Foundations of Deep Learning

Lecture 07

Generalization I

Aurelien Lucchi

Fall 2024

Problem of interest

Hypothesis space Set $\mathcal{F}$ of functions from $\mathcal{X} \subseteq \mathbb{R}^d$ to $\mathcal{Y} \subseteq \mathbb{R}$

Risk Given a data distribution $\mathcal{D}$ over $(\mathcal{X}, \mathcal{Y})$, we define the true risk as

$$R(f) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[l((\mathbf{x}, y), f)]$$

We will mostly discuss a classification setting where $l((\mathbf{x}, y), f) = \mathbf{1}_{f(\mathbf{x})=y}$.

Empirical risk Typically do not have access to the exact data distribution but only a sample set $S = \{(\mathbf{x}_i, y_i\}_{i=1}^{n} \sim D$
$\rightsquigarrow$ define the empirical risk as

$$R_S(f) = \sum_{(x,y)\in S} [l((\mathbf{x}, y), f)].$$

What's a generalization bound?

The typical form of a generalization bound we are interested in is:

$$\underbrace{R(f)}_{\text{True risk}} \leq \underbrace{R_S(f)}_{\text{Empirical risk}} + \underbrace{m(\text{complexity of class of functions}, n)}_{\text{(a function that approaches 0 as n approaches infinity)}} ,$$

where n is the number of training datapoints and m is a function that measures the complexity of a class of functions.

Section 1

Basic Concentration Bound

Applying Hoeffding's inequality

Let $Z_i = (\mathbf{x}_i, y_i)$ be i.i.d. random variables with $f(Z) \in [a, b]$. The quantity we are interested in is

$$R(f) - R_S(f) = \mathbb{E}[f(Z)] - \frac{1}{n}\sum_{i=1}^{n} f(Z_i).$$

Using Hoeffding's inequality, as well as the independence of the random variables $f(Z_i)$, we get:

$$P\left[\left|\mathbb{E}[f(Z)] - \frac{1}{n}\sum_{i=1}^{n} f(Z_i)\right| > \epsilon\right] \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right).$$

Setting the right hand side to be δ, we obtain:

$$\delta = 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right) \implies \epsilon = (b-a)\sqrt{\frac{\log\frac{2}{\delta}}{2n}}$$

Union bound

Therefore, with probability at least $1-\delta$, for a given $f \in \mathcal{F}$,

$$|R_S(f) - R(f)| \leq (b-a)\sqrt{\frac{\log \frac{2}{\delta}}{2n}}.$$

$\longrightarrow$ Result is for **one** given element $f \in \mathcal{F}$.

Generalization to all functions in $\mathcal{F}$

Use union bound:

$$P(C_1 \cup \cdots \cup C_N) \leq \sum_{n=1}^{N} P(C_n) \leq N\delta.$$

Union bound

For simplicity, we take $(b-a)=1$.

We get with probability at least $1-\delta$,

$$\forall f \in \mathcal{F}, R(f_S) \leq R(f^*) + 2\sqrt{\frac{\log N + \log \frac{1}{\delta}}{2n}}.$$

Proof...

Uncountable case

What if the set $\mathcal{F}$ is uncountable?
$\longrightarrow$ Previous approach does not directly work.

There are several **measures of capacity** or size of classes of function:

- Distribution independent: e.g. VC dimension and growth function
- Distribution dependent: VC entropy

Section 2

Uniform Bounds

Principle behind Uniform bounds

Problem:

- Previous results: for each (fixed) function $f \in \mathcal{F}$, there is a collection of sample sets S for which the bound holds
- However these sets S may be different for different functions f
- In other words, for the observed sample, only some of the functions in $\mathcal{F}$ will satisfy the bound.
- This means that f cannot change with different draws of S for the bound to hold.

Solution: consider uniform bounds, i.e. bounds that hold uniformly over all functions in the class:

$$\sup_{f \in \mathcal{F}} R(f) - R_S(f).$$

Generalization bound

Theorem 1

Given a set S of size n and $\mathcal{F} = \{f_1, \dots f_N\}$, let $f^* = \operatorname{argmin}_{f \in \mathcal{F}} R(f)$ and $f_S = \operatorname{argmin}_{f \in \mathcal{F}} R_S(f)$. Then

$$R(f_S) \leq R(f^*) + 2\sqrt{\frac{\log N + \log \frac{2}{\delta}}{2n}}$$

Proof

Section 3

Infinite Case: VC Bound

VC theory

Can we extend our result to the case where the class $\mathcal{F}$ is infinite?

VC Vapnik & Chervonenkis argued that what really matters for a sample $S = \{\mathbf{z}_1, \ldots \mathbf{z}_n\}$ is the set

$$\mathcal{F}_{\mathbf{z}_1, \ldots \mathbf{z}_n} = \{(f(\mathbf{z}_1), \ldots f(\mathbf{z}_n)) | f \in \mathcal{F}\}.$$

$\leadsto$ Size of this set is the total number of possible ways that S can be classified.

VC theory - Example

Example: binary classification $\mathcal{F}_{\mathbf{z}_1,\ldots\mathbf{z}_n}$ is a set of binary vectors $\subset \{-1,+1\}^n$.

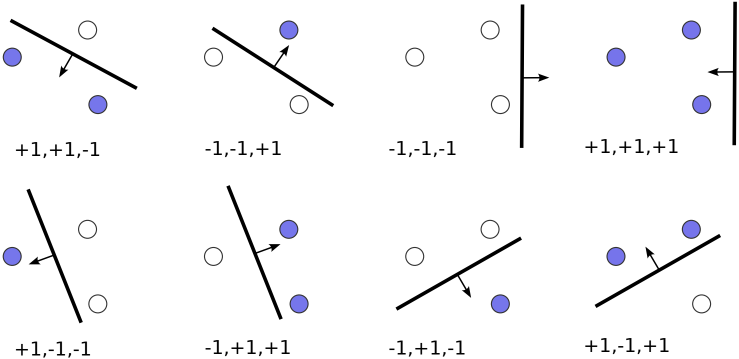

Figure: Illustration of growth function for $n = 3$ datapoints.

VC theory

Definition 2 (Growth Function)

The growth function of $\mathcal{F}$ is equal to

$$S_{\mathcal{F}}(n) = \sup_{(\mathbf{z}_1 \ldots \mathbf{z}_n)} | \mathcal{F}_{\mathbf{z}_1 \ldots \mathbf{z}_n} |$$

⇝ note that we can choose the configuration of the points!

Interpretation Size of this set is the number of possible ways in which the data $\{\mathbf{z}_1 \ldots \mathbf{z}_n\}$ can be classified using functions in $\mathcal{F}$. An upper bound on this number is 2^n.

VC theorem

Theorem 3

For any $\delta > 0$, with probability at least $1 - \delta$,

$$\forall f \in \mathcal{F}, \quad R(f) \leq R_S(f) + 2\sqrt{2\frac{\log S_{\mathcal{F}}(2n) + \log \frac{2}{\delta}}{n}},$$

where $S_{\mathcal{F}}(2n)$ is the growth function of the function class $\mathcal{F}$ defined as

$$S_{\mathcal{F}}(n) = \sup_{z} |\mathcal{F}_{z_1,\ldots z_n}|.$$

Proof.

Proof omitted. See statistical theory book. □

VC dimension

Shattering We say that $\mathcal{F}$ shatters S if $|\mathcal{F}_S| = 2^{|S|}$

VC dimension The VC dimension of a function class $\mathcal{F}$ is the cardinality of the largest set that it can shatter.

Definition 4 (VC dimension)

The VC dimension of a class $\mathcal{F}$ is the largest n such that

$$S_{\mathcal{F}}(n) = 2^n.$$

VC dimension - example

Example: VC dimension of linear classifiers in $\mathbb{R}^d$ is $d + 1$.

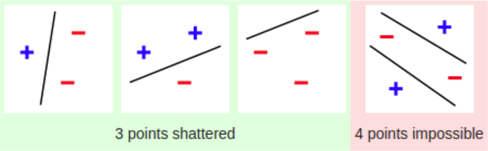

Figure: Illustration of growth function. Source: Wikipedia.

Sauer-Shelah lemma

Lemma 5 (Sauer-Shelah lemma)

Let $\mathcal{F}$ be a class of functions with finite VC-dimension h. Then, $\forall n \in \mathbb{N}$,

$$S_{\mathcal{F}}(n) \leq \sum_{i=0}^{h} \binom{n}{i},$$

and for all $n \geq h$,

$$S_{\mathcal{F}}(n) \leq \left(\frac{en}{h}\right)^{h}.$$

$\rightsquigarrow$ $S_{\mathcal{F}}(n)$ grows very fast as a function of n (faster than exponential), but once the number of datapoints n is larger than the VC dimension h, the growth becomes polynomial.

Sauer-Shelah lemma

Generalization error

One can also use the upper bound on $S_{\mathcal{F}}(n)$ in combination with Theorem 3 in order to obtain the following generalization bound that depends on h and n:

$$\forall f \in \mathcal{F}, \quad R(f) \leq R_S(f) + 2\sqrt{2\frac{h(\log(2n)+1) + \log\frac{2}{\delta}}{n}}.$$

VC theory: Application to Neural Networks

Piecewise linear networks: [BHLM19] proves VC dimension is $\mathcal{O}(WL\log(W))$ where $W =$ number of weights and $L =$ number of layers
$\rightsquigarrow$ Number of datapoints has to scale inversely proportionally to $\mathcal{O}(WL\log(W))$.

Remarks

- Bounds tend to be vacuous for modern deep neural networks that rely on a large number of parameters
- Partly due to generality of VC bounds

Section 4

Rademacher Bound

Rademacher Complexity

Rademacher random variable: σ_i is defined as

$$\sigma_i = \begin{cases} +1, & \text{with prob. } 1/2 \\ -1, & \text{with prob. } 1/2. \end{cases}$$

Definition 6 (Rademacher complexity)

Let $S = \{\mathbf{z}_1, \ldots \mathbf{z}_m\}$ be a set of samples drawn i.i.d. from D. Let $\mathcal{F}$ be a class of functions $f : Z \to \mathbb{R}$. The empirical Rademacher complexity is then defined as

$$\hat{\mathbb{R}}_S(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_i \sigma_i f(z_i) \right].$$

Rademacher Complexity

Remarks

- The supremum measures the maximum correlation between $f(z_i)$ and σ_i over all $f \in \mathcal{F}$
- Taking the expectation over σ, we can say the empirical Rademacher complexity measures the ability of functions from $\mathcal{F}$ to fit random noise.

Expected Rademacher complexity Take the expectation of $\hat{\mathbb{R}}_S(\mathcal{F})$ over all samples of size m, i.e.

$$\mathbb{R}_m(\mathcal{F}) = \mathbb{E}_S[\hat{\mathbb{R}}_S(\mathcal{F})],$$

which measures the expected noise fitting ability of $\mathcal{F}$ over all datasets S.

Rademacher-based uniform convergence

Theorem 7 (Rademacher-based uniform convergence)

Fix the distribution D *and* $\delta \in (0, 1)$. *If* $\mathcal{F} \in \{f : Z \to [0, 1]\}$ *and* $S = \{\mathbf{z}_1, \ldots \mathbf{z}_m\}$. *Then with probability at least* $1 - \delta$ *over the draw of* S,

$$\mathbb{E}_D[f(z)] \leq \hat{\mathbb{E}}_S[f(z)] + 2\mathbb{R}_m(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{m}}$$

Proof sketch

Rademacher bound for Linear Classes

Lemma 8

Let $\mathcal{F}$ be the class of linear predictors, with the L_2-norm of the weights bounded by W_2. Also assume that with probability one that $\|\mathbf{x}\|_2 \leq D$. Then

$$\mathbb{R}(\mathcal{F}) \leq \frac{DW_2}{\sqrt{m}}$$

Note that in the bound derived in Lemma 8, the Lipschitz constant of f w.r.t. to $\mathbf{x}$ appears, since $\|\nabla_{\mathbf{x}} f(\mathbf{z}_i)\|_2 = \|\mathbf{w}\|_2 \leq W_2$.

Composition Lemma

In neural networks, we often use composition of functions. Can we bound the Rademacher complexity of composition of functions?

Lemma 9 (Composition lemma)

For $\mathcal{F} \in \mathbb{R}^d, \phi : \mathbb{R} \to \mathbb{R}$, let $\mathcal{F}' := \{\phi \circ f : f \in \mathcal{F}\}$. If ϕ is L-Lipschitz continuous, i.e. $|\phi(t) - \phi(t')| \leq L|t - t'|$, then for any m,

$$\mathbb{R}(\mathcal{F}') \leq L\mathbb{R}(\mathcal{F}).$$

Proof.

Exercise. □

Application to Neural Networks

One-layer neural network: Based on Lemma 8 and 9, we can derive a bound on the Rademacher complexity of $f(\mathbf{x}) = \sigma(\mathbf{W}\mathbf{x})$ where σ is Lipschitz.

Extension to multiple layers: Use the composition lemma.

Application to Neural Networks

Some words of caution:

- Existing Rademacher bounds typically depend on the product of the weight matrices of a neural network, which lead to loose or even vacuous bounds (i.e. they have no predictive abilities)
- [ZBH$^+$21] showed that neural networks can fit random labels with zero training error. This indicates that these networks can maximize the Rademacher complexity, therefore highlighting some potentially severe shortcomings of such approaches.

This is a very active area of research $\rightsquigarrow$ next time, we will discuss PAC-Bayes bounds.

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian.
Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks.
The Journal of Machine Learning Research, 20(1):2285–2301, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals.
Understanding deep learning (still) requires rethinking generalization.
Communications of the ACM, 64(3):107–115, 2021.