Computational Learning Theory

TM Chapter 7

Outline

- Computational learning theory
- Setting 1: Active learner selects input instances to query teacher
- Setting 2: Teacher selects training examples for learner
- Setting 3: Randomly generated training instances to be labeled by teacher
- Probably approximately correct (PAC) learning
- Vapnik-Chervonenkis (VC) Dimension

Why Study Computational Learning Theory?

What general laws govern/constrain inductive learning?

Computational learning theory aims to relate

- Probability of successful learning
- Number of training examples
- Complexity/size of hypothesis space
- Quality of approximating target concept
- Manner in which training examples are presented

Concept Learning for *EnjoySport*

Given

- Instance space X: Each instance $x \in X$ is represented by input attributes: *Sky*, *AirTemp*, *Humidity*, *Wind*, *Water*, *Forecast*
- Hypothesis space H: Each hypothesis $h \in H$ ($h : X \to \{0, 1\}$) is represented by a conjunction of constraints on input attributes (e.g., $\langle Sunny, ?, ?, Strong, ?, Same \rangle$)
- Unknown target concept/function *EnjoySport*: $c : X \to \{0, 1\}$
- Noise-free training examples D of the form $\langle x, c(x) \rangle$: +ve and −ve training examples of the target concept c

Determine a hypothesis $h \in H$ that is consistent with D

Determine a hypothesis $h \in H$ that is consistent with $\{\langle x, c(x) \rangle\}_{x \in X}$?

Sample Complexity

How many training examples suffice to learn the target concept c?

1. Active learner repeatedly selects input instance x to query a teacher for $c(x)$

2. Teacher (who knows c) selects training examples $\langle x, c(x) \rangle$ for learner

3. Some random process (e.g., nature) repeatedly generates input instance x to query a teacher for $c(x)$

Sample Complexity: Setting 1

Active learner repeatedly selects input instance x to query a teacher for $c(x)$ (assume c is in learner's H)

Optimal query strategy?

- Select input instance x that satisfies exactly half of hypotheses in version space (if possible)
- Version space reduces by half with each training example, hence requiring at least $\lceil \log_2(VS_{H,D}) \rceil$ examples to find target concept c

Sample Complexity: Setting 2

Teacher (who knows c) selects training examples $\langle x, c(x) \rangle$ for learner (assume c is in learner's H)

Optimal teaching strategy? Depends on H used by learner

- Consider H = conjunctions of up to n Boolean literals and their negations

- How many training examples suffice to learn c?

Sample Complexity: Setting 3

Given

- Set X of input instances
- Set H of hypotheses
- Set C of possible target concepts/functions
- Training instances randomly generated by a fixed, unknown probability distribution Q over X

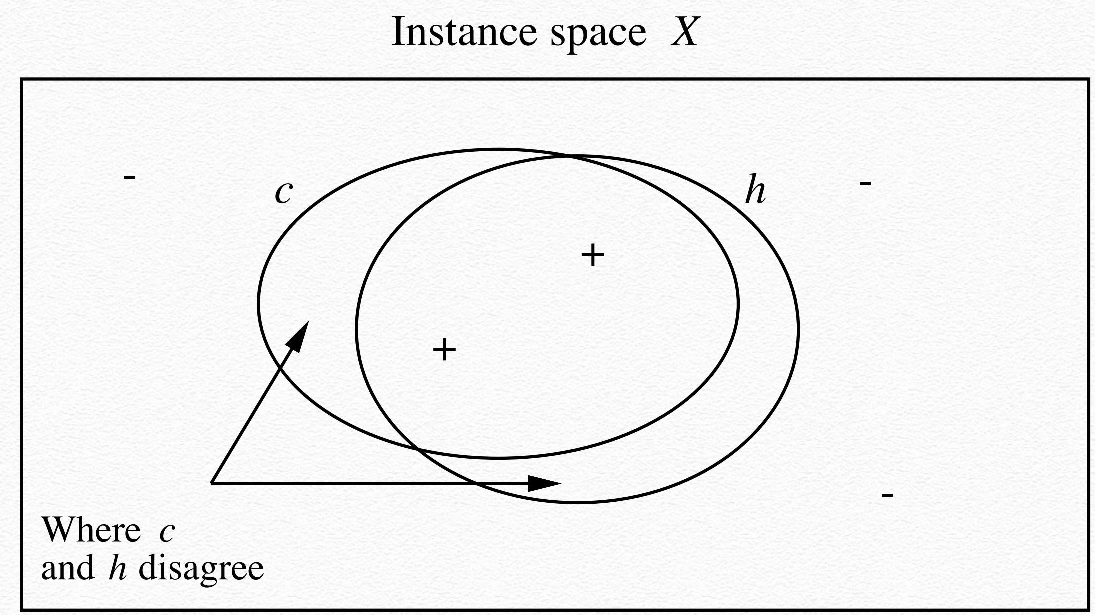
Learner observes a set D of noise-free training examples of the form $\langle x, c(x) \rangle$ of some target concept $c \in C$ where training instance x is randomly sampled from Q to query teacher for $c(x)$

Learner has to output a hypothesis h to approximate c where h is evaluated by its performance on new input instances randomly sampled from Q

True Error of a Hypothesis

Definition. The **true error** $error_Q(h)$ of hypothesis h w.r.t. target concept c and distribution Q is the probability that h misclassifies an input instance x randomly sampled from Q :

$$error_Q(h) = P_{x \sim Q}(h(x) \neq c(x)) \ .$$



Two Notions of Error

True error $error_Q(h)$ of hypothesis h w.r.t. target concept c

- How often $h(x) \neq c(x)$ over input instances randomly sampled from Q

Training error $error_D(h) = (1/|D|) \sum_{\langle x,\, c(x)\rangle \in D} (1 - \delta_{h(x),c(x)})$ of hypothesis h w.r.t. target concept c where $\delta_{h(x),c(x)}$ is of value 1 if $h(x) = c(x)$, and 0 otherwise

- How often $h(x) \neq c(x)$ over training instances

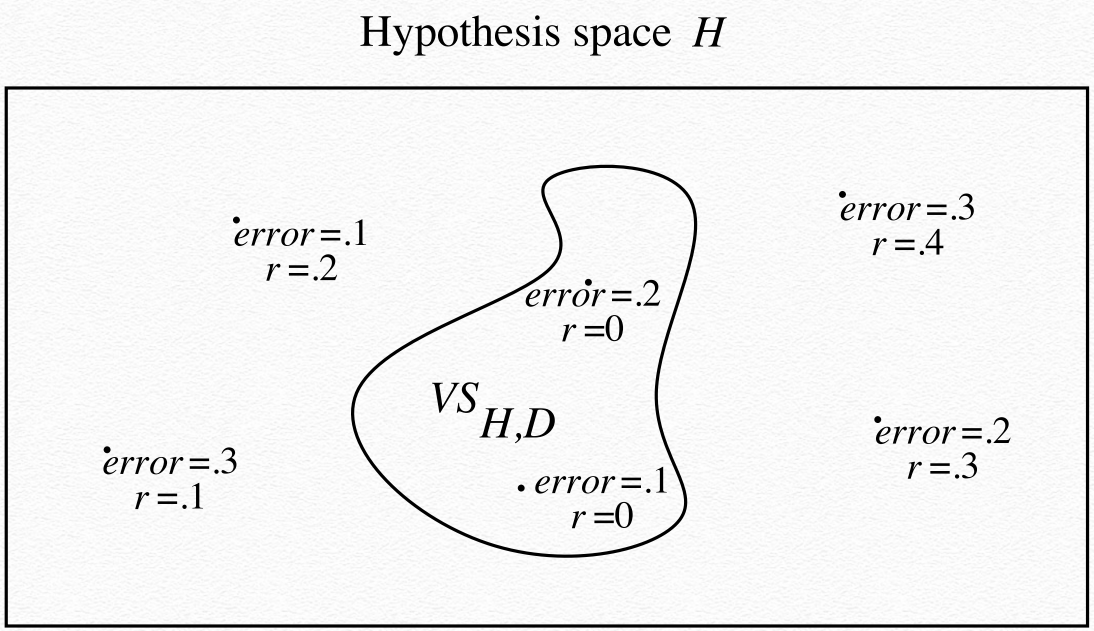
Key question. Can the **true error** of h be bounded given the **training error** of h?

- First consider when **training error** of h is 0 (i.e., $h \in VS_{H,D}$)

Exhausting the Version Space

Definition. The version space $VS_{H,D}$ is said to be **ϵ-exhausted** w.r.t. c and Q iff every hypothesis $h \in VS_{H,D}$ has error less than ϵ w.r.t. c and Q :

$$\forall h \in VS_{H,D} \;\; error_Q(h) < \epsilon \; .$$



How many training examples will ϵ-exhaust $VS_{H,D}$?

Theorem 1 (Haussler 1988). If H is finite and D is a set of independent random examples ($|D| \geq 1$) of some target concept c, then for any $0 \leq \epsilon \leq 1$, the probability that $VS_{H,D}$ is not **ϵ-exhausted** (w.r.t. c) is at most $|H| \exp(-\epsilon |D|)$.

Proof.

1. $VS_{H,D}$ is not ϵ-exhausted iff $\exists h \in H \;\; h \in VS_{H,D} \wedge error_Q(h) \geq \epsilon$ w.r.t. c
2. $error_Q(h) = P_{x \sim Q}(h(x) \neq c(x)) \geq \epsilon$
3. $P_{x \sim Q}(h(x) = c(x)) \leq 1 - \epsilon$. That is, the probability that h with $error_Q(h) \geq \epsilon$ is consistent with one random example is at most $1 - \epsilon$

How many training examples will ϵ-exhaust $VS_{H,D}$?

Proof (Cont'd).

4. The probability that h with $error_Q(h) \geq \epsilon$ is consistent with $|D|$ independent random examples is at most $(1-\epsilon)^{|D|}$:
$$P(h \in VS_{H,D} \wedge error_Q(h) \geq \epsilon) \leq (1-\epsilon)^{|D|}$$

5. $P(\exists h \in H \;\; h \in VS_{H,D} \wedge error_Q(h) \geq \epsilon) \leq |H|(1-\epsilon)^{|D|}$, by union bound

6. $P(VS_{H,D}$ is not ϵ-exhausted$) \leq |H|(1-\epsilon)^{|D|} \leq |H| \exp(-\epsilon|D|)$, by Step 1 and $(1-\epsilon) \leq \exp(-\epsilon)$ for any $0 \leq \epsilon \leq 1$

How many training examples will ϵ-exhaust $VS_{H,D}$?

Theorem 1 (Haussler 1988). If H is finite and D is a set of independent random examples ($|D| \geq 1$) of some target concept c, then for any $0 \leq \epsilon \leq 1$, the probability that $VS_{H,D}$ is not **ϵ-exhausted** (w.r.t. c) is at most $|H| \exp(-\epsilon |D|)$.

Limitation. Bound is loose (useless) due to large (infinite) H

Implication. This bounds the probability that a concept learning algorithm outputs a consistent hypothesis h with $error_Q(h) \geq \epsilon$

To determine the no. $|D|$ of training examples required to reduce this probability to be at most δ,

$$|H| \exp(-\epsilon |D|) \leq \delta \, .$$

Then, $|D| \geq (1/\epsilon) (\ln |H| + \ln (1/\delta))$.

How many training examples will ϵ-exhaust $VS_{H,D}$?

Corollary 1. Let $0 < \epsilon, \delta \leq 1$. If H is finite and D is a set of independent random examples of some target concept c s.t. $|D| \geq (1/\epsilon)\,(\ln|H| + \ln(1/\delta))$, then the probability that $VS_{H,D}$ is **ϵ-exhausted** (w.r.t. c) is at least $1 - \delta$:

$$P(\forall h \in VS_{H,D}\ \ error_Q(h) < \epsilon) \geq 1 - \delta\,.$$

Example 1. H = conjunctions of up to n Boolean literals and their negations. Then, $|H| = 3^n$ and $|D| \geq (1/\epsilon)\,(n \ln 3 + \ln(1/\delta))$

Example 2. H is as given in *EnjoySport* ($|H| = 973$). To guarantee with probability of at least .95 that $VS_{H,D}$ contains only hypotheses with $error_Q(h) < .1$, $|D| \geq (1/.1)\,(\ln 973 + \ln(1/.05)) = 98.76$

PAC Learning

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

Definition. The concept class C is **PAC-learnable** by L using H iff for all $c \in C$, distributions Q over X, and $0 < \epsilon, \delta \leq 1$, the probability that a learner L outputs a hypothesis $h \in H$ with $error_Q(h) \leq \epsilon$ is at least $1 - \delta$ in time that is polynomial in $1/\epsilon$, $1/\delta$, n, and size(c).

Implication 1. Given that C is **PAC-learnable** by L, if L incurs some minimum time to process each training example, then L learns from a polynomial no. of training examples

Implication 2. To show that C is **PAC-learnable** by L, show that each $c \in C$ can be learned from a polynomial no. of training examples using polynomial time per training example

Conjunctions of Boolean Literals are PAC-Learnable

C = conjunctions of up to n Boolean literals and their negations.

Theorem 2. C is **PAC-learnable** by FIND-S using $H = C$.

Proof.

1. For all $c \in C$, $P(\forall h \in VS_{H,D}\ \ error_Q(h) < \epsilon) \geq 1 - \delta$ in no. of training examples that is polynomial in n, $1/\epsilon$, and $1/\delta$, and independent of size(c), by Corollary 1 & Example 1 on page 15
2. For all $c \in C$, the probability that FIND-S outputs $h \in VS_{H,D} \subseteq H$ with $error_Q(h) < \epsilon$ is at least $1 - \delta$ in no. of training examples described in Step 1

Conjunctions of Boolean Literals are PAC-Learnable

C = conjunctions of up to n Boolean literals and their negations.

Theorem 2. C is **PAC-learnable** by FIND-S using $H = C$.

Proof (Cont'd).

3. To process each training example, FIND-S incurs time that is linear in n and independent of $1/\epsilon$, $1/\delta$, and size(c)
4. For all $c \in C$, the probability that FIND-S outputs $h \in VS_{H,D} \subseteq H$ with $error_Q(h) < \epsilon$ is at least $1 - \delta$ in time that is polynomial in n, $1/\epsilon$, and $1/\delta$, and independent of size(c), by Steps 2 and 3
5. C is PAC-learnable by FIND-S, by Implication 2 on page 16