Computational Learning Theories (COLT)

- Issues of COLT

. probability of successful learning

. complexity of hypothesis space

. number of training examples required for learning
(sample complexity)
. accuracy to which target concept is approximated
(generalization bounds)

- Concepts

Σ: alphabet for describing examples
eg. boolean alphabet $\{0,1\}$, real alphabet R
Σ^n: set of n-tuples of elements of Σ

$\Sigma^* = \bigcup_{n=1}^{\infty} \Sigma^n$: set of all non-empty finite strings of elements of Σ

a concept c over alphabet Σ:
$c : X \rightarrow \{0,1\}$ assuming $X \subseteq \Sigma^*$ (example or sample space)

- Training and Learning

C: concept space (a set of concepts)
M: machine
H: hypothesis space - a set of concepts which M determines.

A sample of length m is a sequence of m examples, that is,

$\underline{x} = (x_1, x_2, \cdots, x_m)$ in X^m.

A training sample $\underline{s}$ is an element of $(X \times 0, 1)^m$, that is,

$\underline{s} = ((x_1, b_1), (x_2, b_2), \cdots, (x_m, b_m))$

A learning algorithm L for (C, H): a procedure which accepts $\underline{s}$s for functions in C and output corresponding hypotheses in H, that is,

$L : (X \times \{0, 1\})^m \rightarrow H$

eg. $h = L(\underline{s})$, $h \in H$

A hypothesis $h \in H$ is consistent with $\underline{s}$ if

$h(x_i) = b_i$ for $1 \leqq i \leqq m$

- Probably Approximately Correct (PAC) Learning

. Error of any hypothesis $h \in H$ with respect to a target concept $t \in H$ is defined by

$$er(h,t) = \Pr_{x \in D}\{x \in X | h(x) \neq t(x)\}$$

-> The probability is taken with respect to random draw of x according to the sample distribution D.

Usually, $er(h,t)$ is abbreviated as $er(h)$.

. $S(m,t)$: a set of training samples of length m for a given target concept t where the examples are drawn from X.

. Any sample $\underline{x} \in X^m$ determines a training sample $\underline{s} \in S(m,t)$.

eg. If $\underline{x} = (x_1, x_2, \cdots, x_m)$, then

$$\underline{s} = ((x_1, t(x_1)), (x_2, t(x_2)), \cdots, (x_m, t(x_m))).$$

In other words, there exists $\phi(\underline{x})$ such that

$$\phi : X^m \to S(m,t).$$

. $er(L(\underline{s}))$: the error of the hypothesis when a learning algorithm L is supplied with $\underline{s}$.

. The algorithm L is a probably approximately correct (PAC) learning algorithm for the hypothesis H if a given

(1) a real number δ (confidence parameter, $0 < \delta < 1$) and

(2) a real number ϵ (accuracy parameter, $0 < \epsilon < 1$),

there is a positive integer $m_0 = m_0(\delta, \epsilon)$ such that

(1) for any target concept $t \in H$ and

(2) for any probability distribution D on X,

whenever $m \geqq m_0$ the following probability is satisfied:

$\Pr[\underline{s} \in S(m,t) | er(L(\underline{s})) < \epsilon] > 1 - \delta.$

. potential learnability:

Let $H[\underline{s}]$ be the set of all hypotheses which are consistent with $\underline{s}$, that is,

$H[\underline{s}] = \{h \in H | h(x_i) = t(x_i), 1 \leqq i \leqq m\}.$

Then, L is consistent if and only if $L(\underline{s}) \in H[\underline{s}]$ for all $\underline{s}$.

. the set of ϵ-bad hypotheses for t:

$B_\epsilon = \{h \in H | er(h) \geqq \epsilon\}.$

. H is potentially learnable if there is a positive integer $m_0 = m_0(\delta, \epsilon)$ such that whenever $m \geqq m_0$

$\Pr[\underline{s} \in S(m,t) | H[\underline{s}] \cap B_\epsilon = \varnothing] > 1 - \delta$

for any probability distribution D on X and $t \in H$.

. Theorem:

If H is potentially learnable and L is a consistent learning algorithm for H, then L is PAC.

(proof)

L is consistent -> $L(\underline{s}) \in H[\underline{s}] \ \forall \underline{s}$

H is potentially learnable -> $H[\underline{s}] \cap B_\epsilon = \varnothing$

-> $er(L(\underline{s})) < \epsilon$

-> L is PAC.

. Theorem:

Any finite hypothesis is potentially learnable.

(proof)

Suppose that H is a finite hypothesis and δ, ϵ, t, and D are given. Then, for any $h \in B_\epsilon$

$\Pr[x \in X | h(x) = t(x)] = 1 - er(h) \leqq 1 - \epsilon$

-> $\Pr[\underline{s} \in S(m,t) | h(x_i) = t(x_i),\ 1 \leqq i \leqq m] \leqq (1-\epsilon)^m$

-> $\Pr[\underline{s} \in S(m,t) | H[\underline{s}] \cap B_\epsilon \neq 0] \leqq |H|(1-\epsilon)^m$

This probability is less than δ provided $m \geqq m_0(\delta, \epsilon)$ where

$$m_0(\delta,\epsilon) = \left| \frac{1}{\epsilon} \ln \frac{|H|}{\delta} \right|$$

since

$|H|(1-\epsilon)^m \leqq |H|(1-\epsilon)^{m_0} < |H|e^{-\epsilon m_0} \leqq \delta.$

cf. $(1+x)^m \leqq e^{mx}$

That is, whenever $m \geqq m_0$

$$\Pr[\underline{s} \in S(m,t) | H[\underline{s}] \cap B_\epsilon = \varnothing] > 1-\delta.$$

Therefore, H is potentially learnable.

- The Growth Function

. Let $\underline{x} = (x_1, x_2, \cdots, x_m)$ be a sample of length m of examples from X. Then, the number of classifications of $\underline{x}$ by H is defined by $\Pi_H(\underline{x})$.

. Here, the number of distinct vectors of the form $(h(x_1), h(x_2), \cdots, h(x_m))$ as h runs through all hypotheses of H can be determined by

$$\Pi_H(\underline{x}) \leqq 2^m$$

where the concept is mapping defined by

$$c : X \rightarrow \{0, 1\}.$$

. The growth function is defined by

$$\Pi_H(m) = \max\left\{\Pi_H(\underline{x}) \mid \underline{x} \in X^m\right\}.$$

- The Vapnik-Chervonenkis (VC) Dimension

. A sample $\underline{x}$ of length m is 'shattered' by H if

$$\Pi_H(\underline{x}) = 2^m.$$

That is, H gives all possible classifications of $\underline{x}$.

. The VC dimension of H is the maximum length of a sample shattered by H, that is,

$$VCD(H) = \max\left\{m \mid \Pi_H(m) = 2^m\right\}.$$

. If H is a finite hypothesis space, then

$$VCD(H) \leqq \log_2|H|.$$

. example: linear discriminant function

$$h(\underline{x}) = w_0 + \sum_{i=1}^{n} w_i x_i$$

In this case, $VCD(H) = n+1$.

- Sauer's Lemma (Sauer, 1972)

Let $d \geqq 0$ and $m \geqq 0$ be given natural numbers and

$$\phi_d(m) = \begin{cases} 1 & \text{if } d = 0 \text{ or } m = 0 \\ \phi_d(m-1) + \phi_{d-1}(m-1) & otherwise \end{cases}$$

Then,

$$\phi_d(m) = \sum_{i=0}^{d} \binom{m}{i}.$$

(proof)

(1) If $m = 0$, $\phi_d(0) = \sum_{i=0}^{d} \binom{0}{i} = \binom{0}{0} = 1.$

(2) If $d = 0$, $\phi_0(m) = \binom{m}{0} = 1.$

(3) $$\begin{aligned} \phi_d(m-1) + \phi_{d-1}(m-1) &= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d-1} \binom{m-1}{i} \\ &= \sum_{i=0}^{d} \left[\binom{m-1}{i} + \binom{m-1}{i-1} \right] \\ &= \sum_{i=0}^{d} \binom{m}{i} \\ &= \phi_d(m) \end{aligned}$$

. $\phi_d(m)$ grows polynomially when $m > d$, that is, $0 \leqq \frac{d}{m} < 1$.

$$\left(\frac{d}{m}\right)^d \sum_{i=0}^{d} \binom{m}{i} \leqq \sum_{i=0}^{d} \left(\frac{d}{m}\right)^i \binom{m}{i} \leqq \sum_{i=0}^{m} \left(\frac{d}{m}\right)^i \binom{m}{i} = (1 + \frac{d}{m})^m \leqq e^d$$

cf. $(1+x)^m = \sum_{i=0}^{m} \binom{m}{i} x^i$

This implies that

$$\phi_d(m) = \sum_{i=0}^{d} \binom{m}{i} \leqq e^d \left(\frac{m}{d}\right)^d = \left(\frac{em}{d}\right)^d.$$

. Theorem:

If $d = VCD(H)$, $\Pi_H(m) \leqq \phi_d(m) \leqq \left(\frac{em}{d}\right)^d$.

. The logarithmic growth function

Let $G(m) = \ln \Pi_H(m)$.

Then,

$$G(m) \leqq d(1 + \ln \frac{m}{d}).$$

Note that the above logarithmic growth function is valid for $m > d$.

- VC Dimension of Artificial Neural Networks

The large class of artificial neural networks including sigmoidal functions, radial basis functions, and sigma-pi networks have the following VC dimension bounds (Goldberg and Jerrum, 1993; Sakurai, 1993 and 1995):

$$O(W \log h) \leqq VCD(H) \leqq O(W^2 h^2)$$

where W represents the number of total parameters and h represents the number of hidden units.

- The Upper Bounds of Sample Complexity

. Assuming that H is potentially learnable.
Then, there is a positive integer $m_0 = m_0(\delta, \epsilon)$ such that whenever $m \geqq m_0$,

$$\Pr[\underline{s} \in S(m,t) \mid H[\underline{s}] \cap B_\epsilon = \varnothing] > 1 - \delta \quad \dots \text{ (1)}$$

for any probability distribution D on X and $t \in H$.

. What is m_0 which will guarantee the probability condition of (1)?

. Lemma:

Let H has the finite VC dimension. Then, for $m \geqq 8/\epsilon$, the following inequality holds:

$$\Pr[\underline{s} \in S(m,t) | H[\underline{s}] \cap B_\epsilon \neq \varnothing] \leqq 2\Pi_H(2m)2^{-\frac{\epsilon m}{2}}.$$

. Theorem (Blumer, 1989): upper bound of sample complexity

Suppose H is a hypothesis space of $VCD(H) = d \geqq 1$ and

$$m_0 = m_0(\delta, \epsilon) = \lfloor \frac{4}{\epsilon}(d \log_2 \frac{12}{\epsilon} + \log_2 \frac{2}{\delta}) \rfloor.$$

Then, for any $m \geqq m_0$,

$$\Pr[\underline{s} \in S(m,t) | H[\underline{s}] \cap B_\epsilon \neq \varnothing] \leqq \delta.$$

(proof)

From the lemma,

$$\Pr[\underline{s} \in S(m,t) | H[\underline{s}] \cap B_\epsilon \neq \varnothing] \leqq 2\Pi_H(2m)2^{-\frac{\epsilon m}{2}} \leqq 2\left(\frac{e2m}{d}\right)^d 2^{-\frac{\epsilon m}{2}}.$$

Let

$$2\left(\frac{e2m}{d}\right)^d 2^{-\frac{\epsilon m}{2}} \leqq \delta.$$

Then,

$$d\ln\left(\frac{2e}{d}\right) + d\ln m - \frac{\epsilon m}{2}\ln 2 \leqq \ln\frac{\delta}{2}.$$

Rearranging the above equation, we get

$$\frac{\epsilon m}{2}\ln 2 - d\ln m \geqq d\ln\left(\frac{2e}{d}\right) + \ln\frac{2}{\delta}. \quad \dots \quad (1)$$

Since $\ln x \leqq \ln\frac{1}{c} - 1 + cx$ for any $x > 0$ and $c > 0$,

$$d\ln m \leqq d(\ln\frac{4d}{\epsilon\ln 2} - 1) + \frac{\epsilon\ln 2}{4}m. \quad \ldots \quad (2)$$

Here, we set

$$c = \frac{\epsilon\ln 2}{4d} \quad \text{and} \quad x = m.$$

From (1) and (2),

$$\frac{\epsilon m}{4}\ln 2 \geqq d\ln\left(\frac{2e}{d}\right) + \ln\left(\frac{2}{\delta}\right) + d\ln\left(\frac{4d}{\epsilon\ln 2}\right) - d.$$

Rearranging the above equation, we get

$$m \geqq \frac{4}{\epsilon}(d\log_2\frac{12}{\epsilon} + \log_2\frac{2}{\delta}).$$

Here, note that $8/\ln 2 < 12$.

. example: linear discriminant function

$$h(\underline{x}) = w_0 + \sum_{i=1}^{n} w_i x_i$$

In this case, $VCD(H) = n+1$.

The sufficient number of samples for PAC learning is

$$m_0 = \lceil \frac{4}{\epsilon}((n+1)\log_2\frac{12}{\epsilon} + \log_2\frac{2}{\delta}) \rceil.$$

Let $\epsilon = 0.1$, $\delta = 0.05$ (95% confidence), $n = 2$. Then,

$$m_0 = \lceil \frac{4}{0.1}(3\log_2\frac{12}{0.1} + \log_2\frac{2}{0.05}) \rceil = 656.$$

-> This is quite large number of samples compared to the minimum number of samples (= 4) to learn a linear decision boundary in 2-D space.

- The Lower Bounds of Sample Complexity

. If L is PAC,

$\Pr[\underline{s} \in S(m,t) | er(L(\underline{s})) < \epsilon] > 1-\delta$ or

$\Pr[\underline{s} \in S(m,t) | er(L(\underline{s})) \geqq \epsilon] \leqq \delta$.

. What is the upper bound m_0 of m that does not satisfy the above inequality? That is,

$\Pr[\underline{s} \in S(m,t) | er(L(\underline{s})) \geqq \epsilon] \geqq \delta$.

Here, if $m \leqq m_0$, L can not be PAC.

. In other words, if $m > m_0$, L has the possibility to be PAC.

. Theorem: lower bounds of sample complexity

Suppose L is PAC learning algorithm for H. Then,

$$m(\delta,\epsilon) > \frac{1-\epsilon}{\epsilon} \ln \frac{1}{\delta}$$

for any δ and ϵ between 0 and 1.

(proof)

Let $h = L(\underline{s})$ and $er(h) \geqq \epsilon$. Then,

$\Pr[x \in X | h(x) = t(x)] = 1 - er(h) \leqq 1-\epsilon$ and

$\Pr[\underline{s} \in S(m,t) | h(x_i) = t(x_i), 1 \leqq i \leqq m] \leqq (1-\epsilon)^m$. ... (1)

Let

$\Pr[\underline{s} \in S(m,t) | h(x_i) = t(x_i), 1 \leqq i \leqq m] \geqq \delta$. ... (2)

Then, from (1) and (2),

$(1-\epsilon)^m \geqq \delta.$

-> $m \ln(1-\epsilon) \geqq -\ln\frac{1}{\delta}$

-> $m \leqq -\frac{1}{\ln(1-\epsilon)} \ln\frac{1}{\delta}$

-> $m \leqq \frac{1-\epsilon}{\epsilon} \ln\frac{1}{\delta}$ since $-\ln(1-\epsilon) = \ln(1+\frac{\epsilon}{1-\epsilon}) \leqq \frac{\epsilon}{1-\epsilon}$.

Therefore, if

$m \leqq \frac{1-\epsilon}{\epsilon} \ln\frac{1}{\delta},$

$\Pr[\underline{s} \in S(m,t) | er(L(\underline{s}) \geqq \epsilon] \geqq \delta$ or

$\Pr[\underline{s} \in S(m,t) | er(L(\underline{s})) < \epsilon] < 1-\delta.$

This implies that

$m > \frac{1-\epsilon}{\epsilon} \ln\frac{1}{\delta}$

if L is PAC learning algorithm for H.

. Theorem: lower bounds of sample complexity (Ehrenfeucht, 1989)

For any H of $VCD(H) = d \geqq 1$, and for any PAC learning algorithm L for H,

$m(\delta,\epsilon) > \frac{d-1}{32\epsilon}$

for $\delta \leqq 1/100$ and $\epsilon \leqq 1/8$

. Theorem: lower bounds of sample complexity

Let C be a concept space and H a hypothesis space such that C has the VC dimension at least 1. Suppose L is any PAC learning algorithm for (C, H). Then,

$$m(\delta,\epsilon) > \lfloor \max(\frac{1}{\epsilon}\ln\frac{1}{\delta}, \frac{VCD(C)-1}{32\epsilon}) \rfloor$$

for $\delta \leqq 1/100$ and $\epsilon \leqq 1/8$.

. example: linear discriminant function

$$h(\underline{x}) = w_0 + \sum_{i=1}^{n} w_i x_i$$

In this case, $VCD(H) = n+1$.

Let $\delta = 0.05$, $\epsilon = 0.1$, and $n = 2$.

Then, the lower bound of sample complexity is

$$m_0^L = \lfloor \max(\frac{3-1}{32 \cdot 0.1}, \frac{1}{0.1}\ln\frac{1}{0.05}) \rfloor = 30.$$

Note that

- the upper bound of sample complexity: $m_0^U = 656$ and
- the minimum number of samples to determine the decision boundary: $m_0^* = 4$

- Summary of Sample Complexity

. If H is finite and L is consistent,
L is PAC and

$$m(\delta,\epsilon) = O(\frac{1}{\epsilon}(\ln|H| + \ln\frac{1}{\delta})).$$

. If H has the finite $VCD(H) = d$ and L is consistent,
L is PAC and

$$m(\delta,\epsilon) = O(\frac{1}{\epsilon}(d\ln\frac{1}{\epsilon} + \ln\frac{1}{\delta})).$$

. If L is PAC, C must have the finite $VCD(C) = d$ and

$$m(\delta,\epsilon) = \Omega(\frac{1}{\epsilon}(d + \ln\frac{1}{\delta})).$$

Note that

(1) $f = O(g)$ when there is some constant C such that
$f(x) \leqq Cg(x) \quad \forall x.$

(2) $f = \Omega(g)$ when there is some constant K such that
$f(x) \geqq Kg(x) \quad \forall x.$