Model Theory and Machine Learning

Model Theory and Mathematical Logic

In Honor of Chris Laskowski's 60th Birthday

David Marker

Mathematics, Statistics, and Computer Science
University of Illinois at Chicago

June 22, 2019

HAPPY BIRTHDAY CHRIS!

PAC Learning

Let X be a set and μ a probability measure on X.

A *concept class* $\mathcal{C}$ is a subset of 2^X, though we sometimes think of $\mathcal{C} \subseteq \mathcal{P}(X)$.

We try to learn a concept $c \in \mathcal{C}$ in the following manner.
• Using the distribution μ we choose $x_1, \ldots, x_m$ a sequence of i.i.d. samples from X. Our learning procedure gets input

$$((x_1, c(x_1)), \ldots, (x_m, c(x_m)))$$

a sequence of *test data*.
• Our procedure then produces $h \in \mathcal{C}$.

Our goal is to minimize the error given by

$$\mu(\{x \in X : h(x) \neq c(x)\}.$$

PAC Learning

Definition

Let $\mathcal{C}$ be a concept class on X.
We say that a learning procedure P is *probably approximately correct (PAC)* if for any $\epsilon > 0$ and $\delta > 0$ there is a natural number $m = m(\epsilon, \delta)$ such that for any probability distribution μ on X and any concept $c \in \mathcal{C}$ if we take an i.i.d. sample $x_1, \ldots, x_m$ and test data

$$\sigma = ((x_1, c(x_1), \ldots, (x_m, c(x_m))),$$

and P outputs h, then

$$Pr(\mu(\{x \in X : h(x) \neq c(x)\}) < \epsilon) > 1 - \delta.$$

In other words, given ϵ and δ there is m such that for any probability distribution μ, with high probability the error set is small. It is important to note that the procedure and the choice of m are independent of the probability measure.

Example Learning Rectangles

Let $X = \mathbb{R}^2$ and let $\mathcal{C} = \{[a, b] \times [c, d] : a \leq b, c \leq d\}$. Try to learn $\widehat{R} \in \mathcal{C}$.

Procedure: • test date: random sample S of m points, $S_0 = S \setminus \widehat{R}$, $S_1 = S \cap \widehat{R}$.

• output R the smallest rectangle with $S_1 \subseteq R$.

Let μ be a continuous probability distribution on $\mathbb{R}^2$. Let $\epsilon, \delta > 0$.

Learning Rectangles

Let μ be a continuous probability distribution on $\mathbb{R}^2$. Let $\epsilon, \delta > 0$.
Let B_1 be the smallest rectangle with the same top edge as $\widehat{R}$ with $\mu(B_1) = \epsilon/4$.
Similarly, define B_2, B_3, B_4 on the bottom, right and left.
If $\mu(\widehat{R} \setminus R) > \epsilon$, then some $S \cap B_i = \emptyset$.

$$Pr(S \cap B_i = \emptyset) = (1 - \frac{\epsilon}{4})^m.$$

$$Pr(\mu(\widehat{R} \setminus R) \geq \epsilon) \leq 4(1 - \frac{\epsilon}{4})^m \leq 4e^{-\frac{\epsilon m}{4}}$$

If

$$m \geq \frac{4}{\epsilon} \ln\left(\frac{4}{\delta}\right) \text{ then } Pr(\mu(\widehat{R} \setminus R) \geq \epsilon) < \delta.$$

Note that m does not depend on μ and can be chosen linear in $1/\epsilon$ and in $\ln(1/\delta)$.

VC dimension

We say that $\mathcal{C}$ *shatters* $A \subseteq X$ if

$$\{C \cap A : C \in \mathcal{C}\} = \mathcal{P}(A).$$

Definition

If there is a largest integer d such that $\mathcal{C}$ shatters some set of size d, then we say $d = \mathrm{VCdim}(\mathcal{C})$ is the *VC-dimension* of $\mathcal{C}$.
If $\mathcal{C}$ shatters arbitrarily large finite sets, then we say $\mathrm{VCdim}(\mathcal{C}) = \infty$.

Examples

1) Let $X = \mathbb{R}$. For $a \in \mathbb{R}$ let $C_a = \{x : x \geq a\}$ and $\mathcal{C} = \{C_a : a \in \mathbb{R}\}$, then $\mathrm{VCdim}(\mathcal{C}) = 1$. Suppose $x < y$, then we can not shatter $\{x, y\}$.

2) Let $X = \mathbb{R}^2$ and let $\mathcal{C}$ be the collection of axis-parallel rectangles. Then $\mathrm{VCdim}(\mathcal{C}) = 4$.
It is easy to shatter $\{(0,2),(1,0),(2,3),(3,1)\}$. But no set of size 5 can be shattered. Suppose we have 5 points. Choose 4 points contain ones with maximal and minimal first and second coordinates. Then any axis-parallel rectangle contain those four points contains all five.

3) Let X be the vertices of a random graph. For $a \in X$ let $C_a = \{x : (x, a) \in E\}$ and $\mathcal{C} = \{C_a : a \in X\}$. Then C shatters any finite $A \subset X$ so $\mathrm{VCdim}(A) = \infty$.

VC-dimension and PAC Learning

Lemma

If $\mathrm{VCdim}(\mathcal{C}) = \infty$, then there is no PAC learning procedure for $\mathcal{C}$.

Remarkably, finite VC-dimension is the only constraint on PAC-learnability.

Theorem (Fundamental Theorem of PAC Learning–Valliant)

Let $\mathcal{C}$ be a set system on X with $\mathrm{VCdim}(\mathcal{C}) = d$. Then there is a PAC learning procedure for $\mathcal{C}$.
Indeed, there is a constant k such that for all $\epsilon > 0$ and $\delta > 0$ there is

$$m(\epsilon, \delta) \leq k \frac{d \log(1/\epsilon) + \log(1/\delta)}{\epsilon}$$

Model Theoretic Concept Classes

Definable Families of Definable Sets
Let $\mathcal{M}$ be an $\mathcal{L}$-structure and $\phi(x_1, \ldots, x_n, y_1, \ldots, y_m)$ be a $\mathcal{L}$-formula. For $\overline{b} \in M^m$ let $C_{\overline{b}} = \{\overline{a} \in M^n : \mathcal{M} \models \phi(\overline{a}, \overline{b})\}$ and let $\mathcal{C}_\phi = \{C_{\overline{b}} : \overline{b} \in M^m\}$.

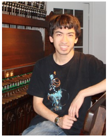
Recall $\phi(\overline{x}, \overline{y})$ has the *independence property* if and only if there are $\overline{a}_0, \overline{a}_1, \ldots$ and $(\overline{b}_A : A \subseteq \omega)$ such that

$$\phi(\overline{a}_i, \overline{b}_A) \Leftrightarrow i \in A.$$

Observation [Laskowski] $\mathcal{C}_\phi$ has infinite VC-dimension if and only if ϕ has the independence property.

For example, any definable family in an o-minimal structure has finite VC-dimension and hence is PAC learnable. The same is true for stable structures, Presburger Arithmetic, the p-adics, algebraically closed valued fields...

On-line Learning

The remaining work I'm talking about today is due to Hunter Chase and James Freitag.

We will look at a second model of machine learning, called *on-line learning*. Once again we have a set X and a concept class $\mathcal{C} \subset 2^X$.

We try to learn $c \in \mathcal{C}$ in the following manner.

For $i = 0, \ldots, M$

We are given x_i;

We choose p_i our guess about $c(x_i)$;

We are told $c(x_i)$;

Go to the next i.

On-line Learning

An *on-line learning procedure* takes $(x_1, c(x_1)), \dots (x_m, c(x_m))$ and x_{m+1} as input, outputs p_{m+1} our guess about $c(x_{m+1})$

The number of mistakes made is $|\{i : p_i \neq c(x_i)\}|$. Our goal is to minimize the number of mistakes.

It's sometimes useful to think of this as a game played against an adversary who gives us the $x_0, \dots, x_M$ and at the end must be able to show there is $c \in \mathcal{C}$ consistent with the answers given.

Definition

We say that $\mathcal{C}$ is *on-line learnable* if there is a learning procedure and an absolute bound B such that for any concept $c \in \mathcal{C}$, any M and any $x_0, \dots, x_M$, the procedure will make at most B errors.

Examples

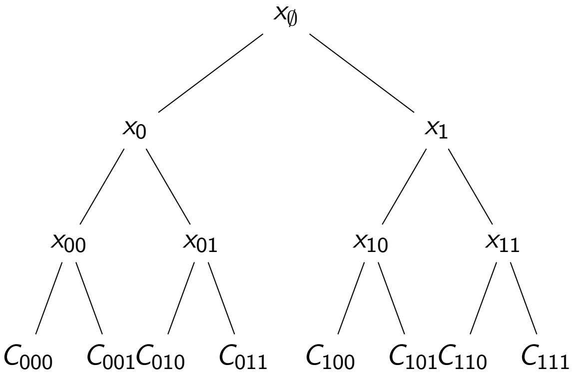
1) Let E be an equivalence relation and let $\mathcal{C}$ be the collection of equivalence classes.
E is on-line learnable. Given x_1 guess no. As long as you are correct keep guessing no. If we are ever wrong we now know x in the equivalence class. In all future rounds we will answer correctly.
This procedure makes at most one mistake.

2) Let $X = \mathbb{R}$ and let $\mathcal{C}$ be the collection of all intervals $(-\infty, a)$.
We claim that for any learning procedure the adversary can choose a sequence where we make a mistake in each round.

Let $x_0 = 1$. If the learner guesses yes at stage yes, let $x_{n+1} = x_n - \frac{1}{2^n}$, while if the learner guesses no, let $x_{n+1} = x_n + \frac{1}{2^n}$. At the end of T stages, the adversary can produce $\widehat{x}$ such that the learner has been wrong at every stage.

labeled trees

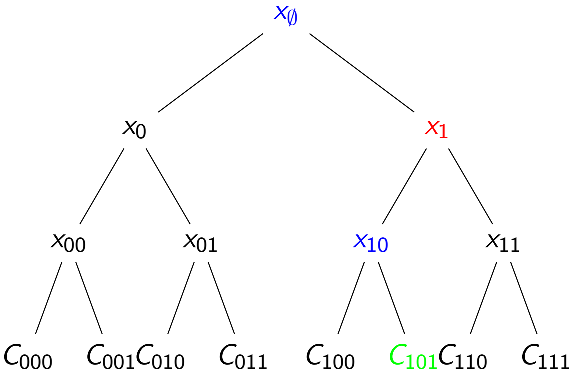
Consider a tree $T \subseteq 2^{\leq n}$ such that every node is either terminal or has two successors. We label T such that every non-terminal node σ is labeled with $C_\sigma \in \mathcal{C}$ and every non-terminal node τ is labeled with $x_\tau \in X$.



well-labeled trees

We say T is *well-labeled* if for all terminal nodes σ and all $I < |\sigma|$

$$x_{\sigma|i} \in C_\sigma \leftrightarrow \sigma(i) = 1$$



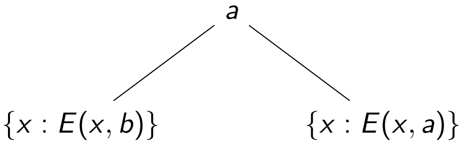
For example $x_\emptyset, x_{10} \in C_{101}, x_1 \notin C_{101}$.

Littlestone Dimension

Definition

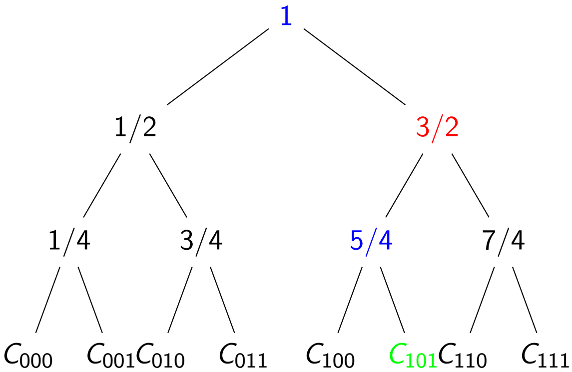
The *Littlestone Dimension* of $\mathcal{C}$, $\mathrm{Ldim}(\mathcal{C})$ is the largest n such that the full binary tree $2^{\leq n}$ can be well labeled. If there is no largest n, then $\mathrm{Ldim}(\mathcal{C}) = \infty$.

Example 1) If E is an equivalence relation on X and $\mathcal{C}$ is the set of equivalence relations then $\mathrm{Ldim}(\mathcal{C}) = 1$. Let $\neg E(a, b)$.



Littlestone Dimension

Example 2 For $a \in \mathbb{Q}$ let $C_a = \{x : x > a\}$ and $\mathcal{C} = \{C_a : a \in \mathbb{Q}\}$. Let $\mathrm{Ldim}(\mathcal{C}) = \infty$. For example



For example we could take $C_{101} = \{x : x > 11/8\}$.

On-Line learning and Littlestone Dimension

Theorem (Littlestone)

There is an on-line learning procedure for $\mathcal{C}$ if and only if $\mathrm{Ldim}(\mathcal{C}) < \infty$. Moreover, if $\mathrm{Ldim}(\mathcal{C}) = k$, there is an on-line procedure learning $\mathcal{C}$ making at most k errors.

($\Rightarrow$) If $\mathrm{Ldim}(\mathcal{C}) = \infty$, given M choose a well-labeled full binary tree of height M. An adversary can consistently tell you that you are wrong in each move and force M errors.

Standard Optimization Algorithm

```
Let $\mathcal{C}_0 = \mathcal{C}$;
For $i = 0, \ldots, M$;
Given $\mathcal{C}_i$ and $x_i$, let $\mathcal{C}_i^j = \{c \in \mathcal{C}_i : c(x_i) = j\}$.;
Choose $j$ such that $\mathrm{Ldim}(\mathcal{C}_i^j)$ is maximal and let $p_i = j$;
Let $C_{i+1} = C_i^{c(x_i)}$;
Next i.
```

Lemma

Suppose $\mathrm{Ldim}(\mathcal{C}) = d$ *and* $a \in X$. *Let* $\mathcal{C}^i = \{c \in \mathcal{C} : c(a) = i\}$ *for* $i = 0, 1$. *Then at most one of* $\mathcal{C}^0$ *and* C^1 *has Littlestone dimension* d.

Each time the algorithm makes an error the Littlestone dimension goes down. Thus we can make at most $\mathrm{Ldim}(\mathcal{C})$-errors.

Model Theory and On-line Learning

Let $\mathcal{M}$ be an $\mathcal{L}$-structure, $\phi(\overline{x}, \overline{y})$ an $\mathcal{L}$-formula and $\mathcal{C}_\phi = \{\{\overline{a} : \phi(\overline{a}, \overline{b})\} : \overline{b} \in M^m\}$.

Observation There is an on-line learning procedure for $\mathcal{C}_\phi$ if and only if ϕ is stable.

$$\mathrm{Ldim}(C_\phi) = \infty \Leftrightarrow \phi^{\mathrm{opp}} \text{ has the binary tree property } \Leftrightarrow \phi \text{ is unstable.}$$

Littlestone dimension $=$ Shelah's 2-rank.

Thus there are on-line learning procedures for definable families in algebraically closed fields, differentially closed fields, separably closed fields, modules, non-abelian free groups....
Few examples of infinite on-line learnable classes were previously known.

Query Learning

We look at a third model of learning introduced by Angluin.
In this model we have a set X a concept class $\mathcal{C}$ and a hypothesis class $\mathcal{H}$ with $\mathcal{C} \subseteq \mathcal{H} \subseteq \mathcal{P}(X)$.
We are trying to learn $c \in \mathcal{C}$. At each stage s:

- we make an *equivalence query* guessing $h_s \in \mathcal{H}$;
- either we succeed if $h_s = c$ or else we are given x_s where $h_s(x_s) \neq c(x_s)$.

We say that $\mathcal{C}$ is *learnable with equivalence queries* from $\mathcal{H}$, if there is a number n and a procedure that will always succeed in at most n-steps. The least such n is $LC^{EQ}(\mathcal{C}, \mathcal{H})$. Otherwise $LC^{EQ}(\mathcal{C}) = \infty$.
$LC^{EQ}(\mathcal{C}, \mathcal{H})$ is the *learning complexity* of $\mathcal{C}$ from $\mathcal{H}$.

Taking $\mathcal{C} = \mathcal{H}$ makes learning very difficult. Let X be an infinite set and $\mathcal{C} = \{\{x\} : x \in X\}$. If we only allowed to make equivalence queries from $\mathcal{C}$ an adversary could keep us from learning C by always returning x as a counterexample when we guess $\{x\}$.

On the other hand if $\mathcal{C} \subset \mathcal{H}$ and $\emptyset \in \mathcal{H}$. We can learn $\{x\}$ in one step by submitting $\emptyset$ as a query.

LC^{EQ} and Littlestone dimension

Lemma

If $\mathrm{Ldim}(\mathcal{C}) \geq d$*, then* $LC^{EQ}(\mathcal{C}, \mathcal{H}) \geq d + 1$*.*

Proof.

We can use a well labeled tree on $2^{\leq d}$ to force $d + 1$ rounds. □

Corollary

If $\mathcal{C}$ *is learnable with equivalence queries from* $\mathcal{H}$*, then* $\mathrm{Ldim}(\mathcal{C})$ *is finite.*

LC^{EQ} and Littlestone dimension when $\mathcal{H} = \mathcal{P}(\mathcal{C})$

Lemma

If $\mathrm{Ldim}(\mathcal{C}) = d$*, then* $LC^{EQ}(\mathcal{C}, \mathcal{P}(\mathcal{C})) \leq d + 1$.

Let $\mathcal{C}_0 = \mathcal{C}$.
Let $\mathcal{C}_i^{(x,j)} = \{c \in \mathcal{C}_i : c(x) = j\}$ for $x \in X$, $j = 0, 1$.
Let $B_i = \{x : \mathrm{Ldim}(\mathcal{C}_i^{x,1}) > \mathrm{Ldim}(\mathcal{C}_i^{(x,0)})\})$.
Submit B_i as a hypothesis. If we receive a counterexample x, let $\mathcal{C}_{i+1} = \{c \in \mathcal{C}_i : c(x) \neq \chi_{B_i}(x)\}$. Then $\mathrm{Ldim}(\mathcal{C}_{i_1}) < \mathrm{Ldim}(\mathcal{C}_i)$.

Consistency Dimension

We say that $f : X \to 2$ is *n-consistent* with $\mathcal{C}$ if for every $A \subseteq X$ with $|A| = n$, there is $c \in \mathcal{C}$ such that $f|A \subseteq c$.

We say $\mathcal{C}$ has *consistency dimension* n with respect to $\mathcal{H}$ if n is least such that whenever $f \in 2^X$ is n-consistent with $\mathcal{C}$, then $f \in \mathcal{H}$ and we let $C(\mathcal{C}, \mathcal{H}) = n$.
If no such n exists, then $C(\mathcal{C}, \mathcal{H}) = \infty$.

Lemma

If $C(\mathcal{C}, \mathcal{H}) > n$, then $LC^{EQ}(\mathcal{C}, \mathcal{H}) > n$

Suppose f is n-consistent, but $f \notin \mathcal{H}$. Suppose we make queries $h_1, h_2, \ldots, h_n$. Our adversary could return $x_1, \ldots, x_n$ with $h_i(x_i) \neq f(x_i)$ but $f|\{x_1, \ldots, x_n\}$ has an extension in $\mathcal{C}$.

Consistency Dimension and Query Learning

Theorem (Chase–Freitag)

$\mathcal{C}$ is learnable with queries from $\mathcal{H}$ if and only if $\mathrm{Ldim}(\mathcal{C}) < \infty$ and $CD(\mathcal{C}, \mathcal{H}) < \infty$.
If $\mathrm{Ldim}(\mathcal{C}) = d$ and $C(\mathcal{C}, \mathcal{H}) = n$, then $LC^{EQ}(\mathcal{C}, \mathcal{H}) \leq n^d$.

Theorem (Chase–Freitag)

If $\mathrm{Ldim}(\mathcal{C}) < \infty$, there is $\mathcal{H}$ with $\mathrm{Ldim}(\mathcal{C}) = \mathrm{Ldim}(\mathcal{H})$ and $C(\mathcal{C}, \mathcal{H}) \leq \mathrm{Ldim}(\mathcal{C}) + 1$.

Finite Cover Property

Definition

$\psi(\overline{x}, \overline{y})$ has the *finite cover property* (FCP) if for every n there is a $p \subseteq \{\psi(\overline{x}, \overline{a}), \neg\psi(\overline{x}, \overline{a}) : \overline{a} \in M\}$ such that every n element subset of p is consistent but p is inconsistent.
Otherwise ψ is NFCP.

Example Let $\mathcal{M}$ be a structure where there is a unique equivalence class of each finite size.
Let $\psi(x, y)$ be $xEy \wedge x \neq y$.

Let $a_1, \ldots, a_n$ list an equivalence class of size n and let p be $\{\psi(x, a_1), \ldots, \psi(x, a_n)\}$.
Then p is $n-1$ consistent but not consistent. Thus ψ is FCP.

Externally Definable Sets

Let $\phi(x_1, \ldots, x_m, y_1, \ldots, y_n)$ be an $\mathcal{L}$-formula and let $\mathcal{C}_\phi$ be the collection of $\{\phi(\mathcal{M}, \overline{b}) : \overline{b} \in M^n\}$.
Let $\mathcal{N}$ be a $|M|^+$-saturated elementary extension of $\mathcal{M}$ and let
$\mathcal{H}_\phi = \{\phi(\mathcal{N}, \overline{b}) \cap M^m : \overline{b} \in N^n\}$.
We call $\mathcal{H}_\phi$ the subsets of M^m *externally definable* by ϕ.

Littlestone dimension is an elementary property thus
$\mathrm{Ldim}(\mathcal{H}_\phi) = \mathrm{Ldim}(\mathcal{C}_\phi)$.

$CD(\mathcal{C}_\phi, \mathcal{H}_\phi) < \infty \Leftrightarrow \phi^{\mathrm{opp}}$ has NFCP

Thus $\mathcal{C}_\phi$ is learnable with queries from $\mathcal{H}_\phi$ if and only if ϕ is stable and ϕ^{opp} is NFCP.
For example, in ACF or DCF we can learn definable families using the corresponding family of externally definable sets.

Membership Queries

We can expand the query learning model by allowing the learner to also make membership queries, i.e., at any stage the learner can ask $x \in C$? for any $x \in X$.

Theorem (Chase–Freitag)

$LC^{EQ+MQ}(\mathcal{C},\mathcal{H}) < \infty$ if and only if $\mathrm{Ldim}(\mathcal{C}) < \infty$ and $C(\mathcal{C},\mathcal{H}) < \infty$.

In this case we can bound $LC^{EQ+MQ}(\mathcal{C},\mathcal{H})$ by $\mathrm{Ldim}(\mathcal{C}) \cdot C(\mathcal{C},\mathcal{H})$ (roughly).

UDTFS

In a completely different direction,

Theorem (Eshel–Kaplan)

The following are equivalent:

- ϕ *is NIP in any completion of* T*.*
- ϕ *has Uniform Definability of Types over Finite sets in* T*.*

While this is a purely model theoretic result, the proof relies on two results from machine learning theory.

References

• Hunter Chase and James Freitag, Model Theory and Machine Learning, arxiv.

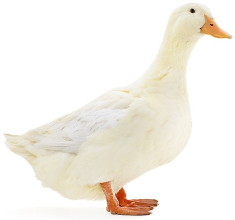
• Hunter Chase and James Freitag, Bounds on Query Learning, arxiv.

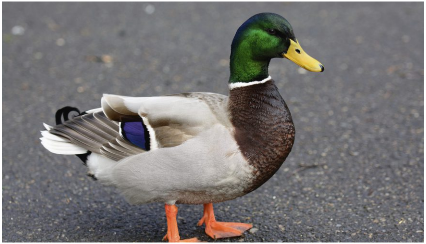
• Shlomo Eschel and Itay Kaplan, On Uniform definability of types over finite sets for NIP formulas, arxiv.

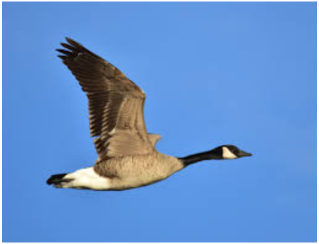
• Michael C. Laskowski, Vapnik-Chervonenkis classes of definable sets, Journal of the London Mathematical Society, 1992.

• Shai Shalev-Shwartz and Shai Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, 2014.

HAPPY BIRTHDAY CHRIS!







Thank You!