Example: linear classifiers $\{x \mapsto \text{sign}(\langle w, x \rangle) : \|w\|_2 \leq 1\}$ and assume margin. Then dimension of w (num. of params in 1-layer NN) never appears in generalization bounds (and can be infinite). This observation already appears in the 60's.

Myth #1: Current theory is lacking because deep neural networks have too many parameters.

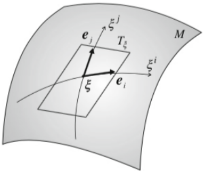
P. Bartlett, "The Sample Complexity of Pattern Classification with Neural Networks: The Size of the Weights is More Important than the Size of the Network," 1998

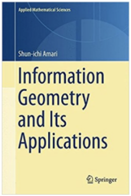
- Margin theory was developed to address this very problem for Boosting and NN (e.g. Koltchinskii & Panchenko '02 and references therein)
- Example: linear classifiers $\{x \mapsto \text{sign}(\langle w, x\rangle) : \|w\|_2 \leq 1\}$ and assume margin. Then dimension of w (num. of params in 1-layer NN) never appears in generalization bounds (and can be infinite). This observation already appears in the 60's.
- In Statistics, one often deals with infinite-dimensional models
- Numer of parameters is rarely the right notion of complexity (true, in classical statistics still the case for linear regression or simple models)
- VC dimension is known to be a loose quantity (distribution-free, only an upper bound)

A study of complexity notions

Our own (arguably incomplete) take on this problem:

T. Liang, T. Poggio, J. Stokes, A.R. “Fisher-Rao Metric, Geometry, and Complexity of Neural Networks,” 2017.

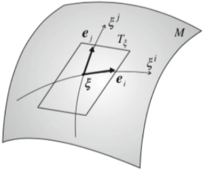


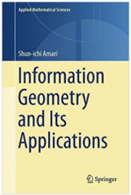


A study of complexity notions

Our own (arguably incomplete) take on this problem:

T. Liang, T. Poggio, J. Stokes, A.R. "Fisher-Rao Metric, Geometry, and Complexity of Neural Networks," 2017.





- Fisher local norm as a common starting point for many measures of complexity currently studied in the literature (see work of Srebro's group and Bartlett et al).
- Information Geometry suggests *Natural Gradient* as the optimization method. Appears to resolve ill-conditioned problems in Shalev-Shwartz et al '17.

Myth #2: To prove good out-of-sample performance, we need to show uniform convergence (a la Vapnik) over some class.

Myth #2: To prove good out-of-sample performance, we need to show uniform convergence (a la Vapnik) over some class.

The oldest counter-example:

Cover and Hart, "Nearest neighbor pattern classification," 1967.

Myth #2: To prove good out-of-sample performance, we need to show uniform convergence (a la Vapnik) over some class.

The oldest counter-example:

Cover and Hart, "Nearest neighbor pattern classification," 1967.

Second (related) issue: uniform vs universal consistency.

Uniform Consistency
There exists a sequence $\{\widehat{y}_t\}_{t=1}^{\infty}$ of estimators, such that for any $\epsilon > 0$, there exists n_ϵ such that for any distribution $P \in \mathcal{P}$ and $n \geq n_\epsilon$,

$$\mathbb{E}\mathbf{L}(\widehat{y}_n) - \inf \mathbf{L}(f) \leq \epsilon$$

Universal Consistency
There exists a sequence $\{\widehat{y}_t\}_{t=1}^{\infty}$ of estimators, such that for any distribution $P \in \mathcal{P}$ and any $\epsilon > 0$, there exists n_ϵ such that for $n \geq n_\epsilon(P)$,

$$\mathbb{E}\mathbf{L}(\widehat{y}_n) - \inf \mathbf{L}(f) \leq \epsilon$$

Myth #2: To prove good out-of-sample performance, we need to show uniform convergence (a la Vapnik) over some class.

The oldest counter-example:

Cover and Hart, "Nearest neighbor pattern classification," 1967.

Second (related) issue: uniform vs universal consistency.

Uniform Consistency
There exists a sequence $\{\widehat{y}_t\}_{t=1}^{\infty}$ of estimators, such that for any $\epsilon > 0$, there exists n_ϵ such that for any distribution $P \in \mathcal{P}$ and $n \geq n_\epsilon$,

$$\mathbb{E}\mathbf{L}(\widehat{y}_n) - \inf \mathbf{L}(f) \leq \epsilon$$

Universal Consistency
There exists a sequence $\{\widehat{y}_t\}_{t=1}^{\infty}$ of estimators, such that for any distribution $P \in \mathcal{P}$ and any $\epsilon > 0$, there exists n_ϵ such that for $n \geq n_\epsilon(P)$,

$$\mathbb{E}\mathbf{L}(\widehat{y}_n) - \inf \mathbf{L}(f) \leq \epsilon$$

Importantly, can interpolate between the two notions using penalization. A few more approaches (e.g. use bracketing entropy) – ask me after the talk.

Myth #3: Sample complexity of neural nets scales exponentially with depth.

Myth #3: Sample complexity of neural nets scales exponentially with depth.

- A common pitfall of making conclusions based on (possibly loose) upper bounds.

Mostly resolved:

N. Golowich, A.R., O. Shamir, "Size-Independent Sample Complexity of Neural Networks," 2017

From 2^d to $\sqrt{d}$ dependence was simply a technical issue. From $\sqrt{d}$ to $O(1)$ requires more work.

Myth #4: If we can fit any set of labels, then Rademacher complexity is too large and, hence, nothing useful can be concluded.

Myth #4: If we can fit any set of labels, then Rademacher complexity is too large and, hence, nothing useful can be concluded.

Related to Myth #2, but let's illustrate with a slightly different technique. Bottom line: we can have a very large overall model, but performance depends on *a posteriori* complexity of the *obtained* solution.

Most trivial example: take a large $\mathcal{F} = \cup_k \mathcal{F}_k$, where $\mathcal{F}_k = \{f : \mathrm{compl}_n(f) \le k\}$ and for simplicity assume $\mathrm{compl}_n(f)$ is positive homogenous. Suppose (this is standard) we have that with high probability

$$\forall f \in \mathcal{F}_1, \qquad \mathbb{E}f - \widehat{\mathbb{E}}f \lesssim \widehat{\mathscr{R}}(\mathcal{F}_1) + \dots$$

where $\widehat{\mathscr{R}}(\mathcal{F}_1)$ is empirical Rademacher. Then with same probability

$$\forall f \in \mathcal{F}, \qquad \mathbb{E}f - \widehat{\mathbb{E}}f \lesssim \mathrm{compl}_n(f) \cdot \widehat{\mathscr{R}}(\mathcal{F}_1) + \dots$$

Conclusion: an *a posteriori* data-dependent guarantee for all f based on complexity of f, yet $\widehat{\mathscr{R}}(\mathcal{F})$ never appears (huge or infinite). If complexity is not positive homogenous, use union bound instead.

So, is there anything left to do? Yes, tons. Perhaps need to ask different questions.

- What are the properties of solutions that optimization methods find in a nonconvex landscape? Is there “implicit regularization” that we can isolate?

  A nice line of work by Srebro and co-authors

- What are the salient features of the random landscape? Uniform deviations for gradients and Hessians?

  Nice work by Montanari and co-authors

- How can one exploit randomness to make conclusions about optimization solutions? (e.g. see the SGLD work of Raginsky et al, as well as papers on escaping saddles)
- What geometric notions can be associated to multi-layer neural nets? How can this geometry be exploited in optimization methods and be reflected in sample complexity?
- Theoretical understanding of adversarial examples.
- etc.