Human Rademacher Complexity

Xiaojin Zhu, Timothy Rogers, Bryan Gibson   University of Wisconsin-Madison, USA   Contact: jerryzhu@cs.wisc.edu

Abstract

We introduce to cognitive psychology a standard tool in machine learning, namely, the Rademacher complexity of the human mind (or, technically, the set of binary classification functions our mind can entertain). Rademacher complexity measures the mind's ability to fit random labels, and thus is a novel measure of human learning capacity. Furthermore, Rademacher complexity can be used to bound a human learner's true error based on her training sample error.

For machine learning researchers, our work serves as a novel and intuitive application of Rademacher complexity and its generalization error bound. It is another example that machine learning and human learning can be studied under the same mathematical principles.

Rademacher Complexity

X: a domain (i.e., stimulus space) with marginal distribution P_X
$x_1, \ldots, x_n \sim P_X$: instances
$F = \{f: X \rightarrow R\}$: a set of functions
Rademacher complexity measures the *capacity* of F.

Definition: For a set of real-valued functions F with domain X, a distribution P_X on X, and a size n, the Rademacher complexity is

$$R(\mathcal{F}, \mathcal{X}, P_X, n) = \mathbb{E}_{\mathbf{x}\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{n}\sum_{i=1}^{n}\sigma_i f(x_i)\right|\right],$$

where the expectation is over training sample $\mathbf{x} = x_1, \ldots, x_n \sim P_X$, and the $\{-1, 1\}$-valued random labels $\boldsymbol{\sigma} = \sigma_1, \ldots, \sigma_n \sim$ Bernoulli(0.5, 0.5).

Comments:

1. Intuition: if for any random training sample $(\mathbf{x}, \boldsymbol{\sigma})$, there always exists $f \in F$ such that $f(\mathbf{x})$ strongly correlates with the random labels $\boldsymbol{\sigma}$, then F is rich and has high capacity.
2. Rademacher complexity remains the same for different classification or regression tasks on X that one might define (i.e., it is insensitive to intended labels y).
3. Rademacher complexity can be estimated by approximating the expectation with sample-average on $(\mathbf{x}, \boldsymbol{\sigma})$.

This work is supported in part by AFOSR grant FA9550-09-1-0313 and the Wisconsin Alumni Research Foundation.

The Rademacher Complexity of the Human Mind

Let $F = H_a$ be the set of binary classification functions on X that the human mind has access to. That is, any $f \in H_a$ defines a particular way a subject categorizes $x \in X$ into label $f(x) \in \{-1, 1\}$. We are interested in the Rademacher complexity of H_a. However, H_a is implicit and as a whole unobservable; the *sup* operation cannot be done explicitly. We propose a "learning the noise" procedure:

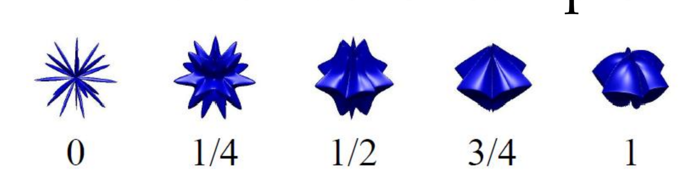
1. On a sheet of paper, show n random training instances $(x_1, \sigma_1) \ldots, (x_n, \sigma_n)$ to a participant for three minutes. The participant is informed that there are only two categories, that order doesn't matter, that they will use what they learned to categorize more instances.
2. The sheet is taken away; Perform a filler task.
3. The participant is given another sheet with $x_1, \ldots, x_n$ in a different order, and asked to categorize them. They do not know these are the same instances in step 1. No time limit. Let their classification labels be $f^*(x_1) \ldots f^*(x_n)$.

Assumption: $\sup_{f\in H_a}\left|\frac{2}{n}\sum_{i=1}^{n}\sigma_i f(x_i)\right| \approx \left|\frac{2}{n}\sum_{i=1}^{n}\sigma_i f^*(x_i)\right|$

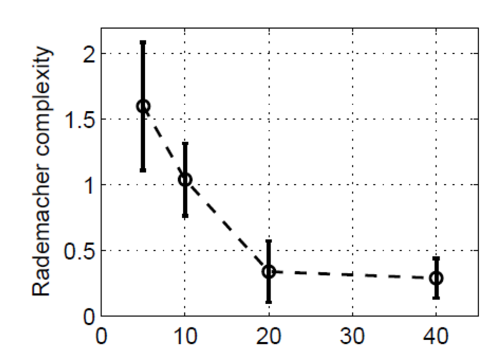
Averaging over m participants:

$$R(H_a, \mathcal{X}, P_X, n) = \frac{1}{m}\sum_{j=1}^{m}\left|\frac{2}{n}\sum_{i=1}^{n}\sigma_i^{(j)} f^{(j)}(x_i^{(j)})\right|$$

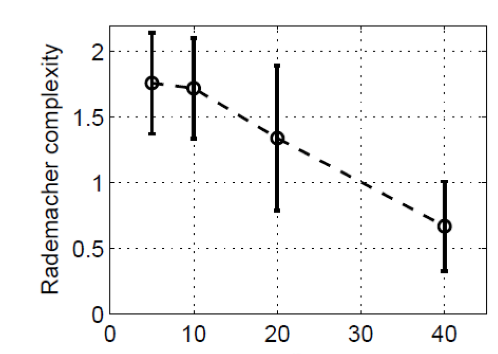
Human Rademacher complexity on two domains from 80 subjects:



rape -5.60   killer -5.55   funeral -5.47   . . .   fun 4.91   laughter 4.95   joy 5.19



$R(H_a, \text{Shape}, \text{uniform}, n)$



$R(H_a, \text{Word}, \text{uniform}, n)$

Observations:

1. R decreases with n.
2. The Word domain has higher R.
3. Post-interviews reveal some participants' f^*
   a) Mnemonics. Training instances (grenade, B), (skull, A), (conflict, A), (meadow, B), (queen, B), ➔ "a queen was sitting in a meadow and then a grenade was thrown (B = before), then this started a conflict ending in bodies & skulls (A = after)."
   b) Idiosyncratic and imperfect rules: whether the item "tastes good," "relates to motel service," or "physical vs. abstract."

Bounding Human Generalization Error

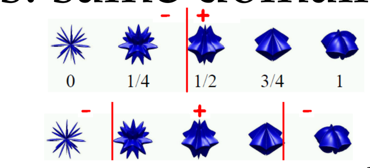
Consider any binary categorization task with joint probability P_{XY}
The observed training sample error of f: $\hat{e}(f) = \frac{1}{n}\sum_{i=1}^{n}(y_i \neq f(x_i))$
The true error of f: $e(f) = \mathbb{E}_{(x,y)\overset{iid}{\sim}P_{XY}}[(y \neq f(x))]$

Rademacher complexity can bound the "amount of overfitting" (Bartlett & Mendelson): with probability at least 1-δ, every function $f \in F$ satisfies

$$e(f) - \hat{e}(f) \leq \frac{R(\mathcal{F}, \mathcal{X}, P_X, n)}{2} + \sqrt{\frac{\ln(1/\delta)}{2n}}$$

In particular, the bound holds for the classifier f^* used by a human. Meaning: if the RHS is large, good training performance may not guarantee good test performance.

Example tasks: same domain, but different classification goals.



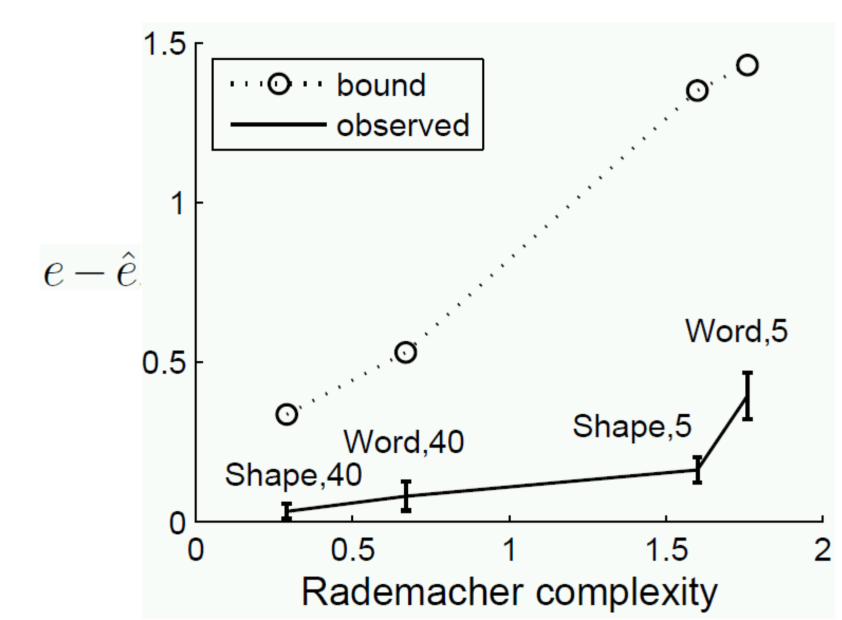
Shape-+   WordEmotion (pos/neg)
Shape-+-   WordLength (>5)

Same procedure, except replacing random σ with true label y, and step 3 containing n training and 100 test instances. 40 subjects.
The bound always holds: (δ=0.05)

condition	ID	$\hat{e}$	bound e	e
Shape-+ n=5	81	0.00	1.35	0.05
	82	0.00	1.35	0.22
	83	0.00	1.35	0.10
	84	0.00	1.35	0.09
	85	0.00	1.35	0.07
Shape-+ n=40	86	0.05	0.39	0.04
	87	0.03	0.36	0.14
	88	0.03	0.36	0.03
	89	0.00	0.34	0.04
	90	0.00	0.34	0.01
Shape-+- n=5	91	0.00	1.35	0.23
	92	0.00	1.35	0.27
	93	0.00	1.35	0.21
	94	0.00	1.35	0.40
	95	0.20	1.55	0.18
Shape-+- n=40	96	0.12	0.46	0.16
	97	0.32	0.66	0.50
	98	0.15	0.49	0.08
	99	0.15	0.49	0.11
	100	0.03	0.36	0.10

condition	ID	$\hat{e}$	bound e	e
WordEmotion n=5	101	0.00	1.43	0.58
	102	0.00	1.43	0.46
	103	0.00	1.43	0.04
	104	0.00	1.43	0.03
	105	0.00	1.43	0.31
WordEmotion n=40	106	0.70	1.23	0.65
	107	0.00	0.53	0.04
	108	0.00	0.53	0.00
	109	0.62	1.15	0.53
	110	0.00	0.53	0.05
WordLength n=5	111	0.00	1.43	0.46
	112	0.00	1.43	0.69
	113	0.00	1.43	0.55
	114	0.00	1.43	0.26
	115	0.00	1.43	0.57
WordLength n=40	116	0.12	0.65	0.51
	117	0.45	0.98	0.55
	118	0.00	0.53	0.00
	119	0.15	0.68	0.29
	120	0.15	0.68	0.37

Furthermore, the bound and the *actual* amount of overfitting agree on the trend:



A few overfitting f^*:

- Subject 102 "anything related to emitting light"
- Subject 111 "things you can go inside"
- Subject 114 "odd number of syllables"