Title: Rademacher complexity

URL: https://en.wikipedia.org/wiki/Rademacher_complexity

PageID: 14529261

Categories: Category:Machine learning, Category:Measures of 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 ⊆ R m {\displaystyle A\subseteq \mathbb {R} ^{m}} , the Rademacher complexity of A is defined as follows: [1] [2] : 326

where σ 1 , σ 2 , … , σ m {\displaystyle \sigma _{1},\sigma _{2},\dots ,\sigma _{m}} are independent random variables drawn from the Rademacher distribution i.e. Pr (σ i = + 1) = Pr (σ i = − 1) = 1 / 2 {\displaystyle \Pr(\sigma _{i}=+1)=\Pr(\sigma _{i}=-1)=1/2} for i = 1 , 2 , … , m {\displaystyle i=1,2,\dots ,m} , and a = (a 1 , … , a m) {\displaystyle a=(a_{1},\ldots ,a_{m})} . Some authors take the absolute value of the sum before taking the supremum, but if A {\displaystyle A} is symmetric this makes no difference.

Rademacher complexity of a function class

Let S = { z 1 , z 2 , … , z m } ⊂ Z {\displaystyle S=\{z_{1},z_{2},\dots ,z_{m}\}\subset Z} be a sample of points and consider a function class F {\displaystyle {\mathcal {F}}} of real-valued functions over Z {\displaystyle Z} . Then, the empirical Rademacher complexity of F {\displaystyle {\mathcal {F}}} given S {\displaystyle S} is defined as:

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

where F ■ S {\displaystyle {\mathcal {F}}\circ S} denotes function composition , i.e.:

The worst case empirical Rademacher complexity is Rad ¯ m (F) = sup S = { z 1 , … , z m } Rad S ■ (F) {\displaystyle {\overline {\operatorname {Rad} }}_{m}({\mathcal {F}})=\sup _{S=\{z_{1},\dots ,z_{m}\}}\operatorname {Rad} _{S}({\mathcal {F}})} Let P {\displaystyle P} be a probability distribution over Z {\displaystyle Z} . The Rademacher complexity of the function class F {\displaystyle {\mathcal {F}}} with respect to P {\displaystyle P} for sample size m {\displaystyle m} is:

where the above expectation is taken over an identically independently distributed (i.i.d.) sample S = (z 1 , z 2 , … , z m) {\displaystyle S=(z_{1},z_{2},\dots ,z_{m})} generated according to P {\displaystyle 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 {\displaystyle \sigma _{i}} under the expectation, so that this quantity in the sum is maximised.

The Rademacher complexity of a set Rad ■ (A) := 1 m E σ [sup a ∈ A ∑ i = 1 m σ i a i] {\displaystyle \operatorname {Rad} (A):={\frac {1}{m}}\mathbb {E} _{\sigma }\left[\sup _{a\in A}\sum _{i=1}^{m}\sigma _{i}a_{i}\right]} can be rewritten as

Rad ■ (A) := 1 m 2 m ∑ σ ∈ { − 1 / m , + 1 / m } m [sup a ∈ A ■ σ , a ■] {\displaystyle \operatorname {Rad} (A):={\frac {1}{{\sqrt {m}}2^{m}}}\sum _{\sigma \in \{-1/{\sqrt {m}},+1/{\sqrt {m}}\}^{m}}\left[\sup _{a\in A}\langle \sigma ,a\rangle \right]} Each term in the summation is the farthest distance of the set A {\displaystyle A} from the origin, along a unit-length direction σ {\displaystyle \sigma } . The directions are along the vertices of a hypercube . Thus, we can also write it as Rad ■ (A) := 1 2 m 1 2 m – 1 ∑ σ ∈ { – 1 / m , + 1 / m } m / { – 1 , + 1 } [sup a ∈ A ■ σ , a ■ – inf a ∈ A ■ σ , a ■] {\displaystyle \operatorname {Rad} (A):={\frac {1}{2{\sqrt {m}}}}{\frac {1}{2^{m-1}}}\sum _{\sigma \in \{-1/{\sqrt {m}},+1/{\sqrt {m}}\}^{m}/\{-1,+1\}}\left[\sup _{a\in A}\langle \sigma ,a\rangle -\inf _{a\in A}\langle \sigma ,a\rangle \right]} Here, the set { – 1 / m , + 1 / m } m / { – 1 , + 1 } {\displaystyle \{-1/{\sqrt {m}},+1/{\sqrt {m}}\}^{m}/\{-1,+1\}} denotes half of the vertices of a hypercube, selected so that each diagonal has exactly one vertex selected.

In words, this states that 2 m Rad ■ (A) {\displaystyle 2{\sqrt {m}}\operatorname {Rad} (A)} is precisely the average width of the set A {\displaystyle A} along all diagonal directions of a hypercube.

Examples

Singleton sets have 0 width in any direction, so it has Rademacher complexity 0. [3] : 56

If A = { (1 , 1) , (1 , 2) } ⊂ R 2 {\displaystyle A=\{(1,1),(1,2)\}\subset \mathbb {R} ^{2}} , then it has average width 1 / 2 {\displaystyle 1/{\sqrt {2}}} along the two diagonal directions of the square, so it has Rademacher complexity 1 / 4 {\displaystyle 1/4} .

The unit cube [0 , 1] m {\displaystyle [0,1]^{m}} has constant width m {\displaystyle {\sqrt {m}}} along the diagonal directions, so it has Rademacher complexity 1 / 2 {\displaystyle 1/2} . Similarly, the unit cross-polytope { x ∈ R m : ■ x ■ 1 ≤ 1 } {\displaystyle \{x\in \mathbb {R} ^{m}:\|x\|_{1}\leq 1\}} has constant width 2 / m {\displaystyle 2/{\sqrt {m}}} along the diagonal directions, so it has Rademacher complexity 1 / m {\displaystyle 1/m} .

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

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

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: If F {\displaystyle {\mathcal {F}}} is a set of functions with range within [0 , 1] {\displaystyle [0,1]} , then [2] : 326 [4]

Furthermore, the representativeness is concentrated around its expectation: [4] For any ■ {\displaystyle \epsilon } , with probability ≥ 1 – 2 e – 2 ■ 2 m {\displaystyle \geq 1-2e^{-2\epsilon

^{2}m}} , Rep P ■ (F , S) ∈ E S ∼ P m [Rep P ■ (F , S)] ± ■ {\displaystyle \operatorname {Rep} _{P}({\mathcal {F}},S)\in \mathbb {E} _{S\sim P^{m}}[\operatorname {Rep} _{P}({\mathcal {F}},S)]\pm \epsilon }

Bounding the generalization error

The Rademacher complexity is a theoretical justification for empirical risk minimization .

When the error function is binary (0-1 loss), for every δ > 0 {\displaystyle \delta >0} ,

with probability at least 1 − δ {\displaystyle 1-\delta } . [2] : 328

There exists a constant c > 0 {\displaystyle c>0} , such that when the error function is squared ■ (y ^ , y) := (y ^ − y) 2 {\displaystyle \ell ({\hat {y}},y):=({\hat {y}}-y)^{2}} , and the function class F {\displaystyle {\mathcal {F}}} consists of functions with range within [− 1 , + 1] {\displaystyle [-1,+1]} , then for any δ > 0 {\displaystyle \delta >0} L P (f) − L S (f) ≤ c [L S (f) + (ln ■ m) 4 Rad ¯ m (F) 2 + ln ■ (1 / δ) m] , ∀ f ∈ F {\displaystyle L_{P}(f)-L_{S}(f)\leq c\left[L_{S}(f)+(\ln m)^{4}{\overline {\operatorname {Rad} }}_{m}({\mathcal {F}})^{2}+{\frac {\ln(1/\delta)}{m}}\right],\quad \forall f\in {\mathcal {F}}} with probability at least 1 − δ {\displaystyle 1-\delta } . [4] : Theorem 2.2

Oracle inequalities

Let the Bayes risk L ∗ = inf f L P (f) {\displaystyle L^{*}=\inf _{f}L_{P}(f)} , where f {\displaystyle f} can be any measurable function.

Let the function class F {\displaystyle {\mathcal {F}}} be split into "complexity classes" F r {\displaystyle {\mathcal {F}}_{r}} , where r ∈ R {\displaystyle r\in \mathbb {R} } are levels of complexity. Let p r {\displaystyle p_{r}} be real numbers. Let the complexity measure function p {\displaystyle p} be defined such that p (f) := min { p r : f ∈ F r } {\displaystyle p(f):=\min\{p_{r}:f\in {\mathcal {F}}_{r}\}} .

For any dataset S {\displaystyle S} , let f ^ {\displaystyle {\hat {f}}} be a minimizer of L S (f) + p (f) {\displaystyle L_{S}(f)+p(f)} . If sup f ∈ F r | L P (f) − L S (f) | ≤ p r , ∀ r {\displaystyle \sup _{f\in {\mathcal {F}}_{r}}|L_{P}(f)-L_{S}(f)|\leq p_{r},\quad \forall r} then we have the oracle inequality L (f ^) − L ∗ ≤ inf r (inf f ∈ F r L (f) − L ∗ + 2 p r) {\displaystyle L({\hat {f}})-L^{*}\leq \inf _{r}\left(\inf _{f\in {\mathcal {F}}_{r}}L(f)-L^{*}+2p_{r}\right)} Define f r ∗ ∈ arg ■ min f ∈ F r L (f) {\displaystyle f_{r}^{*}\in \arg \min _{f\in {\mathcal {F}}_{r}}L(f)} . If we further assume r ≤ s implies F r ⊆ F s and p r ≤ p s {\displaystyle r\leq s{\text{ implies }}{\mathcal {F}}_{r}\subseteq {\mathcal {F}}_{s}{\text{ and }}p_{r}\leq p_{s}} and ∀ r , sup f ∈ F r (L P (f) − L P (f r ∗) − 2 (L S (f) − L S (f r ∗))) ≤ 2 p r / 7 sup f ∈ F r (L S (f) − L S (f r ∗) − 2 (L P (f) − L P (f r ∗))) ≤ 2 p r / 7 {\displaystyle {\begin{aligned}\forall r,\sup _{f\in {\mathcal {F}}_{r}}\left(L_{P}(f)-L_{P}\left(f_{r}^{*}\right)-2\left(L_{S}(f)-L_{S}\left(f_{r}^{*}\right)\right)\right)&\leq 2p_{r}/7\\\sup _{f\in {\mathcal {F}}_{r}}\left(L_{S}(f)-L_{S}\left(f_{r}^{*}\right)-2\left(L_{P}(f)-L_{P}\left(f_{r}^{*}\right)\right)\right)&\leq 2p_{r}/7\end{aligned}}} then we have the oracle inequality L P (f ^) − L ∗ ≤ inf r (inf f ∈ F r L P (f) − L ∗ + 3 p r) {\displaystyle L_{P}({\widehat {f}})-L^{*}\leq \inf _{r}\left(\inf _{f\in {\mathcal {F}}_{r}}L_{P}(f)-L^{*}+3p_{r}\right)}

[4] : Theorem 2.3

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 A ⊂ R m {\displaystyle A\subset \mathbb {R} ^{m}} . [2] : 329–330

If all vectors in A {\displaystyle A} are translated by a constant vector a 0 ∈ R m {\displaystyle a_{0}\in \mathbb {R} ^{m}} , then Rad(A) does not change.

If all vectors in A {\displaystyle A} are multiplied by a scalar c ∈ R {\displaystyle c\in \mathbb {R} } , then Rad(A) is multiplied by | c | {\displaystyle |c|} .

Rad ■ (A + B) = Rad ■ (A) + Rad ■ (B) {\displaystyle \operatorname {Rad} (A+B)=\operatorname {Rad} (A)+\operatorname {Rad} (B)} . [3] : 56

(Kakade & Tewari Lemma) If all vectors in A {\displaystyle A} are operated by a Lipschitz function , then Rad(A) is (at most) multiplied by the Lipschitz constant of the function. In particular, if all vectors in A {\displaystyle A} are operated by a contraction mapping , then Rad(A) strictly decreases.

The Rademacher complexity of the convex hull of A {\displaystyle A} equals Rad(A).

(Massart Lemma) The Rademacher complexity of a finite set grows logarithmically with the set size. Formally, let A {\displaystyle A} be a set of N {\displaystyle N} vectors in R m {\displaystyle \mathbb {R} ^{m}} , and let a ¯ {\displaystyle {\bar {a}}} be the mean of the vectors in A {\displaystyle A} . Then:

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

Bounds related to the VC dimension

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

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

With more advanced techniques (Dudley's entropy bound and Haussler's upper bound [5]) one can show, for example, that there exists a constant C {\displaystyle C} , such that any class of { 0 , 1 } {\displaystyle \{0,1\}} -indicator functions with Vapnik–Chervonenkis dimension d {\displaystyle d} has Rademacher complexity upper-bounded by C d m {\displaystyle C{\sqrt {\frac {d}{m}}}} .

Bounds related to linear classes

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

Define A 2 = { (w ⋅ x 1 , … , w ⋅ x m) ■ ■ w ■ 2 ≤ 1 } = {\displaystyle A_{2}=\{(w\cdot x_{1},\ldots ,w\cdot x_{m})\mid \|w\|_{2}\leq 1\}=} the set of dot-products of the vectors in S {\displaystyle S} with vectors in the unit ball . Then:

Define A 1 = { (w ⋅ x 1 , … , w ⋅ x m) ■ ■ w ■ 1 ≤ 1 } = {\displaystyle A_{1}=\{(w\cdot x_{1},\ldots ,w\cdot x_{m})\mid \|w\|_{1}\leq 1\}=} the set of dot-products of the vectors in S {\displaystyle S} with vectors in the unit ball of the 1-norm. Then:

Bounds related to covering numbers

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

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

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

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

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 {\displaystyle g_{i}} instead of σ i {\displaystyle \sigma _{i}} , where g i {\displaystyle g_{i}} are Gaussian i.i.d. random variables with zero-mean and variance 1, i.e. g i ∼ N (0 , 1) {\displaystyle 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 ⊆ R n {\displaystyle A\subseteq \mathbb {R} ^{n}} then it holds that [6] : G (A) 2 log ■ n ≤ Rad (A) ≤ π 2 G (A) {\displaystyle {\frac {G(A)}{2{\sqrt {\log {n}}}}}\leq {\text{Rad}}(A)\leq {\sqrt {\frac {\pi }{2}}}G(A)} Where G (A) {\displaystyle 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 log ■ d {\displaystyle {\sqrt {\log d}}} (which can be shown by applying known properties of suprema of a set of subgaussian random variables). [6]

References

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