



Rademacher complexity

In [computational learning theory](#) ([machine learning](#) and [theory of computation](#)), **Rademacher complexity**, named after [Hans Rademacher](#), measures richness of a class of sets with respect to a [probability distribution](#). The concept can also be extended to real valued functions.

Definitions

Rademacher complexity of a set

Given a set $A \subseteq \mathbb{R}^m$, the **Rademacher complexity** of A is defined as follows:^{[1][2]:326}

$$\text{Rad}(A) := \frac{1}{m} \mathbb{E}_{\sigma} \left[\sup_{a \in A} \sum_{i=1}^m \sigma_i a_i \right]$$

where $\sigma_1, \sigma_2, \dots, \sigma_m$ are independent random variables drawn from the [Rademacher distribution](#) i.e. $\Pr(\sigma_i = +1) = \Pr(\sigma_i = -1) = 1/2$ for $i = 1, 2, \dots, m$, and $a = (a_1, \dots, a_m)$. Some authors take the absolute value of the sum before taking the supremum, but if A is [symmetric](#) this makes no difference.

Rademacher complexity of a function class

Let $S = \{z_1, z_2, \dots, z_m\} \subset Z$ be a sample of points and consider a function class \mathcal{F} of real-valued functions over Z . Then, the **empirical Rademacher complexity** of \mathcal{F} given S is defined as:

$$\text{Rad}_S(\mathcal{F}) = \frac{1}{m} \mathbb{E}_{\sigma} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^m \sigma_i f(z_i) \right]$$

This can also be written using the previous definition:^{[2]:326}

$$\text{Rad}_S(\mathcal{F}) = \text{Rad}(\mathcal{F} \circ S)$$

where $\mathcal{F} \circ S$ denotes [function composition](#), i.e.:

$$\mathcal{F} \circ S := \{(f(z_1), \dots, f(z_m)) \mid f \in \mathcal{F}\}$$

Let P be a probability distribution over Z . The **Rademacher complexity** of the function class \mathcal{F} with respect to P for sample size m is:

$$\text{Rad}_{P,m}(\mathcal{F}) := \mathbb{E}_{S \sim P^m} [\text{Rad}_S(\mathcal{F})]$$

where the above expectation is taken over an [identically independently distributed](#) (i.i.d.) sample $S = (z_1, z_2, \dots, z_m)$ generated according to P .

Intuition

The Rademacher complexity is typically applied on a function class of models that are used for classification, with the goal of measuring their ability to classify points drawn from a probability space under arbitrary labellings. When the function class is rich enough, it contains functions that can appropriately adapt for each arrangement of labels, simulated by the random draw of σ_i under the expectation, so that this quantity in the sum is maximised.

Examples

1. A contains a single vector, e.g., $A = \{(a, b)\} \subset \mathbb{R}^2$. Then:

$$\text{Rad}(A) = \frac{1}{2} \cdot \left(\frac{1}{4} \cdot (a + b) + \frac{1}{4} \cdot (a - b) + \frac{1}{4} \cdot (-a + b) + \frac{1}{4} \cdot (-a - b) \right) = 0$$

The same is true for every singleton hypothesis class.^{[3]:56}

2. A contains two vectors, e.g., $A = \{(1, 1), (1, 2)\} \subset \mathbb{R}^2$. Then:

$$\begin{aligned}\text{Rad}(A) &= \frac{1}{2} \cdot \left(\frac{1}{4} \cdot \max(1+1, 1+2) + \frac{1}{4} \cdot \max(1-1, 1-2) + \frac{1}{4} \cdot \max(-1+1, -1+2) + \frac{1}{4} \cdot \max(-1-1, -1-2) \right) \\ &= \frac{1}{8}(3+0+1-2) = \frac{1}{4}\end{aligned}$$

Using the Rademacher complexity

The Rademacher complexity can be used to derive data-dependent upper-bounds on the learnability of function classes. Intuitively, a function-class with smaller Rademacher complexity is easier to learn.

Bounding the representativeness

In machine learning, it is desired to have a training set that represents the true distribution of some sample data \mathcal{S} . This can be quantified using the notion of **representativeness**. Denote by \mathbf{P} the probability distribution from which the samples are drawn. Denote by \mathbf{H} the set of hypotheses (potential classifiers) and denote by \mathbf{F} the corresponding set of error functions, i.e., for every hypothesis $\mathbf{h} \in \mathbf{H}$, there is a function $\mathbf{f}_{\mathbf{h}} \in \mathbf{F}$, that maps each training sample (features,label) to the error of the classifier \mathbf{h} (note in this case hypothesis and classifier are used interchangeably). For example, in the case that \mathbf{h} represents a binary classifier, the error function is a 0–1 loss function, i.e. the error function $\mathbf{f}_{\mathbf{h}}$ returns 0 if \mathbf{h} correctly classifies a sample and 1 else. We omit the index and write \mathbf{f} instead of $\mathbf{f}_{\mathbf{h}}$ when the underlying hypothesis is irrelevant. Define:

$L_{\mathbf{P}}(\mathbf{f}) := \mathbb{E}_{z \sim \mathbf{P}}[\mathbf{f}(z)]$ – the expected error of some error function $\mathbf{f} \in \mathbf{F}$ on the real distribution \mathbf{P} ;

$L_{\mathcal{S}}(\mathbf{f}) := \frac{1}{m} \sum_{i=1}^m \mathbf{f}(z_i)$ – the estimated error of some error function $\mathbf{f} \in \mathbf{F}$ on the sample \mathcal{S} .

The representativeness of the sample \mathcal{S} , with respect to \mathbf{P} and \mathbf{F} , is defined as:

$$\text{Rep}_{\mathbf{P}}(\mathbf{F}, \mathcal{S}) := \sup_{\mathbf{f} \in \mathbf{F}} (L_{\mathbf{P}}(\mathbf{f}) - L_{\mathcal{S}}(\mathbf{f}))$$

Smaller representativeness is better, since it provides a way to avoid overfitting: it means that the true error of a classifier is not much higher than its estimated error, and so selecting a classifier that has low estimated error will ensure that the true error is also low. Note however that the concept of representativeness is relative and hence can not be compared across distinct samples.

The expected representativeness of a sample can be bounded above by the Rademacher complexity of the function class:^{[2]:326}

$$\mathbb{E}_{\mathcal{S} \sim \mathbf{P}^m} [\text{Rep}_{\mathbf{P}}(\mathbf{F}, \mathcal{S})] \leq 2 \cdot \mathbb{E}_{\mathcal{S} \sim \mathbf{P}^m} [\text{Rad}(\mathbf{F} \circ \mathcal{S})]$$

Bounding the generalization error

When the Rademacher complexity is small, it is possible to learn the hypothesis class \mathbf{H} using empirical risk minimization.

For example, (with binary error function),^{[2]:328} for every $\delta > 0$, with probability at least $1 - \delta$, for every hypothesis $\mathbf{h} \in \mathbf{H}$:

$$L_{\mathbf{P}}(\mathbf{h}) - L_{\mathcal{S}}(\mathbf{h}) \leq 2 \text{Rad}(\mathbf{F} \circ \mathcal{S}) + 4 \sqrt{\frac{2 \ln(4/\delta)}{m}}$$

Bounding the Rademacher complexity

Since smaller Rademacher complexity is better, it is useful to have upper bounds on the Rademacher complexity of various function sets. The following rules can be used to upper-bound the Rademacher complexity of a set $\mathbf{A} \subset \mathbb{R}^m$.^{[2]:329–330}

1. If all vectors in \mathbf{A} are translated by a constant vector $\mathbf{a}_0 \in \mathbb{R}^m$, then $\text{Rad}(\mathbf{A})$ does not change.
2. If all vectors in \mathbf{A} are multiplied by a scalar $\mathbf{c} \in \mathbb{R}$, then $\text{Rad}(\mathbf{A})$ is multiplied by $|\mathbf{c}|$.
3. $\text{Rad}(\mathbf{A} + \mathbf{B}) = \text{Rad}(\mathbf{A}) + \text{Rad}(\mathbf{B})$.^{[3]:56}
4. (Kakade & Tewari Lemma) If all vectors in \mathbf{A} are operated by a Lipschitz function, then $\text{Rad}(\mathbf{A})$ is (at most) multiplied by the Lipschitz constant of the function. In particular, if all vectors in \mathbf{A} are operated by a contraction mapping, then $\text{Rad}(\mathbf{A})$ strictly decreases.
5. The Rademacher complexity of the convex hull of \mathbf{A} equals $\text{Rad}(\mathbf{A})$.
6. (Massart Lemma) The Rademacher complexity of a finite set grows logarithmically with the set size. Formally, let \mathbf{A} be a set of N vectors in \mathbb{R}^m , and let $\bar{\mathbf{a}}$ be the mean of the vectors in \mathbf{A} . Then:

$$\text{Rad}(\mathcal{A}) \leq \max_{a \in \mathcal{A}} \|a - \bar{a}\| \cdot \frac{\sqrt{2 \log N}}{m}$$

In particular, if \mathcal{A} is a set of binary vectors, the norm is at most \sqrt{m} , so:

$$\text{Rad}(\mathcal{A}) \leq \sqrt{\frac{2 \log N}{m}}$$

Bounds related to the VC dimension

Let \mathcal{H} be a set family whose VC dimension is d . It is known that the growth function of \mathcal{H} is bounded as:

$$\text{for all } m > d + 1: \text{Growth}(\mathcal{H}, m) \leq (em/d)^d$$

This means that, for every set h with at most m elements, $|\mathcal{H} \cap h| \leq (em/d)^d$. The set-family $\mathcal{H} \cap h$ can be considered as a set of binary vectors over \mathbb{R}^m . Substituting this in Massart's lemma gives:

$$\text{Rad}(\mathcal{H} \cap h) \leq \sqrt{\frac{2d \log(em/d)}{m}}$$

With more advanced techniques (Dudley's entropy bound and Haussler's upper bound^[4]) one can show, for example, that there exists a constant C , such that any class of $\{0, 1\}$ -indicator functions with Vapnik–Chervonenkis dimension d has Rademacher complexity upper-bounded by $C\sqrt{\frac{d}{m}}$.

Bounds related to linear classes

The following bounds are related to linear operations on \mathcal{S} – a constant set of m vectors in \mathbb{R}^n .^{[2]:332–333}

1. Define $\mathcal{A}_2 = \{(w \cdot x_1, \dots, w \cdot x_m) \mid \|w\|_2 \leq 1\}$ = the set of dot-products of the vectors in \mathcal{S} with vectors in the unit ball. Then:

$$\text{Rad}(\mathcal{A}_2) \leq \frac{\max_i \|x_i\|_2}{\sqrt{m}}$$

2. Define $\mathcal{A}_1 = \{(w \cdot x_1, \dots, w \cdot x_m) \mid \|w\|_1 \leq 1\}$ = the set of dot-products of the vectors in \mathcal{S} with vectors in the unit ball of the 1-norm. Then:

$$\text{Rad}(\mathcal{A}_1) \leq \max_i \|x_i\|_\infty \cdot \sqrt{\frac{2 \log(2n)}{m}}$$

Bounds related to covering numbers

The following bound relates the Rademacher complexity of a set \mathcal{A} to its external covering number – the number of balls of a given radius r whose union contains \mathcal{A} . The bound is attributed to Dudley.^{[2]:338}

Suppose $\mathcal{A} \subset \mathbb{R}^m$ is a set of vectors whose length (norm) is at most c . Then, for every integer $M > 0$:

$$\text{Rad}(\mathcal{A}) \leq \frac{c \cdot 2^{-M}}{\sqrt{m}} + \frac{6c}{m} \cdot \sum_{i=1}^M 2^{-i} \sqrt{\log(N_{c \cdot 2^{-i}}^{\text{ext}}(\mathcal{A}))}$$

In particular, if \mathcal{A} lies in a d -dimensional subspace of \mathbb{R}^m , then:

$$\forall r > 0 : N_r^{\text{ext}}(\mathcal{A}) \leq (2c\sqrt{d}/r)^d$$

Substituting this in the previous bound gives the following bound on the Rademacher complexity:

$$\text{Rad}(\mathcal{A}) \leq \frac{6c}{m} \cdot \left(\sqrt{d \log(2\sqrt{d})} + 2\sqrt{d} \right) = O\left(\frac{c\sqrt{d \log(d)}}{m}\right)$$

Gaussian complexity

Gaussian complexity is a similar complexity with similar physical meanings, and can be obtained from the Rademacher complexity using the random variables g_i instead of σ_i , where g_i are Gaussian i.i.d. random variables with zero-mean and variance 1, i.e. $g_i \sim \mathcal{N}(0, 1)$. Gaussian and Rademacher complexities are known to be equivalent up to logarithmic factors.

Equivalence of Rademacher and Gaussian complexity

Given a set $A \subseteq \mathbb{R}^n$ then it holds that^[5]:

$$\frac{G(A)}{2\sqrt{\log n}} \leq \text{Rad}(A) \leq \sqrt{\frac{\pi}{2}} G(A)$$

Where $G(A)$ is the Gaussian Complexity of A. As an example, consider the rademacher and gaussian complexities of the L1 ball. The Rademacher complexity is given by exactly 1, whereas the Gaussian complexity is on the order of $\sqrt{\log d}$ (which can be shown by applying known properties of suprema of a set of subgaussian random variables).^[5]

References

1. Balcan, Maria-Florina (November 15–17, 2011). "Machine Learning Theory – Rademacher Complexity" (<https://www.cs.cmu.edu/~ninamf/ML11/lect1117.pdf>) (PDF). Retrieved 10 December 2016.
 2. Chapter 26 in Shalev-Shwartz, Shai; Ben-David, Shai (2014). *Understanding Machine Learning – from Theory to Algorithms*. Cambridge University Press. ISBN 9781107057135.
 3. Mohri, Mehryar; Rostamizadeh, Afshin; Talwalkar, Ameet (2012). *Foundations of Machine Learning*. US, Massachusetts: MIT Press. ISBN 9780262018258.
 4. Bousquet, O. (2004). Introduction to Statistical Learning Theory. *Biological Cybernetics*, 3176(1), 169–207. doi:10.1007/978-3-540-28650-9_8 (https://doi.org/10.1007%2F978-3-540-28650-9_8)
 5. Wainwright, Martin (2019). *High-dimensional statistics : a non-asymptotic viewpoint* (<https://www.worldcat.org/oclc/1089254580>). Cambridge, United Kingdom. pp. Exercise 5.5. ISBN 978-1-108-62777-1. OCLC 1089254580 (<https://search.worldcat.org/oclc/1089254580>).
- Peter L. Bartlett, Shahar Mendelson (2002) *Rademacher and Gaussian Complexities: Risk Bounds and Structural Results*. Journal of Machine Learning Research 3 463–482
 - Giorgio Gnecco, Marcello Sanguineti (2008) *Approximation Error Bounds via Rademacher's Complexity*. Applied Mathematical Sciences, Vol. 2, 2008, no. 4, 153–176
-

Retrieved from "https://en.wikipedia.org/w/index.php?title=Rademacher_complexity&oldid=1220065433"