Machine Learning (CS 181): 23. Computational Learning Theory

David C. Parkes and Sasha Rush

Spring 2017

Contents

Contents

A Theory of Learning

What might we want from a useful theory?

- Complexity theory: which problems are computable in polynomial time, which not?
- Learning theory:
  - Which hypotheses are learnable with a polynomial number of examples, which not?
  - . . . and in a polynomial amount of time?

Ok, but what does it mean to be learnable? Certainly we'll need the training data to "look like" the future.

The framework

- Inputs $\mathbf{x} \in \{0,1\}^m$ sampled according to distribution p (for training and test):

$$\mathbf{x} \sim p$$

- True hypothesis h, so that the target value y for input $\mathbf{x}$ is

$$y = h(\mathbf{x}) \in \{0,1\}$$

- Classification problem, no noise.
- Let H denote the hypothesis space ($h \in H$).

Can we learn h exactly?

- Can we learn h exactly from a finite sample $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$?
- No!

Can we learn h exactly?

- Can we learn h exactly from a finite sample $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$?
- No!

Can we learn h approximately?

- Define error of a learned hypothesis $\hat{h}$:

$$\text{error}_p(\hat{h}) = \sum_{\mathbf{x}} p(\mathbf{x}) \mathbb{I}[h(\mathbf{x}) \neq \hat{h}(\mathbf{x})]$$

  where $\mathbb{I}[\cdot]$ is the indicator function.

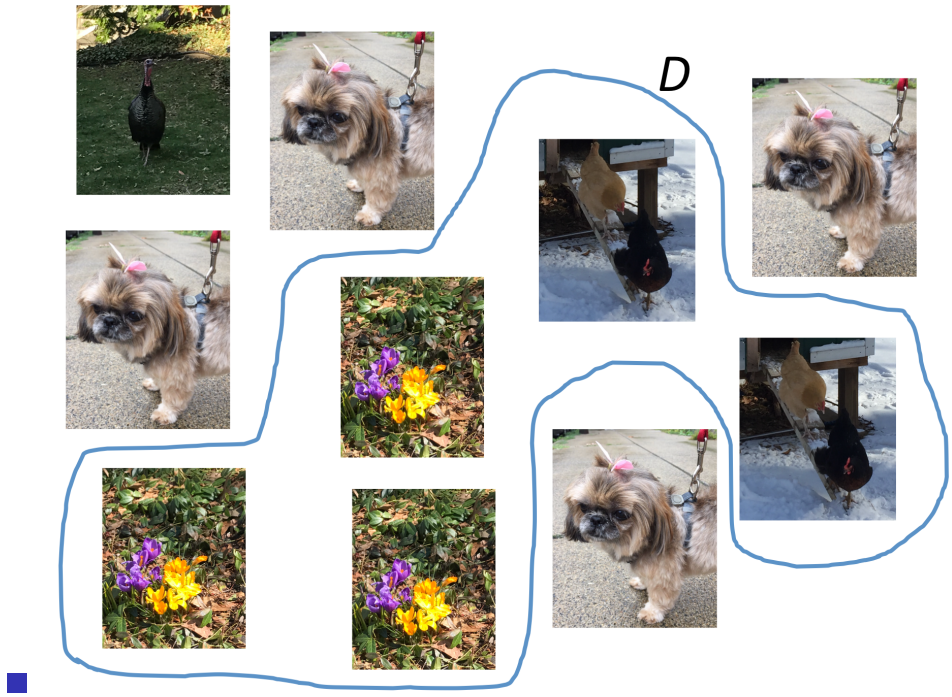
- Can we achieve

$$error_p(\hat{h}) \leq \epsilon,$$

  for small $\epsilon > 0$?

Can we learn h approximately?

- Can we achieve $error_p(\hat{h}) \leq \epsilon$, for small $\epsilon > 0$? No!



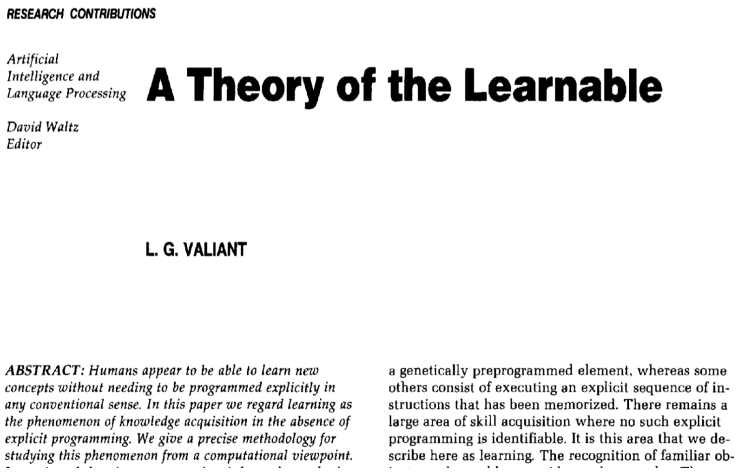
What we can do!

Probably, Approximately Correct learning:

- For any distribution p, for any true hypothesis $h \in H$,
  - Learn classifier $\hat{h}$ with $error_p(\hat{h}) \leq \epsilon$ with probability $\geq 1 - \delta$, for small $\epsilon > 0$, small $\delta > 0$.
  - Do this with a polynomial number of examples, and a polynomial amount of computation.

Understand which hypotheses spaces H are PAC learnable. (How can this be possible? What about turkeys?)

The PAC Model (Valiant'84)



RESEARCH CONTRIBUTIONS

Artificial Intelligence and Language Processing

David Waltz Editor

A Theory of the Learnable

L. G. VALIANT

ABSTRACT: Humans appear to be able to learn new concepts without needing to be programmed explicitly in any conventional sense. In this paper we regard learning as the phenomenon of knowledge acquisition in the absence of explicit programming. We give a precise methodology for studying this phenomenon from a computational viewpoint.

a genetically preprogrammed element, whereas some others consist of executing an explicit sequence of instructions that has been memorized. There remains a large area of skill acquisition where no such explicit programming is identifiable. It is this area that we describe here as learning. The recognition of familiar ob-

A hypothesis space H is PAC-learnable if, for most training sets, we can learn an approximately correct hypothesis, and do so in polynomial time.

Our agenda

1. Define a consistent learner: returns a hypothesis that agrees with every example in D whenever such a hypothesis exists.
2. Sample complexity: how many examples are needed, such that every hypothesis consistent with n examples will be probably, approximately correct?
3. PAC learnability: suppose H has polynomial sample complexity, can a consistent hypothesis be efficiently computed?
4. Extend to infinite hypothesis spaces. (The VC dimension!)

Bonus: a few remarks about my research, and then Sasha about LA181.

Contents

Sample complexity (1 of 7)

- What is the probability that some hypothesis, h_1, is consistent with a random example?
  - $1 - error_p(h_1)$
- What is the probability that some hypothesis, h_1, is consistent with n random examples?
  - $(1 - error_p(h_1))^n$
- Suppose a hypothesis is 'bad', such that $error_p(h_1) > \epsilon$. Then:

$$\text{Pr(prob bad } h_1 \text{ consistent with } D) < (1 - \epsilon)^n$$

Sample complexity (2 of 7)

- We want to understand the probability that one or more bad hypotheses are consistent with D.
- Let A_k = event that bad hypothesis k is consistent with D
- Union bound:

$$\Pr(A_1 \vee A_2 \vee \ldots \vee A_\ell) \leq \Pr(A_1) + \Pr(A_2) + \ldots + \Pr(A_\ell)$$

- For some bad h_k, we have

$$\Pr(\text{prob bad } h_k \text{ consistent with } D) < (1-\epsilon)^n$$

- For ℓ bad hypotheses:

$$\Pr(\text{one or more bad hyp. const. with } n \text{ instances})$$
$$< \ell(1-\epsilon)^n \leq |H|(1-\epsilon)^n$$

Sample complexity (3 of 7)

- We have

$$\Pr(\text{one or more bad hyp. const. with } n \text{ instances}) < |H|(1-\epsilon)^n$$

- We want this bad event to occur with probability less than δ, and find n such that,

$$|H|(1-\epsilon)^n \leq \delta$$

- For consistent learners to provide the ϵ-, δ- guarantee, it is sufficient for:

$$n \geq \frac{1}{\epsilon}\left[\ln|H| + \ln\frac{1}{\delta}\right]$$

- If $|H|$ grows exponentially in m (# attributes), then the sample complexity is polynomial in m, $1/\epsilon$, $1/\delta$.

Sample Complexity (4 of 7)

- $H_1 =$ monotone conjunctive, e.g. $h(\mathbf{x}) = x_1 \wedge x_4$
- $|H_1| = 2^m$, and thus $n \geq \frac{1}{\epsilon}(m \ln 2 + \ln(\frac{1}{\delta}))$.
- Consistent learner: drop all negated literals in positive examples.

$\mathbf{x}$					y	$\hat{h}$
1	1	0	1	1	1	$x_1 x_2 x_4 x_5$
0	1	0	0	0	0	ok
1	1	0	1	0	1	$x_1 x_2 x_4$
0	0	1	1	1	0	ok
1	1	1	0	0	1	$x_1 x_2$

- Complexity $O(mn)$. PAC learnable too!

Sample complexity (5 of 7)

- Conjecture 1: if H has polynomial sample complexity, then so does any $H' \subset H$.
- Conjecture 2: if H is PAC learnable, then so is any $H' \subset H$.

Sample complexity (6 of 7)

- $H_2 =$ conjunctive formula, e.g. $h(\mathbf{x}) = x_1 \wedge \neg x_3 \wedge x_5$
  - $|H_2| = 3^m$, and so polynomial sample complexity.
  - Also PAC learnable.
- $H_3 = k$-CNF, e.g. 2-CNF would be $(x_1 \vee \neg x_3) \wedge (x_2 \vee x_4) \wedge \ldots$
  - $(2m)^k$ unique clauses, and so $|H_3| = 2^{O(m^k)}$, and sample complexity is poly. in m^k, and thus poly. in m for fixed k.
  - Also PAC learnable.
- $H_4 = k$-term DNF, e.g. $x_1 x_4 \neg x_6 x_8 \vee x_2 x_7 \vee \ldots$ (at most k conj. terms)
  - $|H_4| = (3^m)^k = 3^{mk}$. Thus, polynomial sample complexity.
  - Not PAC learnable, and even though $H_4 \subseteq H_3$!

Sample complexity (7 of 7)

- Conjecture 1: if H has polynomial sample complexity, then so does any $H' \subset H$. **True**.
- Conjecture 2: if H is PAC learnable, then so is any $H' \subset H$. **False**.

Contents

But what about an infinite hypothesis space?

- We have:
$$n \geq \frac{1}{\epsilon}\left[\ln |H| + \ln \frac{1}{\delta}\right]$$
- What if there is an infinite hypothesis space, i.e. $|H| = \infty$?
- Adopt the VC dimension, $VC(H)$ (Vapnik-Chervonenkis'71). Provides the effective number of degrees of freedom in H.
- Obtain:
$$n > \frac{1}{\epsilon}\left[8\, VC(H) \log_2 \frac{13}{\epsilon} + 4 \log_2 \frac{2}{\delta}\right]$$
- Simplifying, the following number of examples are sufficient:
$$n \geq O(\frac{1}{\epsilon}\left[VC(H) \ln \frac{1}{\epsilon} + \ln \frac{1}{\delta}\right])$$

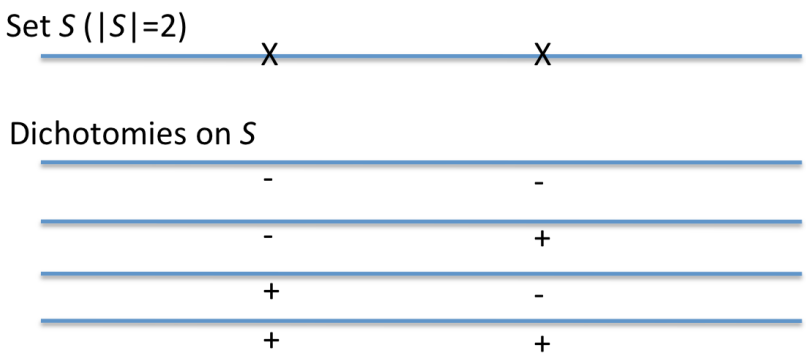
VC dimension (1 of 8)

- Fix some set of points S ($|S| = n$). A dichotomy on S assigns some to be 'positive' and some to be 'negative'.
- A hypothesis space H shatters a particular set S if, for every dichotomy, there is a $h \in H$ that represents the dichotomy.

VC dimension (1 of 8)

- Fix some set of points $S \subseteq X^n$. A dichotomy on S assigns some to be 'positive' and some to be 'negative'.
- A hypothesis space H shatters a particular set S if, for every dichotomy, there is a $h \in H$ that represents the dichotomy.

e.g., consider $X = \mathbb{R}$, and H=closed intervals.

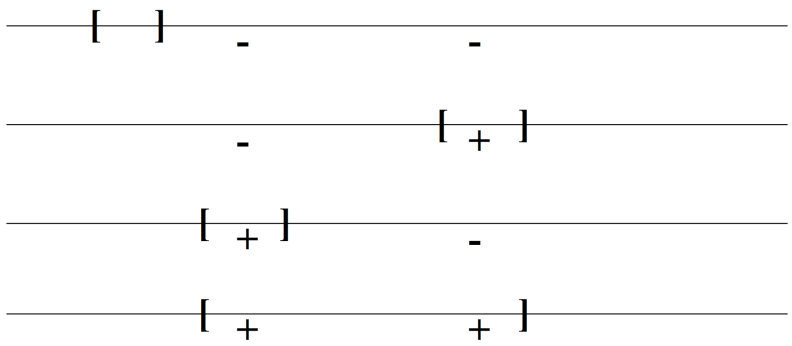


VC dimension (2 of 8)

Definition (VC dimension)

$VC(H)$ is the size of the largest set S shattered by H.

$H =$ closed intervals. Two points that can be shattered:

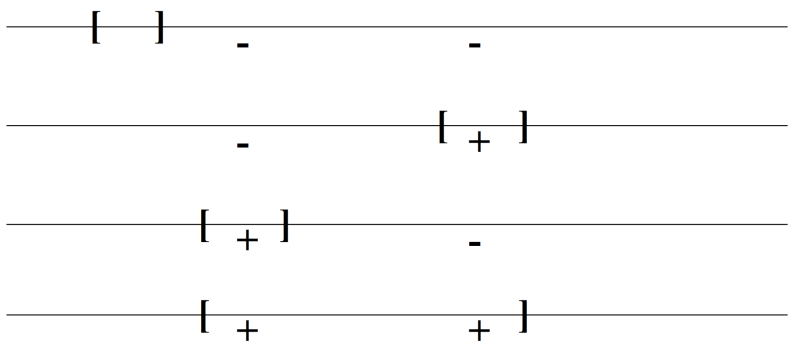


VC dimension (2 of 8)

Definition (VC dimension)

$VC(H)$ is the size of the largest set S shattered by H.

$H =$ closed intervals. Two points that can be shattered:



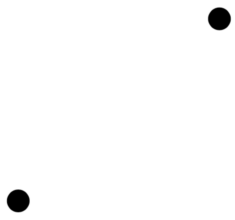
but no set of three points can be shattered (and thus $VC(H) = 2$):

VC dimension (3 of 8)

- Suppose $\mathbf{x} \in \mathbb{R}^2$. Consider H = set of linearly separable hypotheses:

$$h_{\mathbf{w}}(\mathbf{x}) = \begin{cases} + & f_{\mathbf{w}}(\mathbf{x}) \geq 0 \\ - & \text{o.w.} \end{cases}$$

- Can linear discriminant H shatter two points?



yes

VC dimension (4 of 8)

- Can linear discriminant H shatter three points in $\mathbb{R}^2$?

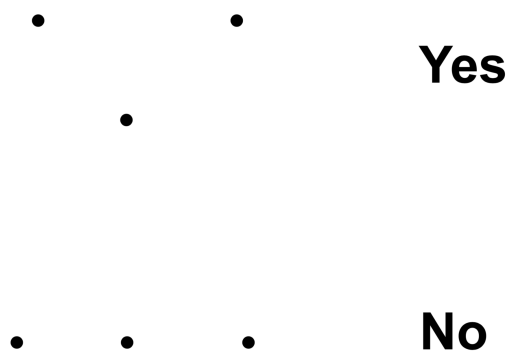
- Depends. Sometimes yes, sometimes no. What does this say about $VC(H)$?

VC dimension (5 of 8)

Definition (VC dimension)

$VC(H)$ is the size of the largest set S shattered by H.

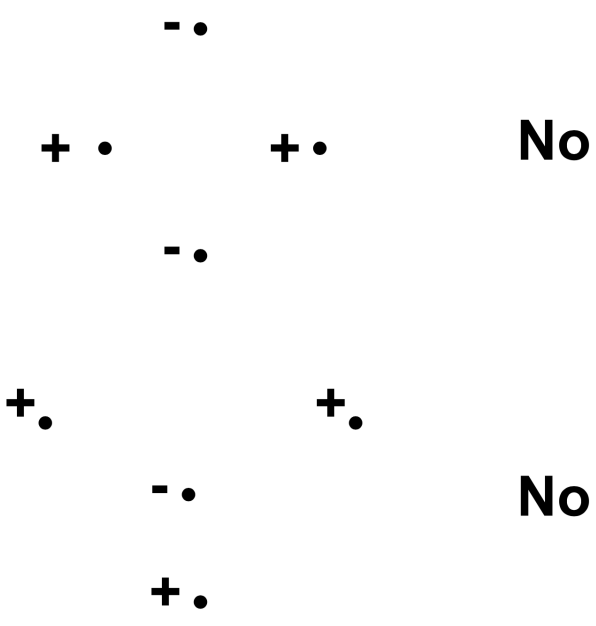
Can linear discriminant H shatter three points in $\mathbb{R}^2$?



Conclude that $VC(H)$ is at least 3.

VC dimension (6 of 8)

Can linear discriminant H shatter four points in $\mathbb{R}^2$?



This covers all cases. **Therefore,** $VC(H) = 3$.

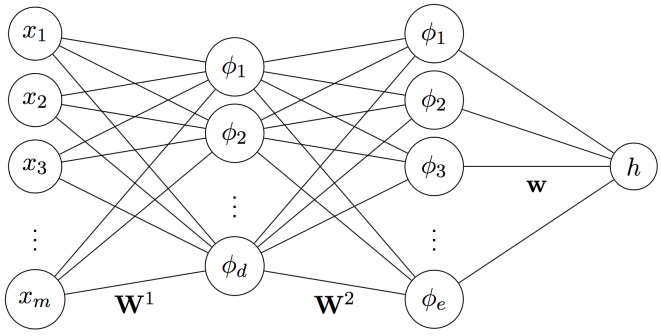
VC dimension (7 of 8)

- More generally, consider $X = \mathbb{R}^m$
- H: set of linearly separable hypotheses

$$h_{\mathbf{w}}(\mathbf{x}) = \begin{cases} 1 & f_{\mathbf{w}}(\mathbf{x}) \geq 0 \\ 0 & \text{o.w.} \end{cases}$$

- Can show that $VC(H) = m + 1$.

VC dimension (8 of 8)



- What about a neural network with N non-input nodes? Have:

$$VC(H) \leq 2kN \log_2(eN),$$

  where e is base of natural log, and VC dim. nodes $\leq k$.
- With linear-threshold activations and $d = e = m$, we have $k = m + 1$,

$$VC(H) \leq 2(m+1)(2m+1) \log_2(e(2m+1)) = O(m^2 \ln m)$$

- Since number of weights w is $O(m^2)$, this is $VC(H) = O(w \ln w)$.
- Networks with sigmoidal activations are less well understood.

Contents

Conclusion

- Sample complexity: how many examples are needed such that every consistent hypothesis will be ϵ-accurate with probability $\geq 1 - \delta$
- PAC-learnability: can this be done in polynomial time?
- VC-dimension: a measure of the effective dimension of an infinite hypothesis space, can be used to get a bound on sample complexity
- (Entire graduate class (CS228) on computational learning theory!)