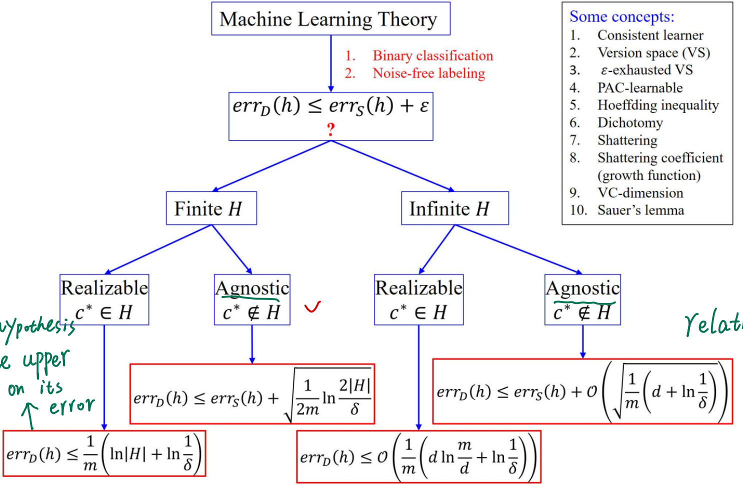
Discussions on Learning Theory

April 20, 2020

Big Picture of Today's Discussion

We recap some important concepts in the learning theory, which has been fully discussed in our classes.



learner's hypothesis

absolute upper bound on its error

relative

What General Laws Constrain Inductive Learning

Generalizabitity of learning

In machine learning it's really $\boxed{\text{generalization error}}$ that we care about, but most learning algorithms fit their models to the training set.

- Sample Complexity
  - How many training examples are sufficient to learn target concept?
- Computational Complexity
  - Resources required to learn target concept? Running time
- We want theory to relate
  - Training examples.
    - Quantity m
    - Quality
    - How presented
  - Complexity of hypothesis space. $\mathcal{H}$
  - Accuracy to which target function is approximated. ϵ
  - Probability of successful learning. δ

Problem Setting for Learning from Data

$\mathcal{X}$ $\mathcal{D}$

Given:

n pairs

- **Training data** $S = ((x_1, y_1), \dots, (x_n, y_n)) \subset \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are termed input space and output space, respectively.
- Set of **hypothesis** $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}$. $\{0, 1\}$
- Set of possible target functions $C = \{c : \mathcal{X} \to \mathcal{Y}\}$. labeling Function
- Noise-free label $c(x)$.

Goal: Learner must output a hypothesis $h \in \mathcal{H}$ estimating c such that

$$h = \arg\min_{h \in \mathcal{H}} \boxed{\mathrm{error}_{train}(h)}.$$

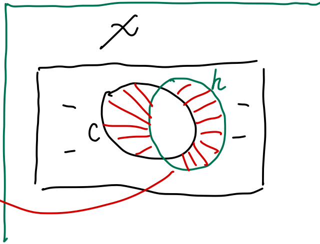
- learner is trying to figure out the unknown labeling function.

with zero error on S
whenever $\mathcal{H}$ contains such h.

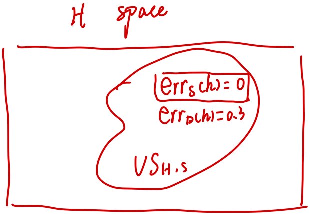
- Consistent Learner
  - outputs hypothesis h that $\boxed{\text{perfectly fits the training data } S,}$

$$h(x) = c^*(x), \qquad \forall x \in S.$$

- $\text{error}_{\text{train}}(h) = \Pr_{x \in \underline{S}} [h(x) \neq c(x)]$
- $\text{error}_{\text{true}}(h) = \Pr_{x \in \underline{D}} [h(x) \neq c(x)]$



Version Space

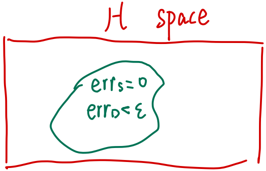


- Version Space (VS)
  - set of all hypotheses $h \in H$ that correctly classify the training data S,

$$VS_{H,S} = \{h \in H | \forall x \in S, \boxed{h(x)} = c^*(x)\}.$$

- A consistent learner must produce a hypothesis in the $VS_{H,S}$ for H given S.

ϵ-exhausted Version Space



Definition

The version space $\mathrm{VS}_{\mathrm{H,D}}$ with respect to training data D is said to be **ϵ-exhausted** if every hypothesis h in $\mathrm{VS}_{\mathrm{H,D}}$ has true error less than ϵ.

$$(\forall h \in \mathrm{VS}_{\mathrm{H,D}})\ \mathrm{error}_{\mathrm{true}}(h) < \epsilon.$$

ϵ-exhausted Version Space

· One can never be sure that $VS_{H,S}$ is ϵ-exhausted.
but one can bound the probabity that it is not.

How many examples will ϵ-exhaust the VS?

Theorem: [Haussler, 1988].

If the hypothesis space H is finite, and D is a sequence of $m \geq 1$ independent random examples of some target concept c, then for any $0 \leq \epsilon \leq 1$, the probability that the version space with respect to H and D is not ϵ-exhausted (with respect to c) is less than

$$|H|e^{-\epsilon m}$$

Interesting! This bounds the probability that any consistent learner will output a hypothesis h with $error(h) \geq \epsilon$

Any(!) learner that outputs a hypothesis consistent with all training examples (i.e., an h contained in $VS_{H,D}$)

PAC-learnable (Probably Approximately Correct)

PAC – Framework.

Formal Def.

Consider a class C of possible target concepts defined over a set of instances X of length n, and a learner L using hypothesis space H.

> *Definition:* C is **PAC-learnable** by L using H if for all $c \in C$, distributions $\mathcal{D}$ over X, ϵ such that $0 < \epsilon < 1/2$, and δ such that $0 < \delta < 1/2$,
>
> learner L will with probability at least $(1 - \delta)$ output a hypothesis $h \in H$ such that $\boxed{error_{\mathcal{D}}(h) \leq \epsilon}$ in time that is polynomial in $1/\epsilon$, $1/\delta$, n and $size(c)$.

Sufficient condition:

Holds if learner L requires only a polynomial number of training examples, and processing per example is polynomial

a feasonable amount of Computation!

- With high probability learns a close approximation to the C^*.
- ϵ, δ, at least $(1-\delta)$, a learer can learn a concept with error at most ϵ

HW3.

Hoeffding bounds

Consider coin of bias p flipped m times.
Let N be the observed # heads. Let $\varepsilon \in [0,1]$.

Hoeffding bounds:

- $\Pr[N/m > p + \varepsilon] \leq e^{-2m\varepsilon^2}$, and
- $\Pr[N/m < p - \varepsilon] \leq e^{-2m\varepsilon^2}$.

$\Pr\left[\left|\frac{N}{m} - p\right| > \varepsilon\right] \leq 2e^{-2m\varepsilon^2}$.

Exponentially decreasing tails

- Tail inequality: bound probability mass in tail of distribution (how concentrated is a random variable around its expectation).

· Remark.

if we use $\frac{N}{m}$, $\underbrace{\boxed{\text{\# heads}}}_{\text{r.v.}}$ Emperical average, as estimate of p, then the probability of our being far from the true value is small with sufficiently large m.

Dichotomy

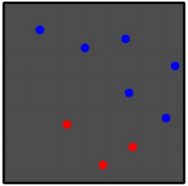
$h : S \to \{+1, -1\}$

Given a dataset $S=\{x_1,\ldots,x_m\}$,

$(h(x_1),\ldots,h(x_m))$ tuple

A dichotomy of S

1. *If H is diverse, we get many different dichotomies.*
2. *If H contains many similar function, we only get a few dichotomies.*



dichotomy

- $\mathcal{H}[S]$ 是在 $\underbrace{S=\{x_1,\cdots,x_m\}}_{\text{specified set.}}$ 所有 dichotomy 的一个集合.

$$\mathcal{H}(S) = \{(h(x_1),\cdots h(x_m)) : h \in \mathcal{H}\}.$$

Shattering Coefficient (Growth Function)

- H[m] - max number of ways to $\boxed{\text{split m points}}$ using concepts in H

$$\underbrace{H[m]}_{} = \max_{|S|=m} |H[S]| \leq 2^m.$$

For arbitrary S. 在 m 个 points 上能产生的最多的 dichotomy 的数量.

H 能在 S 上产生所有的 dichotomy.

Shattering, VC-dimension

Definition: H shatters S if $|\mathrm{H}[\mathrm{S}]| = 2^{|S|}$.

A set of points S is shattered by H is there are hypotheses in H that split S in all of the $2^{|S|}$ possible ways, all possible ways of classifying points in S are achievable using concepts in H.

Definition: VC-dimension (Vapnik-Chervonenkis dimension)

The VC-dimension of a hypothesis space H is the cardinality of the largest set S that can be shattered by H.

If arbitrarily large finite sets can be shattered by H, then $\mathrm{VCdim}(\mathrm{H}) = \infty$

$\mathrm{VC\,dim}(\mathcal{H}) = \max\{ |S| : S \text{ shattered by } \mathcal{H} \}$.

$\mathrm{VC\,dim}(h) = d$

① Lower bound: Show there exists a set of d points that can be shattered.

② upper bound: show no set of $d+1$ points that can be shattered

$\max_{|S|=m} |H[S]|$

Sauer's Lemma:

Let d = VCdim(H) $< \infty$

- $m \leq d$, then $\underline{H[m] = 2^m}$
- m>d, then $H[m] = O(m^d) \implies$ $\boxed{H[m] \leq \sum_{i=0}^{d} \binom{m}{i} = \Phi_d(m) = O(m^d)}$

- This Lemma States that H[m] grows exponentially fast for $m \leq d$.

But then only grows like a polynominal for $m > d$.

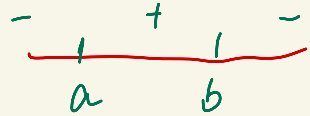
- VC dim(H) $< \infty$, $m \to \infty$, implies "Learnability".

e.g: The interval example.

$\mathcal{H}$ is the class of intervals on the $\mathbb{R}$

given $h = [a, b]$ with $a, b \in \mathbb{R}$, $a \leq b$

$$h(x) = \begin{cases} +1, & \text{if } a \leq x \leq b, \\ -1, & \text{otherwise} \end{cases}$$



$$\mathcal{H}(m) = \binom{m}{2} + \binom{m}{1} + 1 = \binom{m}{2} + m + 1$$

Sauer's lemma is tight!

Proof:

- At least 1 +1 follows by at least −1. $\binom{m}{2}$
- At least 1 +1 but no negative −1 followed the +1. $\binom{m}{1}$
- all the −1.

∎

$\mathcal{H}(m) \le \Phi_d(m)$

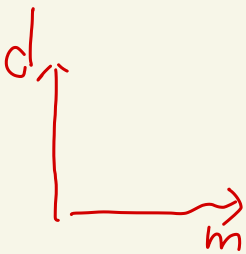
- Base case (for two variable).

- $m=0$, for any d. $\Phi_d(m) = \sum_{i=0}^{d} \binom{0}{i}$

  $\mathcal{H}(0) \le 1$. Since we can label 0 points at most 1 ways.

- $d=0$, for any m. $\Phi_d(m) = \sum_{i=0}^{0} \binom{m}{i} = 1$.

  $\mathcal{H}(m) = 1$ since VC dim($\mathcal{H}$) $= 0$ implies we lable everything with the same lable.



- Inductive step:

inclued the first item,

$$\boxed{\binom{m}{k} = \binom{m-1}{k} + \binom{m-1}{k-1}.}$$

$$\binom{m}{k} = 0 \text{ if } k<0 \text{ or } k>m$$

Assume lemma holds $\bar{m}+\bar{d} < m+d$.

Given $\underline{S=(x_1, \cdots, x_m)}$, we want to show $|\mathcal{H}[S]| \le \Phi_d(m)$.

key ideas:

Construct two new Hypothesis: $\mathcal{H}_1$ and $\mathcal{H}_2$.

$\mathcal{H}_1$ 和 $\mathcal{H}_2$ 在 $\boxed{S' = \{x_1, \cdots, x_{m-1}\}}$ 忽略了 $\underline{x_m}$

$\mathcal{H}_1$: 是 $\underline{\mathcal{H}}$ 的一个最小子集但能 $\underline{S'}$ shattering,

$\mathcal{H}_2$: 如果有 $\underline{\text{2个 hypothesis label}}$ S' 以相同的方式, 一个就在 $\mathcal{H}_1$ 里, 一个就在 $\mathcal{H}_2$ 里.

e.g:

$\mathcal{H}$

	x_1	x_2	x_3	x_4	x_5
h_1	0	1	1	0	0
h_2	0	1	1	0	1
h_3	0	1	1	1	0
h_4	1	0	0	1	0
h_5	1	0	0	1	1
h_6	1	0	0	0	1

$\mathcal{H}_1$

x_1	x_2	x_3	x_4
0	1	1	0
0	1	1	1
1	0	0	1
1	0	0	0

$\mathcal{H}_2$

x_1	x_2	x_3	x_4
0	1	1	0
1	0	0	1

- if a set is shattered by $\mathcal{H}_1$, then it is also shattered by $\mathcal{H}$. Because can generate $\mathcal{H}$ by using the same as x_i's when we generate $\mathcal{H}_1$. Thus,

$$\text{VC dim}(\mathcal{H}_1) \le \text{VC dim}(\mathcal{H}) = d.$$

· if a set T is shattered by $\mathcal{H}_2$. then

$T \cup \{x_m\}$ is shattered by $\mathcal{H}$.

Because. $\mathcal{H}$ 中会有两个对应的 Hypothesis

和 $\mathcal{H}_2$ 有着相同的 element. 通过加上 $x_m=1$

and $x_m=0$. Thus,

$$VC\dim(\mathcal{H}) \geq VC\dim(\mathcal{H}_2)+1$$

Now, by inductive,

$$\begin{aligned}
|\mathcal{H}(S)| &= |\mathcal{H}_1(S')| + |\mathcal{H}_2(S')| \\
&\leq \sum_{i=0}^{d}\binom{m-1}{i} + \sum_{i=0}^{d-1}\binom{m-1}{i} \\
&= \sum_{i=0}^{d}\binom{m-1}{i} + \boxed{\sum_{i=0}^{d}\binom{m-1}{i-1}} \\
&= \sum_{i=0}^{d}\binom{m}{i} = \Phi_d(m)
\end{aligned}$$

$\blacksquare$