Ref: [2014] [S. Shalev, etal] Understanding Machine Learning.

ML P1

Chap2 A Gentle Start.

2.1 A Formal Model.

$\boxed{\text{Domain set}}$: $\boxed{\mathcal{X}}$, also called $\boxed{\text{instance space}}$.
pts in $\mathcal{X}$: $\boxed{\text{instances}}$.

$\boxed{\text{Label set}}$: $\boxed{\mathcal{Y}}$

$\boxed{\text{Training set}}$: $\boxed{S} = ((x_1, y_1), \cdots, (x_m, y_m))$, a finite seq. of pairs in $\mathcal{X} \times \mathcal{Y}$.

Each (x_i, y_i) is called a $\boxed{\text{training example}}$.

S is also called a $\boxed{\text{training set}}$ (although not really a set since we might have repetitions).

The Learner's Output: (Domain / Label / Training : Learner's Inputs).

A function $\boxed{h : \mathcal{X} \to \mathcal{Y}}$, called a $\boxed{\text{predictor}}$, a $\boxed{\text{hypothesis}}$, or a $\boxed{\text{classifier}}$.

Notation : $\boxed{A(S)}$, where $\boxed{A}$ is a learning algorithm.

A "simple" data generating model : (to be modified later)

$\boxed{\mathcal{D}}$: a probability distribution on $\mathcal{X}$.

$\exists$ "correct" labelling function $\boxed{f : \mathcal{X} \to \mathcal{Y}}$. (to be relaxed later).

x_i are drawn from $\mathcal{X}$ according to $\mathcal{D}$.

$y_i = f(x_i)$.

Measure of success :

$$\boxed{L_{\mathcal{D},f}(h)} := \mathbb{P}_{x \sim \mathcal{D}} [h(x) \neq f(x)] = \mathcal{D}(\{x \mid f(x) \neq h(x)\}),$$

called the $\boxed{\text{error of the classifier}}$ h.

Rmk: The $\boxed{\text{generalization error}}$, the $\boxed{\text{risk}}$, or the $\boxed{\text{true error}}$ of h are synonyms of $L_{\mathcal{D},f}(h)$.

L refers to "loss" of the learner.

Note: The learner is blind to $\mathcal{D}$.

2.2 Empirical Risk Minimization.

Notation : $\boxed{h_S} : \mathcal{X} \to \mathcal{Y}$: the output.

Def

$$\boxed{L_S(h)} := \frac{|\{i \in [m] \mid h(x_i) \neq y_i\}|}{m}, \text{ called the}$$

$\boxed{\text{training error}}$, $\boxed{\text{empirical error}}$, or $\boxed{\text{empirical risk}}$.

$\boxed{\text{Empirical Risk Minimization}}$ ($\boxed{\text{ERM}}$) is the learning paradigm searching for h that minimizes $L_S(h)$.

Rmk :

ERM may fail if one is not careful; overfitting may arise.

2.3 Empirical Risk Minimization w/ Inductive Bias.

A common sol'n to overfitting is to apply ERM to a restricted search space.

Def

$\boxed{\mathcal{H}}$:= a set of certain functions $\mathcal{X} \to \mathcal{Y}$, called a $\boxed{\text{hypothesis class}}$.

$\boxed{\text{ERM}_{\mathcal{H}}}$ is the algorithm applying ERM only on $\mathcal{H}$.

Thus $\text{ERM}_{\mathcal{H}}(S) \in \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, L_S(\mathcal{H})$.

Rmk :

① The restriction of applying ERM on a certain $\mathcal{H}$ is called an $\boxed{\text{inductive bias}}$.

② The choice of which $\mathcal{H}$ to look at is based on some prior knowledge about the problem.

③ Fundamental question : Over which $\mathcal{H}$ will $\text{ERM}_{\mathcal{H}}$ not result in overfitting?

2.3.1 Finite Hypothesis class.

Goal : Prove that $\text{ERM}_{\mathcal{H}}$ will not overfit if $|\mathcal{H}| < \infty$ and $|S|$ is sufficiently large.

Notation : $\boxed{h_S} := \text{ERM}_{\mathcal{H}}(S)$

Def

The assumption that "$\exists\, h^* \in \mathcal{H}$ s.t. $L_{\mathcal{D},f}(h^*) = 0$" is called the $\boxed{\text{realizability assumption}}$.

Rmk : This assumption is to be relaxed later.

$\boxed{\text{The i.i.d. assumption}}$: S comes from sampling $\mathcal{X}$ iid w.r.t. $\mathcal{D}$. This assumption guarantees S to be representative of $\mathcal{D}$ in probability sense.

Notation : $\boxed{S \sim \mathcal{D}^m}$, where $m = |S|$.

Notation :

① $\boxed{\delta}$: probability of "getting a nonrepresentative sample"
$(1-\delta)$ is called the $\boxed{\text{confidence parameter}}$.

② $\boxed{\varepsilon}$: $\boxed{\text{accuracy parameter}}$, measuring the quality of predictn.

The event "$L_{\mathcal{D},f}(h) > \varepsilon$" is interpreted as a failure of the learner.

To prove the main thm, we begin w/ a simple lemma.

Lemma 2.2 (Union Bound)

$$\mathcal{D}(A \cup B) \le \mathcal{D}(A) + \mathcal{D}(B).$$

Cor 2.3

$\mathcal{H}$: a finite hypothesis class.

$\delta \in (0,1)$, $\varepsilon > 0$.

m : an integer w/ $m \ge \dfrac{\log(|\mathcal{H}|/\delta)}{\varepsilon}$.

$f : \mathcal{X} \to \mathcal{Y}$: a labeling function.

Assume S is a training set from sampling $\mathcal{X}$ iid w.r.t. a distribution $\mathcal{D}$ and labelling by f of size m, and assume also the realizability assumptn.

Then, for every ERM hypothesis h_S,

$L_{(\mathcal{D},f)}(h_S) \le \varepsilon$ w/ probability at least $1-\delta$.

<Pf>

Let $A = \{S = (x_1, \dots, x_m) \mid L_{\mathcal{D},f}(h_S) > \varepsilon\}$.

Our goal is to figure out condition on m s.t.

$\mathcal{D}^m(A) \le \delta$.

the collection of "bad" hypotheses.

Let $\mathcal{H}_B := \{h \in \mathcal{H} \mid L_{(\mathcal{D},f)}(h) > \varepsilon\}$ and

the samplings that mislead the learner

$M := \{S = (x_1, \dots, x_m) \mid \exists\, h \in \mathcal{H}_B \text{ s.t. } L_S(h) = 0\}$.

(B refers to "Bad" and M refers to "misleading")

For $S \in A$, $L_{\mathcal{D},f}(h_S) > \varepsilon$. $\Rightarrow h_S \in \mathcal{H}_B$.

By realizability assumption, $L_{\mathcal{D},f}(h_S) = 0$, $\forall S$.

Thus $S \in M$ and $A \subseteq M$. $(*_1)$

Note that M can be expressed as

$$M = \bigcup_{h \in \mathcal{H}_B} \{S = (x_1, \dots, x_m) \mid L_S(h) = 0\}. \quad (*_2)$$

By iid assumption,

$$\mathcal{D}^m(\{S = (x_1, \dots, x_m) \mid L_S(h) = 0\})$$
$$= \prod_{i=1}^{m} \mathcal{D}(\{x \mid h(x) = f(x)\}) = \prod_{i=1}^{m} (1 - L_{\mathcal{D},f}(h)) \quad (*_3)$$

In particular, for $h \in \mathcal{H}_B$, $(*_3) < (1-\varepsilon)^m$. $(*_4)$

Thus

$$\mathcal{D}^m(A) \overset{(*_1)}{\le} \mathcal{D}^m(M) \overset{(*_2)}{\le} \sum_{h \in \mathcal{H}_B} \mathcal{D}^m(\{S = (x_1, \dots, x_m) \mid L_S(h) = 0\})$$

$$\overset{(*_3),\,(*_4),\,\text{L2.2}}{\le} \sum_{h \in \mathcal{H}_B} (1-\varepsilon)^m = |\mathcal{H}_B| \cdot (1-\varepsilon)^m \le |\mathcal{H}| \cdot (1-\varepsilon)^m.$$

Note that $1 - \varepsilon \le e^{-\varepsilon}$, $\forall\, \varepsilon \ge 0$.

Thus $\mathcal{D}^m(A) \le |\mathcal{H}| \cdot e^{-m\varepsilon}$.

To make $\mathcal{D}^m(A) \le \delta$, we may make

$|\mathcal{H}| \cdot e^{-m\varepsilon} \le \delta$. i.e. $|\mathcal{H}|/\delta \le e^{m\varepsilon}$

i.e. $m \ge \dfrac{\log(|\mathcal{H}|/\delta)}{\varepsilon}$, as assumed. (#)

Chap 3 A Formal Learning Model.

3.1 PAC Learning.

PAC means Probably Approximately Correct.

Def

A hypothesis class $\mathcal{H}$ is PAC learnable if

$\exists\, m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm s.t.

$\forall\, \varepsilon, \delta \in (0,1)$ and every dist. $\mathcal{D}$ over $\mathcal{X}$, and

$\forall$ labelling function $f : \mathcal{X} \to \{0,1\}$,

if the realizability assumption holds, then when running the algorithm on $m \ge m_{\mathcal{H}}(\varepsilon, \delta)$ iid examples generated by $\mathcal{D}$ and labelled by f, the returned hypothesis h satisfies

$L_{(\mathcal{D},f)}(h) \le \varepsilon$ w/ prob. at least $1-\delta$. (over all possible examples)

Rmk:

① The accuracy parameter $\varepsilon \leftrightarrow$ "approximately correct".

② The confidence parameter $\delta \leftrightarrow$ "probably".

Def

$\mathcal{H}$: hypothesis class that is PAC learnable.

The sample complexity of learning $\mathcal{H}$ is the minimal $m_{\mathcal{H}}$ satisfying conditions in the def of PAC learnable.

Cor 3.2

Every finite hypothesis class $\mathcal{H}$ is PAC learnable w/ sample complexity $m_{\mathcal{H}}(\varepsilon, \delta) \le \left\lceil \dfrac{\log(|\mathcal{H}|/\delta)}{\varepsilon} \right\rceil$.

3.2 A More General Learning Model.

3.2.1 Releasing the Realizability Assumption — Agnostic PAC Learning.

Motivation:

Possible that we have $(x,y),(x,y') \in S$ w/ $y \neq y'$.

Sol'n:

Replacing target labelling function w/ a "data-labels generating distribution".

From now on, $\boxed{\mathcal{D}}$:= a prob. dist. over $\mathcal{X} \times \mathcal{Y}$.

i.e. the joint distribution on $\mathcal{X} \times \mathcal{Y}$.

Rmk:

$$\mathcal{D}(x,y) = \underset{\text{(conditional)}}{\mathcal{D}((x,y)\mid x)} \cdot \underset{\text{(marginal)}}{\mathcal{D}_x}.$$

Def (redefine)

$$\boxed{L_{\mathcal{D}}(h)} := \underset{(x,y)\sim\mathcal{D}}{P}[h(x)\neq y] := \mathcal{D}(\{(x,y) \mid h(x)\neq y\}).$$

Rmk:

$$\boxed{L_S(h)} := \frac{|\{i \in [m] \mid h(x_i) \neq y_i\}|}{m}, \text{ as before.}$$

Def

$\mathcal{D}$: distribution on $\mathcal{X} \times \{0,1\}$.

Define $\boxed{f_{\mathcal{D}}(x)} = \begin{cases} 1, & \text{if } P[y=1 \mid x] \geq 1/2 \\ 0, & \text{o.w.} \end{cases}$

$f_{\mathcal{D}}$ is called the Bayes optimal predictor.

Prop

For every classifier g, $L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g)$.

Thus, $f_{\mathcal{D}}$ is optimal.

case1 $P[y=1 \mid x] \geq 1/2$.
case2 $P[y=1 \mid x] < 1/2$.
Compare $P[f_{\mathcal{D}}(x) \neq y]$ and $P[g(x) \neq y]$.

Rmk:

$\because$ $\mathcal{D}$ is unknown $\therefore$ We can not compute $f_{\mathcal{D}}$.

Def 3.3

A hypothesis class $\mathcal{H}$ is agnostic PAC learnable if $\exists\, m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm s.t. $\forall\, \varepsilon, \delta \in (0,1)$, $\mathcal{D}$: dist. over $\mathcal{X} \times \mathcal{Y}$, when running the algorithm on $m \geq m_{\mathcal{H}}(\varepsilon,\delta)$ iid examples generated by $\mathcal{D}$, the returned hypothesis h satisfies

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon, \text{ w/ prob. at least } 1-\delta$$

(over all m training examples).

End of W1

3.2.2 The Scope of Learning Problems Modeled

For different learning tasks, we need different model, ex: multiclass classification, regression.

Rmk:

① For multiclass classification, the measurement of quality of a hypothesis is as before.

② For regression, we may use the expected square difference

$$L_{\mathcal{D}}(h) := \underset{(x,y)\sim\mathcal{D}}{E}(h(x)-y)^2, \text{ instead.}$$

Motivation:

We need to generalize the formalism of measure of success to accommodate a wide range of learning tasks.

Def

A function $\ell : \mathcal{H} \times Z \to \mathbb{R}_+$ is called a loss function where $\boxed{\mathbb{R}_+}$:= {nonnegative real numbers}.

Given a distribution $\mathcal{D}$ on Z, the risk function is

$$\boxed{L_{\mathcal{D}}(h)} := \underset{z\sim\mathcal{D}}{E}[\ell(h,z)].$$

For $z_1, \cdots, z_m \in Z$ drawn iid wrt $\mathcal{D}$ from Z, the empirical risk is, where $S = (z_1, \cdots, z_m)$,

$$\boxed{L_S(h)} := \frac{1}{m}\sum_{i=1}^{m} \ell(h, z_i).$$

Example:

① (0-1 loss)

$Z := \mathcal{X} \times \mathcal{Y}$.

$$\ell_{0-1}(h,(x,y)) := \begin{cases} 0, & \text{if } h(x) = y \\ 1, & \text{if } h(x) \neq y. \end{cases} \quad (= \chi_{\{h(x)\neq y\}}).$$

The $L_{\mathcal{D}}(h)$ defined here (in this subsect'n) is the same as before.

② (square loss)

$Z := \mathcal{X} \times \mathcal{Y}$

$$\ell_{sq}(h,(x,y)) := (h(x)-y)^2.$$

Then $L_{\mathcal{D}}(h)$ is the expected square difference.

Rmk:

As demonstrated, the newly defined $L_{\mathcal{D}}(h)$ is a generalizat'n.

Def 3.4

A hypothesis class $\mathcal{H}$ is agnostic PAC learnable w.r.t. a set Z and a loss function $\ell : \mathcal{H} \times Z \to \mathbb{R}_+$ if $\exists\, m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm s.t. $\forall\, \varepsilon, \delta \in (0,1)$, dist. $\mathcal{D}$ over Z, when running the algorithm on $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ iid examples generated by $\mathcal{D}$, the returned $h \in \mathcal{H}$ satisfies

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon$$ w/ prob. at least $1-\delta$ (over the choice of m training examples).

where

$$L_{\mathcal{D}}(h) := \mathop{E}_{z \sim \mathcal{D}} [\ell(h,z)].$$

Rmk:

① Strictly speaking, we need to require, for each fixed $h \in \mathcal{H}$, $\ell(h,z)$ is $\mathcal{D}$-measurable as a function of z. (Of course, a σ-algebra on Z needs to be preassigned.)

② The authors talk about "representation indep. learning" (or "improper learning") while the learning defined previously is called proper learning.

Chap 4 Learning via Uniform Convergence.

4.1 Uniform Convergence is Suff. for Learnability.

Def 4.1

A training set S is called ε-representative (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss func'n ℓ and dist. $\mathcal{D}$) if

$$|L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon, \ \forall\, h \in \mathcal{H}.$$

Lemma 4.2

S : $\varepsilon/2$-representative.

$h_S \in \mathrm{ERM}_{\mathcal{H}}(S)$.

Then

$$L_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon.$$

<Pf>.

$$L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \varepsilon/2 = \min_{h \in \mathcal{H}} L_S(h) + \varepsilon/2$$

$$\leq \min_{h \in \mathcal{H}} \left(L_{\mathcal{D}}(h) + \varepsilon/2\right) + \varepsilon/2$$

$$= \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon. \ \#$$

Rmk:

By L 4.2, to guarantee $\mathcal{H}$ to be PAC learnable, it suffices to guarantee, w/ prob. at least $1-\delta$, S is $\varepsilon/2$-representative.

Def 4.3 (Unif. Con.)

A hypothesis class $\mathcal{H}$ is said to have the Uniform convergence property (w.r.t. domain Z, loss function ℓ) if

$\exists\, m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$ s.t. $\forall\, \varepsilon, \delta \in (0,1)$, dist. $\mathcal{D}$ over Z, if S is a sample of $m \geq m_{\mathcal{H}}^{UC}(\varepsilon, \delta)$ examples iid from $\mathcal{D}$, then S is ε-representative w/ prob. $\geq 1-\delta$.

Rmk:

The adj. "uniform" refers to the unif. of $m_{\mathcal{H}}^{UC}$ over all members of $\mathcal{H}$ and all dist. $\mathcal{D}$ over Z.

Cor 4.4

$\mathcal{H}$: a hypothesis class w/ UC property.

Then $\mathcal{H}$ is agnostic PAC learnable w/ sample complexity $m_{\mathcal{H}}(\varepsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\varepsilon/2, \delta)$.

Moreover, $\mathrm{ERM}_{\mathcal{H}}$ is a successful agnostic PAC learner for $\mathcal{H}$.

4.2 Finite Classes are Agnostic PAC learnable.

We are going to prove the statement as the title suggests. The following lemma is used:

Lemma 4.5 (Hoeffding's inequality)

$\theta_1, \cdots, \theta_m$: iid r.v.

$E[\theta_i] = \mu$, and $P[a \leq \theta_i \leq b] = 1$, $\forall\, i$.

Then, $\forall\, \varepsilon > 0$,

$$P\left[\left| \frac{1}{m} \sum_{i=1}^{m} \theta_i - \mu \right| > \varepsilon \right] \leq 2 e^{-2m\varepsilon^2/(b-a)^2}.$$

Cor 4.6

$\mathcal{H}$: finite hypothesis class.

Z : domain.

$\ell : \mathcal{H} \times Z \to [a,b]$: a loss function.

Then $\mathcal{H}$ has the unif. con. property w/

$$m_{\mathcal{H}}^{UC}(\varepsilon,\delta) \le \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\varepsilon^2/(b-a)^2} \right\rceil.$$

Therefore (by Cor 4.4),

$\mathcal{H}$ is agnostic PAC learnable using ERM w/ sample complexity

$$m_{\mathcal{H}}(\varepsilon,\delta) \le m_{\mathcal{H}}^{UC}(\varepsilon/2,\delta) \le \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\varepsilon^2/(b-a)^2} \right\rceil.$$

<pf>

Two steps :

Step 1 : Applying the union bound.

Let

$$B = \{ \underset{\text{in } (\mathcal{X}\times\mathcal{Y})^m \text{ w/ } \mathcal{D}^m}{S} : \exists h \in \mathcal{H} \text{ s.t. } |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon \}.$$

Then

$$B = \bigcup_{h\in\mathcal{H}} \{ S : |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon \}.$$

Thus,

$$\mathcal{D}^m(B) \le \sum_{h\in\mathcal{H}} \mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}).$$

Step 2 : Applying a measure concentrat'n ineq.

(Here, we use Hoeffding's).

Denote $S = (z_1,\cdots,z_m)$. Fix $h\in\mathcal{H}$.

$\theta_i := \ell(h,z_i)$, $\mu := \underset{z\sim\mathcal{D}}{E}[\ell(h,z)] = E[\ell(h,z_i)], \forall i$.

Note that $L_{\mathcal{D}}(h) = \mu$ and $L_S(h) = \frac{1}{m}\sum_{i=1}^m \theta_i$.

$\because$ $z_1,\cdots,z_m$ are iid

$\therefore$ $\theta_1,\cdots,\theta_m$ are iid.

Thus, by Hoeffding's ineq,

$$\mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}) = P\left[\left|\frac{1}{m}\sum_{i=1}^m \theta_i - \mu\right| > \varepsilon\right]$$

$$\le 2e^{-2m\varepsilon^2/(b-a)^2}.$$

$$\Rightarrow \mathcal{D}^m(B) \le 2\cdot|\mathcal{H}|\cdot e^{-2m\varepsilon^2/(b-a)^2}.$$

To make $\mathcal{D}^m(B) < \delta$, it suffices to make $2e^{-m\varepsilon^2/(b-a)^2} < \delta$. i.e. $m > \frac{\log(2|\mathcal{H}|/\delta)}{2\varepsilon^2/(b-a)^2}$. #

Rmk : ("Discretizat'n Trick").

Though the assumption $|\mathcal{H}| < \infty$ seems unrealistic, due to limitat'n of finite expression of numbers in computers, it becomes realistic.

Rmk :

① Hypothesis classes w/ unif. con. property are also called $\boxed{\text{Glivenko-Cantelli classes}}$.

② (see Chap 6) In binary classification problems, $UC \Leftrightarrow$ learnable.

(Not the case for general learning problems).

Chap 5 The Bias Complexity Tradeoff.

End of W2

Motivation :

Is there a universal learner ?

5.1 The No-Free-Lunch Theorem.

Thm 5.1 (No-Free-Lunch)

A : an algorithm for binary classification w.r.t. 0-1 loss and over domain $\mathcal{X}$.

m : an integer w/ $m < |\mathcal{X}|/2$.

Then $\exists$ dist. $\mathcal{D}$ over $\mathcal{X}\times\{0,1\}$ s.t.

1. $\exists f : \mathcal{X}\to\{0,1\}$ w/ $L_{\mathcal{D}}(f) = 0$.
2. $L_{\mathcal{D}}(A(S)) \ge 1/8$ w/ prob. $\ge 1/7$ over the choice of $S\sim\mathcal{D}^m$.

Rmk :

Intuitively, the thm states that for every learner (the algorithm A), there exists a task on which it fails, while there is another learner that succeeds.

5.1.1 No-Free-Lunch and Prior Knowledge.

Rmk :

A preassigned hypothesis class $\mathcal{H}$ represents a prior knowledge. Thus, by taking $\mathcal{H}$ = {all functions from $\mathcal{X}$ to $\{0,1\}$}, we are in a state of lack of prior knowledge.

Cor 5.2

$\mathcal{X}$: an infinite domain set.

$\mathcal{H} :=$ {all functions $\mathcal{X}\to\{0,1\}$}.

Then $\mathcal{H}$ is not PAC learnable.

⟨Pf⟩

Suppose $\mathcal{H}$ is PAC learnable.

Then $\exists$ algorithm A and $m_{\mathcal{H}}:(0,1)^2 \to \mathbb{N}$ w/ the "PAC learnable properties".

Take $\varepsilon < 1/8$ and $\delta < 1/7$.

Let $\mathcal{D}$ be a distributn on $\mathcal{X} \times \{0,1\}$ corresponding to A as in the No-Free-Lunch Thm.

PAC learnability implies that

$L_{\mathcal{D}}(A(S)) < \varepsilon < 1/8$ w/ prob. $> 1-\delta > 6/7$ over the choice of $S \sim \mathcal{D}^m$.

i.e.

$$\mathcal{D}^m(\{S \mid L_{\mathcal{D}}(A(S)) < 1/8\}) > 6/7.$$

$$\Leftrightarrow 1 - \mathcal{D}^m(\{S \mid L_{\mathcal{D}}(A(S)) \geq 1/8\}) > 6/7.$$

$$\Leftrightarrow \mathcal{D}^m(\{S \mid L_{\mathcal{D}}(A(S)) \geq 1/8\}) < 1/7$$

$$\Leftrightarrow \sim(\mathcal{D}^m(\{S \mid L_{\mathcal{D}}(A(S)) \geq 1/8\}) \geq 1/7)$$

$\Leftrightarrow \sim(L_{\mathcal{D}}(A(S)) \geq 1/8$ w/ prob. $\geq 1/7$ over choice of $S \sim \mathcal{D}^m)$

This contradicts w/ 2. of No-Free-Lunch Thm. #

Rmk:

By Cor 5.2, w/o using any prior knowledge will fail learning. We may impose prior knowledge via restricting $\mathcal{H}$.

5.2 Error Decomposition.

Q: How should we choose a good hypothesis class?

① $\mathcal{H}$ should be large enough to include a hypothesis w/ small $\min_{h\in\mathcal{H}} L_{\mathcal{D}}(h)$.

② $\mathcal{H}$ should be small enough to be learnable. (as suggested by No-Free-Lunch Thm.)

Def $\mathcal{H}$: a hypothesis class

Given $h_S \in ERM_{\mathcal{H}}(S)$.

Write $L_{\mathcal{D}}(h_S) = \varepsilon_{app} + \varepsilon_{est}$, where

$\varepsilon_{app} := \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h)$ and $\varepsilon_{est} := L_{\mathcal{D}}(h_S) - \varepsilon_{app}$.

ε_{app} is called the approximation error.

ε_{est} is called the estimation error.

Rmk:

① ε_{app} is the minimum risk achievable by $\mathcal{H}$.
i.e. it is exactly the inductive bias.
If $\mathcal{H}$: realizable, then $\varepsilon_{app} = 0$.
In agnostic case, ε_{app} may be large.

② ε_{est} results because $L_{\mathcal{D}}(h_S)$ is just an estimation of $\min_{h\in\mathcal{H}} L_{\mathcal{D}}(h)\ (= \varepsilon_{app})$.

Observation:

	ε_{app}	ε_{est}	$L_{\mathcal{D}}(h_S) = \varepsilon_{app} + \varepsilon_{est}$
$\mathcal{H}\uparrow$	$\downarrow$	? might $\uparrow$	?
$\mathcal{H}\downarrow$	$\uparrow$	? might $\downarrow$	?

Def

The tradeoff on $\mathcal{H}$ we face when trying to minimize $L_{\mathcal{D}}(h_S)$ is called the bias-complexity tradeoff.

Rmk:

Too large $\mathcal{H}$ may result in overfitting while
" small $\mathcal{H}$ may result " underfitting.

Chap 6 The VC Dimension.

Q: Which classes $\mathcal{H}$ are PAC learnable?

6.1 Infinite-Size Class Can Be Learnable.

Def

A function of the form $h_a: \mathbb{R} \to \{0,1\}$ w/ $h_a(x) = \mathbb{1}_{[x<a]}$ is called a threshold function.

Lemma 6.1

$\mathcal{H} := \{$all threshold functions$\}$.

Then $\mathcal{H}$ is PAC learnable, using ERM, w/ sample complexity $m_{\mathcal{H}}(\varepsilon,\delta) \leq \lceil \log(2/\delta)/\varepsilon \rceil$.

Rmk:

The "PAC learnable" here is the very first (non-agnostic) learnability w/ specific target f.

⟨Pf⟩ (Intuition).

Let S be a sample wrt a dist. $\mathcal{D}$ over $\mathcal{X}$,

$b_0 := \max\{x : (x,1) \in S\}$. ($\max(\emptyset) := -\infty$).

$b_1 := \min\{x : (x,0) \in S\}$. ($\min(\emptyset) = \infty$).

Then $\nexists$ b w/ $(b,0) \in S$ or $(b,1) \in S$, and

$h_{b_S} \in ERM_{\mathcal{H}}(S)$, where b_S is arbitrary in (b_0, b_1).

Assume $a^* \in \mathbb{R}$ w/ $L_{\mathcal{D}}(h_{a^*}) = 0$.

"Then", as $|S| = m \uparrow$, b_S is closer to a^* w/ high prob.

Therefore, $\mathcal{H}$ is PAC learnable. #

6.2 The VC-Dimension.

Motivation:

As suggested by L 6.1, while $|\mathcal{H}| < \infty$ is sufficient for learnability, it is NOT necessary.

Def 6.2, 6.3

$\mathcal{H}$: a hypothesis class of fun. from $\mathcal{X}$ to $\{0,1\}$.

$C = \{c_1, \cdots, c_m\} \subseteq \mathcal{X}$, i.e. C is a finite subset of $\mathcal{X}$.

① The restriction of $\mathcal{H}$ to C is the set

$$\mathcal{H}_C = \{f|_C : C \to \{0,1\} \mid f \in \mathcal{H}\}.$$

② $\mathcal{H}$ is said to shatter C if $\mathcal{H}_C$ is the set of all possible functions from C to $\{0,1\}$.

i.e. $|\mathcal{H}_C| = 2^{|C|}$.

Ex 6.2

$\mathcal{H} :=$ {all threshold functions}.

① $C = \{c_1\} \subseteq \mathbb{R}$.

Then $h_{c_1-1}(c_1) = 0$ and $h_{c_1+1}(c_1) = 1$.

Thus $\mathcal{H}$ shatters C.

② $C = \{c_1, c_2\}$ w/ $c_1 < c_2$.

Then $\nexists$ $h \in \mathcal{H}$ s.t. $h(c_1) = 0$ and $h(c_2) = 1$.

$\Rightarrow$ $\mathcal{H}$ does not shatter C. #

Cor 6.4 (of Proof of No-Free-Lunch Thm).

$\mathcal{H}$: a hypothesis class of fun. from $\mathcal{X}$ to $\{0,1\}$.

m: a training set size.

$\exists$ $C \subseteq \mathcal{X}$ of size $2m$ s.t. C is shattered by $\mathcal{H}$.

Then, $\forall$ algorithm A, $\exists$ dist. $\mathcal{D}$ on $\mathcal{X} \times \{0,1\}$ and $\exists$ $h \in \mathcal{H}$ s.t.

(1) $L_{\mathcal{D}}(h) = 0$, but

(2) $L_{\mathcal{D}}(A(S)) \geq 1/8$ w/ prob. $\geq 1/7$ (over the choice of $S \sim \mathcal{D}^m$).

Rmk:

Roughly speaking, Cor 6.4 tells us that if $\mathcal{H}$ shatters some set C of size $2m$, then we cannot learn $\mathcal{H}$ using m examples.

Def 6.5

The VC-dimension of a hypothesis class $\mathcal{H}$, denoted $VCdim(\mathcal{H})$ is the maximal size of a set $C \subseteq \mathcal{X}$ that can be shattered by $\mathcal{H}$.

i.e.

$VCdim(\mathcal{H}) = \max\{|C| : C \subseteq \mathcal{X}, \mathcal{H} \text{ shatters } C\}$.

$VCdim(\mathcal{H}) = \infty$ if $\forall$ $n \in \mathbb{N}$, $\exists$ C w/ $|C| = n$ s.t. $\mathcal{H}$ shatters C.

Rmk:

VC refers to Vapnik-Chervonenkis.

Thm 6.6

$\mathcal{H}$: a hypothesis class w/ $VCdim(\mathcal{H}) = \infty$.

Then $\mathcal{H}$ is not PAC learnable.

<Pf>.

For training size m, $\exists$ $C \subseteq \mathcal{X}$ w/ size $2m$ s.t. $\mathcal{H}$ shatters C.

By Cor 6.4, the result follows. #

Rmk:

We shall see later in this Chap that the converse is also true.

6.3 Examples.

6.3.1 Threshold fun.

$VCdim(\mathcal{H}) = 1$.

6.3.2 Intervals

$VCdim(\mathcal{H}) = 2$.

6.3.3 Axis-Aligned Rectangles (on $\mathbb{R}^2$).

$VCdim(\mathcal{H}) = 4$.

6.3.4 Finite Classes.

$\mathcal{H}$: a finite hypothesis class.

For $C \subseteq \mathcal{X}$, $|\{\text{all fun. } C \to \{0,1\}\}| = 2^{|C|}$.

Thus, if $2^{|C|} > |\mathcal{H}|$, then $\mathcal{H}$ cannot shatter C.

$\Rightarrow \mathrm{VCdim}(\mathcal{H}) \le \log_2(|\mathcal{H}|) < \infty$.

Rmk: ① For finite $\mathcal{H}$, $\mathrm{VCdim}(\mathcal{H})$ may be significantly smaller than $\log_2(|\mathcal{H}|)$.

② $\mathrm{VCdim}(\mathcal{H})$ is not necessarily the number of parameters defining $\mathcal{H}$. (see 6.3.5)

6.4 The Fundamental Theorem of PAC learning.

Thm 6.7 (The Fundamental Thm of Statistical Learning).

$\mathcal{H}$: a hypothesis class of fun. from $\mathcal{X}$ to $\{0,1\}$, w/ 0-1 loss function.

Then T.F.A.E.

(1) $\mathcal{H}$ has the UC property.

(2) ERM is a successful agnostic PAC learner for $\mathcal{H}$.

(3) $\mathcal{H}$: agnostic PAC learnable.

(4) $\mathcal{H}$: PAC learnable.

(5) ERM is a successful PAC learner for $\mathcal{H}$.

(6) $\mathrm{VCdim}(\mathcal{H}) < \infty$.

Thm 6.8 (The F.T. of S.L. - Quantitative Version).

$\mathcal{H}$: hypothesis class of fun. $\mathcal{X} \to \{0,1\}$. w/ 0-1 loss.

$\mathrm{VCdim}(\mathcal{H}) = d < \infty$.

Then $\exists$ abs. const. C_1, C_2 s.t.

(1) $\mathcal{H}$ has UC property w/

$$C_1 \frac{d + \log(1/\delta)}{\varepsilon^2} \le m_{\mathcal{H}}^{UC}(\varepsilon, \delta) \le C_2 \frac{d + \log(1/\delta)}{\varepsilon^2}.$$

(2) $\mathcal{H}$: agnostic PAC learnable w/

$$C_1 \frac{d + \log(1/\delta)}{\varepsilon^2} \le m_{\mathcal{H}}(\varepsilon, \delta) \le C_2 \frac{d + \log(1/\delta)}{\varepsilon^2}.$$

(3) $\mathcal{H}$: PAC learnable w/

$$C_1 \frac{d + \log(1/\delta)}{\varepsilon} \le m_{\mathcal{H}}(\varepsilon, \delta) \le C_2 \frac{d + \log(1/\delta)}{\varepsilon}.$$

<Pf> in Chap 28.

Rmk:

Similar F.T. holds for "some" other learning problems but "not all".

6.5 Proof of Thm 6.7.

So far, we have

$$\begin{tikzcd}
(1) \arrow[r, "C4"] & (2) \arrow[r, "trivial"] \arrow[dr, "trivial"'] & (3) \arrow[r, "trivial"] & (4) \arrow[d, "\text{Thm 6.6 (i.e. No-Free-Lunch)}"] \\
 & & (5) \arrow[r, "\text{Thm 6.6}"'] & (6)
\end{tikzcd}$$

We are going to prove $(6) \to (1)$.

6.5.1 Sauer's Lemma and Growth Function.

Def 6.9 $\mathcal{H}$: a hypothesis class.

The growth function of $\mathcal{H}$, denoted $\tau_{\mathcal{H}}: \mathbb{N} \to \mathbb{N}$, is

$$\tau_{\mathcal{H}}(m) := \max_{\substack{C \subseteq \mathcal{X} \\ |C| = m}} |\mathcal{H}_C|.$$

Rmk:

If $\mathrm{VCdim}(\mathcal{H}) = d$, then $\forall m \le \delta$, $\tau_{\mathcal{H}}(m) = 2^m$, exponential in m.

However, when $m > d$, we have $\tau_{\mathcal{H}}(m)$ polynomial in m:

Lemma 6.10 (Sauer-Shelah-Perles).

$\mathcal{H}$: hypothesis class w/ $\mathrm{VCdim}(\mathcal{H}) \le d < \infty$.

Then, $\forall m$,

$$\tau_{\mathcal{H}}(m) \le \sum_{i=0}^{d} \binom{m}{i}.$$

In particular, if $m > d+1$, then

$$\tau_{\mathcal{H}}(m) \le (em/d)^d.$$

End of W3