EE6550 Machine Learning

Lecture One – Part II
The PAC Learning Framework

Chung-Chin Lu

Department of Electrical Engineering

National Tsing Hua University

February 13, 2017

The Contents of This Lecture - Part II

- The PAC learning framework.
- Sample complexity, finite $\mathcal{H}$, consistent case.
- Sample complexity, finite $\mathcal{H}$, inconsistent case.

Fundamental Questions in Machine Learning

- What can be learned efficiently?
- What is inherently hard to learn?
- How many examples are needed to learn successfully?
- Is there a general model of learning?

What Will Be Learned? – Concept Class

- Input space $\mathscr{I}$: the population of all possible items.
  - $(\mathscr{I}, \mathcal{F}, P)$: a probability space associated with the population of all items, where the probability function P is usually unknown to the learner.
  - Example: $\mathscr{I} = \mathbb{R}^2$ is the set of all points in the plane. $\mathcal{F} = \mathcal{B}^2$ is the collection of all 2-dimensional Borel subsets of $\mathbb{R}^2$, including triangular areas, rectangular areas, disks, etc.
- Label space $\mathscr{Y}$: the set of all possible labels.
  - $(\mathscr{Y}, \mathcal{G})$: a measurable space associated with the label space $\mathscr{Y}$.
  - If $\mathscr{Y}$ is countable, $\mathcal{G}$ is commonly chosen to be $2^{\mathscr{Y}}$.
  - Example: $\mathscr{Y} = \{0, 1\}$ for binary classification and $2^{\mathscr{Y}} = \{\emptyset, \{0\}, \{1\}, \mathscr{Y}\}$.

- A concept $c : \mathscr{I} \to \mathscr{Y}$: a measurable function from the input space to the label space.
  - c is a $\mathscr{Y}$-valued random variable.
  - Example: Let R be an axis-aligned rectangular area in the plane, a member in $\mathcal{B}^2$. Define a concept

$$c(\omega) = \begin{cases} 1, & \text{if } \omega \in R, \\ 0, & \text{otherwise.} \end{cases}$$

    * c is the indicator of the rectangular area R, i.e., $c = I_R$.
    * The concept c to learn is the rectangular area R in the plane.

- Concept class $\mathcal{C}$: a set of concepts we may wish to learn.
  - Example: $\mathcal{C}$ = the set of concepts of all axis-aligned rectangular areas in the plane.

Generalization Error or Risk

- c: a fixed but unknown target concept in the concept class $\mathcal{C}$.
- $f : \mathscr{I} \to \mathscr{Y}'$: an arbitrary measurable function from the input space to the output space to approximate the concept c.
  - $(\mathscr{Y}', \mathcal{G}')$: a measurable space associated with the output space $\mathscr{Y}'$.
  - If $\mathscr{Y}'$ is countable, $\mathcal{G}'$ is commonly chosen to be $2^{\mathscr{Y}'}$.
  - f is a $\mathscr{Y}'$-valued random variable.

The generalization error (or risk) or true error of an approximation f to the concept c is defined as

$$R(f) \triangleq \mathop{E}_{\omega \sim P}[L(f(\omega), c(\omega))].$$

- Assume that the loss function $L : \mathscr{Y}' \times \mathscr{Y} \to \mathbb{R}$ is measurable, i.e., $L^{-1}(I) = \{(y', y) \in \mathscr{Y}' \times \mathscr{Y} \mid L(y', y) \in I\}$ is a member of the product σ-algebra $\mathcal{G}' \times \mathcal{G}$ for every interval I in $\mathbb{R}$.
- As a measurable function of r.v.s $f(\omega)$ and $c(\omega)$, $L(f(\omega), c(\omega))$ is a random variable.
- Both the probability function P and the target concept c are unknown.
- $R(f)$ is not directly accessible to the learner.
- Example: $L(y', y) = 1_{y' \neq y}$ so that

$$R(f) = \mathop{E}_{\omega \sim P}[L(f(\omega), c(\omega))] = \mathop{E}_{\omega \sim P}[1_{f(\omega) \neq c(\omega)}] = P(f(\omega) \neq c(\omega)).$$

Bayes Error

- c: a fixed but unknown target concept in the concept class $\mathcal{C}$.

The Bayes error of learning the concept c is the least possible generalization error to learn c,

$$R^* \triangleq \inf_{f \text{ is a } \mathscr{Y}'\text{-valued r.v.}} R(f).$$

- In general, R^* is not accessible to the learner.
- If $\mathscr{Y}' = \mathscr{Y}$ and $L(y, y) = 0$ for all labels y, then $R^* = 0$.
- A hypothesis h with $R(h) = R^*$ is called a Bayes hypothesis.

Best-In-Class Hypotheses

- c: a fixed but unknown target concept in the concept class $\mathcal{C}$.
- $\mathcal{H}$: the hypothesis set chosen.
  - A hypothesis h in $\mathcal{H}$ is a $\mathscr{Y}'$-valued random variable.
- $R^*_{\mathcal{H}} \triangleq \min_{h\in\mathcal{H}} R(h)$: the least generalization error w.r.t. c achievable by some hypotheses in the hypothesis set $\mathcal{H}$.

A hypothesis h^* in $\mathcal{H}$ is called best-in-class w.r.t. c if

$$R(h^*) = R^*_{\mathcal{H}}.$$

- In general, $R^*_{\mathcal{H}}$ and h^* are not accessible to the learner.
- If $\mathcal{H} = \mathcal{C}$ and $L(y, y) = 0$ for all labels y, then $R^*_{\mathcal{H}} = R^* = 0$ and $h^* = c$ is a best-in-class hypothesis w.r.t. c.

ϵ-Best-In-Class Hypotheses

- c: a fixed but unknown target concept in the concept class $\mathcal{C}$.
- $\mathcal{H}$: the hypothesis set chosen.
  - A hypothesis h in $\mathcal{H}$ is a $\mathscr{Y}'$-valued random variable.
- $R^*_{\mathcal{H}} \triangleq \inf_{h \in \mathcal{H}} R(h)$: the least generalization error w.r.t. c asymptotically achievable by hypotheses in the hypothesis set $\mathcal{H}$.

A hypothesis h^*_ϵ in $\mathcal{H}$ is called ϵ-best-in-class w.r.t. c if

$$|R(h^*_\epsilon) - R^*_{\mathcal{H}}| \leq \epsilon.$$

- In general, $R^*_{\mathcal{H}}$ and h^*_ϵ are not accessible to the learner.

Estimation and Approximation

- c: a fixed but unknown target concept in the concept class $\mathcal{C}$.
- R^*: the Bayes error of learning the concept c.
- $\mathcal{H}$: the hypothesis set chosen.
- h^*: a best-in-class hypothesis in $\mathcal{H}$.
- h: a hypothesis in $\mathcal{H}$.

The difference of the true error of a hypothesis h from the Bayes error R^* of learning the concept c is

$$R(h) - R^* = \underbrace{R(h) - R(h^*)}_{\text{Estimation}} + \underbrace{R(h^*) - R^*}_{\text{Approximation}}.$$

- The approximation part only depends on $\mathcal{H}$.
- The estimation part is where we can hope to bound.

Formulation of Learning Problem

- c: a fixed but unknown target concept in the concept class $\mathcal{C}$.
- $\mathcal{H}$: the hypothesis set.
- $S = (\omega_1, \ldots, \omega_m)$: a sample of m items, drawn i.i.d. from the population according to P, with labels $(c(\omega_1), \ldots, c(\omega_m))$.

To learn the concept c from the labeled sample S, the learner's task is to use the labeled sample S to select a hypothesis h_S in the hypothesis set $\mathcal{H}$ that has a "small" generalization error with respect to the concept c and then is a "good" approximation to c.

- But the learner does not know how far the true error $R(h_S)$ is from the least generalization error $R^*_{\mathcal{H}}$ over $\mathcal{H}$.

Empirical Error

- c: a fixed but unknown target concept in the concept class $\mathcal{C}$.
- $S = (\omega_1, \ldots, \omega_m)$: a sample of m items, drawn i.i.d. from the population according to P, with labels $(c(\omega_1), \ldots, c(\omega_m))$.
- h: an arbitrary hypothesis in the hypothesis set $\mathcal{H}$.

The empirical error or risk of a hypothesis h w.r.t. the concept c on the labeled sample S is defined as

$$\hat{R}_S(h) \triangleq \frac{1}{m} \sum_{i=1}^{m} L(h(\omega_i), c(\omega_i)).$$

- The learner can measure the empirical error of a hypothesis w.r.t. the unknown concept on the labeled sample.

The Sample Space Ω_m of Size m

- The sample space Ω_m of size m: the set of all samples $S = (\omega_1, \ldots, \omega_m)$ of m items from the population $\mathscr{I}$.
- The σ-algebra $\mathcal{F}_m$: the product $\underbrace{\mathcal{F} \times \cdots \times \mathcal{F}}_{m \text{ times}}$ of m copies of the σ-algebra $\mathcal{F}$.
- The probability function P_m: the product $\underbrace{P \times \cdots \times P}_{m \text{ times}}$ of m copies of the probability function P, i.e.,

$$P_m(E_1 \times \cdots \times E_m) = P(E_1) \cdots P(E_m)$$

for all members $E_1, \ldots, E_m$ in $\mathcal{F}$.

Projections ϕ_i

- $\phi_i : \Omega_m \to \mathscr{I}$: the ith projection function from the sample space to the input space, defined as

$$\phi_i(S) = \phi_i((\omega_1, \ldots, \omega_m)) = \omega_i$$

  for all sample $S = (\omega_1, \ldots, \omega_m) \in \Omega_m$ and for all $1 \leq i \leq m$.

- ϕ_i is measurable and then is an $\mathscr{I}$-valued random variable.

$\phi_1, \phi_2, \ldots, \phi_m$ Are I.I.D. R.V.s

Proof. Let $E_1, \ldots, E_m$ be members in $\mathcal{F}$. Since

$$(\phi_i \in E_i) = \phi_i^{-1}(E_i) = \mathscr{I} \times \cdots \times E_i \times \cdots \times \mathscr{I},$$

the joint event $(\phi_1 \in E_1, \phi_2 \in E_2, \ldots, \phi_m \in E_m)$ is

$$\phi_1^{-1}(E_1) \cap \phi_2^{-1}(E_2) \cap \cdots \cap \phi_m^{-1}(E_m) = E_1 \times E_2 \times \cdots \times E_m$$

so that

$$\begin{aligned}
& P_m(\phi_1 \in E_1, \phi_2 \in E_2, \ldots, \phi_m \in E_m) \\
= \quad & P_m(E_1 \times E_2 \times \cdots \times E_m) \\
= \quad & P(E_1) \cdot P(E_2) \cdots P(E_m) \\
= \quad & P_m(E_1 \times \mathscr{I} \times \cdots \times \mathscr{I}) \cdot P_m(\mathscr{I} \times E_2 \times \cdots \times \mathscr{I}) \\
& \cdots P_m(\mathscr{I} \times \mathscr{I} \times \cdots \times E_m) \\
= \quad & P_m(\phi_1 \in E_1) \cdot P_m(\phi_2 \in E_2) \ldots P_m(\phi_m \in E_m).
\end{aligned}$$

Thus $\phi_1, \phi_2, \ldots, \phi_m$ are statistically independent. For any E in $\mathcal{F}$,

$$P_m(\phi_i \in E) = P_m(\mathscr{I} \times \cdots \times E \times \cdots \times \mathscr{I}) = P(E)$$

so that ϕ_i's are identically distributed. $\square$

- The probability distributions of the projections ϕ_i's are the same as P.

$\hat{R}_S(h)$ Is a Random Variable

- c: a fixed but unknown target concept in the concept class $\mathcal{C}$.
- $S = (\omega_1, \ldots, \omega_m)$: a sample of m items, drawn i.i.d. from the population according to P, with labels $(c(\omega_1), \ldots, c(\omega_m))$.
- h: an arbitrary hypothesis in the hypothesis set $\mathcal{H}$.
- $\phi_i(S) = \phi_i((\omega_1, \ldots, \omega_m)) = \omega_i$: the ith projection function.
- $h(\omega_i) \triangleq h(\phi_i(S))$, $c(\omega_i) \triangleq c(\phi_i(S))$: measurable functions from the sample space to the output space.

The empirical error of h w.r.t. c on a labeled sample S

$$\hat{R}_S(h) = \frac{1}{m} \sum_{i=1}^{m} L(h(\omega_i), c(\omega_i)) = \frac{1}{m} \sum_{i=1}^{m} L(h(\phi_i(S)), c(\phi_i(S)))$$

is a measurable function from Ω_m to $\mathbb{R}$, i.e., a random variable.

$$\boxed{\underset{S \sim P_m}{E}[\hat{R}_S(h)] = R(h)}$$

- The expectation of empirical error of a hypothesis h w.r.t. the target concept c on a labeled sample S of size m is equal to the generalization error of h w.r.t. the target concept c.
- Observation: since r.v.'s ϕ_i have the same probability distribution P, r.v.'s $h(\phi_i(S))$ $(c(\phi_i(S)))$ have the same probability distribution as the r.v. $h(\omega)$ $(c(\omega))$.

Proof.

$$
\begin{aligned}
& \mathop{E}_{S\sim P_m}[\hat{R}_S(h)] \\
= \quad & \mathop{E}_{S\sim P_m}\left[\frac{1}{m}\sum_{i=1}^{m} L(h(\phi_i(S)), c(\phi_i(S)))\right] \\
= \quad & \frac{1}{m}\sum_{i=1}^{m} \mathop{E}_{S\sim P_m}[L(h(\phi_i(S)), c(\phi_i(S)))] \\
= \quad & \frac{1}{m}\sum_{i=1}^{m} \mathop{E}_{\omega\sim P}[L(h(\omega), c(\omega))] \\
& \text{since } h(\phi_i(S))\text{'s } (c(\phi_i(S))\text{'s}) \text{ have the same probability} \\
& \text{distribution as } h(\omega)\ (c(\omega)) \\
= \quad & \mathop{E}_{\omega\sim P}[L(h(\omega), c(\omega))] = R(h).
\end{aligned}
$$

$\square$

Empirical Risk Minimization (ERM)

- c: a fixed but unknown target concept in the concept class $\mathcal{C}$.
- $S = (\omega_1, \ldots, \omega_m)$: a sample of m items, drawn i.i.d. from the population according to P, with labels $(c(\omega_1), \ldots, c(\omega_m))$.
- $\mathcal{H}$: the hypothesis set.

The learner will return a hypothesis among all hypotheses in $\mathcal{H}$ which minimizes the empirical error,

$$h_S = \arg\min_{h \in \mathcal{H}} \hat{R}(h).$$

- Overfitting may occur, i.e., h_S matches to the training data sample S too well so that it may have large generalization error.
  - The hypothesis set $\mathcal{H}$ may be too complex.
  - The sample size may not be large enough.

Structural Risk Minimization (SRM)

- c: a fixed but unknown target concept in the concept class $\mathcal{C}$.
- $S = (\omega_1, \ldots, \omega_m)$: a sample of m items, drawn i.i.d. from the population according to P, with labels $(c(\omega_1), \ldots, c(\omega_m))$.
- $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \cdots \subseteq \mathcal{H}_n \subseteq \cdots$: an increasing sequence of hypothesis sets.

The learner will return a hypothesis among all hypotheses in $\cup_{n=1}^{\infty} \mathcal{H}_n$ which minimizes the empirical error plus a complexity measure of $\mathcal{H}_n$ and the sample size m,

$$h_S = \arg \min_{h \in \mathcal{H}_n, n \in \mathbb{N}} [\hat{R}(h) + \text{complexity}(\mathcal{H}_n, m)].$$

- Theoretical guarantees: consistency under general assumptions.
- Computational complexity: typically hard problems.

Probably Approximately Correct (PAC) Learning

- Definition: A concept class $\mathcal{C}$ is PAC-learnable if there exists a learning algorithm $\mathbb{A}$, which returns $h_S \in \mathcal{H}$ to approximate an unknown target concept $c \in C$ on a labeled sample S of size m,

  $$h_S = \mathbb{A}(S; c, \mathcal{H}),$$

  such that for any $\epsilon > 0$, $\delta > 0$, $c \in \mathcal{C}$ and P, we have

  $$P_m(R(h_S) \leq \epsilon) \geq 1 - \delta,$$

  provided that the sample size m is

  $$m \geq \text{poly}(1/\epsilon, 1/\delta, n, \text{size}(c))$$

  for a fixed polynomial, where

  – $O(n)$: cost of computational representation of an item ω.

  – $O(\text{size}(c))$: cost of computational representation of a c.

- When such an algorithm $\mathbb{A}$ exists, it is called a PAC-learning algorithm for $\mathcal{C}$.

Efficient PAC Learning

- Definition: A concept class $\mathcal{C}$ is efficiently PAC-learnable if
  - $\mathcal{C}$ is PAC-learnable by a learning algorithm $\mathbb{A}$,
  - $\mathbb{A}$ further runs in $\text{poly}(1/\epsilon, 1/\delta, n, \text{size}(c))$.
- When such an algorithm $\mathbb{A}$ exists, it is called an efficient PAC-learning algorithm for $\mathcal{C}$.

Remarks

- Concept class $\mathcal{C}$ is known to the algorithm $\mathbb{A}$.
- But a specific target concept $c \in \mathcal{C}$ is unknown to $\mathbb{A}$.
- Hypothesis set $\mathcal{H}$ is built in the algorithm $\mathbb{A}$.
- Distribution-free model: no assumption on the probability function P.
- Both training and test samples are drawn i.i.d. from the population according to P, which is unknown to $\mathbb{A}$.
- The mapping $S \mapsto R(h_S)$ is measurable so that $R(h_S)$ is a random variable.
- High probable: at least $1 - \delta$.
- Approximately correct: true error at most ϵ.

Example 2.1: Learning Axis-Aligned Rectangular Areas

- Problem: learn with small error an unknown axis-aligned rectangular area R using as small a labeled training sample as possible.
- Input space $\mathscr{I} = \mathbb{R}^2$, the plane.
- Label space $\mathscr{Y} = \{0, 1\}$.
- Concept class $\mathcal{C}$ = the set of all axis-aligned rectangular area in the plane.
- We will show that this concept class $\mathcal{C}$ is PAC-learnable.

Example 2.1: A Learning Algorithm $\mathbb{A}$

- $R \in \mathcal{C}$: an unknown target axis-aligned rectangular area to learn.
- $S = (\omega_1, \ldots, \omega_m)$: a labeled sample of size m.
- The hypothesis set is $\mathcal{H} = \mathcal{C} =$ the set of all axis-aligned rectangular area.
- $R'_S = \mathbb{A}(S; R, \mathcal{H}) =$ the tightest axis-aligned rectangular area containing the points in the sample S labeled with 1.

Example 2.1: Error Analysis (1)

- The loss function is $L(y', y) = 1_{(y' \neq y)},\ \forall\, y', y \in \{0, 1\}$.
- The generalization error of a hypothesis R' w.r.t. a concept R is

$$R(R') = \mathop{E}_{\omega \sim P}[1_{(1_{R'}(\omega) \neq 1_R(\omega))}] = \mathop{E}_{\omega \sim P}[1_{R' \Delta R}(\omega)] = P(R' \Delta R),$$

  - $R' \Delta R \triangleq (R' \setminus R) \cup (R \setminus R')$: the symmetric difference of two events R' and R.
  - A point $\omega \in R' \setminus R$ will make a false positive.
  - A point $\omega \in R \setminus R'$ will make a false negative.
- Since $R'_S \subseteq R$, the error region $R'_S \Delta R = R \setminus R'_S$ is included in R and R'_S does not produce any false positive.
- $R(R'_S) = P(R'_S \Delta R) = P(R \setminus R'_S) = P(R) - P(R'_S)$.

Example 2.1: Error Analysis (2)

- The self-empirical error is
$\hat{R}_S(R'_S) = \frac{1}{m} \sum_{i=1}^{m} 1_{1_{R'_S}(\omega_i) \neq 1_R(\omega_i)} = 0.$
- With zero self-empirical error for all labeled sample S, both the hypothesis R'_S and the learning algorithm $\mathbb{A}$ are called consistent.
- If $P(R) \leq \epsilon$, then the generalization error
$R(R'_S) = P(R) - P(R'_S) \leq P(R) \leq \epsilon$ for all labeled sample S.
- Assume $P(R) > \epsilon$. Let r_1, r_2, r_3, r_4 be the four smallest sub-rectangular areas of R along the four sides of R such that $P(r_i) = \frac{\epsilon}{4}$.
- That the event $(R(R'_S) > \epsilon) = (P(R) - P(R'_S) > \epsilon)$ occurs implies that R'_S misses at least one of four r_i's.

Example 2.1: Error Analysis (3)

- Thus we have

$$
\begin{aligned}
P_m(R(R'_S) > \epsilon) &\leq P_m(\cup_{i=1}^4 (R'_S \cap r_i = \emptyset)) \\
&\leq \sum_{i=1}^4 P_m(R'_S \cap r_i = \emptyset) \text{ by the union bound} \\
&\leq 4(1 - \epsilon/4)^m \\
&< 4e^{-m\epsilon/4} \text{ by } 1 - x < e^{-x} \text{ for all } x \in \mathbb{R} \setminus \{0\}
\end{aligned}
$$

- Set $4e^{-m\epsilon/4} \leq \delta$ if and only if set $m \geq \frac{4}{\epsilon} \ln \frac{4}{\delta}$.
- For any $\epsilon > 0$, $\delta > 0$, $R \in \mathcal{C}$ and P, if $m \geq \frac{4}{\epsilon} \ln \frac{4}{\delta}$, we have

$$P_m(R(R'_S) > \epsilon) < \delta.$$

Example 2.1: PAC-Learnability

- The concept class $\mathcal{C}$ of axis-aligned rectangular areas is PAC-learnable.
- $\mathbb{A}$ is a PAC-learning algprithm.
- The sample complexity of PAC-learning axis-aligned rectangular areas is in $O(\frac{4}{\epsilon} \ln \frac{4}{\delta})$.
- An equivalent statement: with probability at least $1 - \delta$ and a sample size m, the generalization error of the PAC-learning algorithm is upper bounded as:

$$R(R'_S) \leq \frac{4}{m} \ln \frac{4}{\delta}$$

by setting $\delta = 4e^{-m\epsilon/4}$ and solving ϵ.

The Contents of This Lecture - Part II

- The PAC learning framework.
- Sample complexity, finite $\mathcal{H}$, consistent case.
- Sample complexity, finite $\mathcal{H}$, inconsistent case.

Learning Bound for Finite $\mathcal{H}$ - Consistent Case

Theorem 2.1: Let

- $\mathscr{I}$: input space, which is general.
- $\mathscr{Y} = \{0, 1\}$: label space with loss function $L(y', y) = 1_{y' \neq y}$.
- $\mathcal{H} = \mathcal{C}$: finite hypothesis set and concept class.
- $\mathbb{A}$: consistent learning algorithm.
  - $h_S = \mathbb{A}(S; c, \mathcal{H})$ is consistent for any i.i.d. sample S of size m and any target concept c, i.e., $\hat{R}_S(h_S) = 0$.

Then for any $\epsilon > 0, \delta > 0$, we have

$$P_m(R(h_S) \leq \epsilon) \geq 1 - \delta,$$

provided that

$$m \geq \frac{1}{\epsilon}\left(\ln |\mathcal{H}| + \ln \frac{1}{\delta}\right).$$

Proof. Since $\hat{R}_S(h_S) = 0$ for every returned hypothesis h_S, the event $(R(h_S) > \epsilon) = (R(h_S) > \epsilon, \hat{R}_S(h_S) = 0)$ implies the event that there exists a hypothesis $h \in \mathcal{H}$ with $R(h) > \epsilon$ such that $\hat{R}_S(h) = 0$, i.e., $\cup_{h \in \mathcal{H} \text{ with } R(h) > \epsilon}(\hat{R}_S(h) = 0)$. By union bound, we have

$$
\begin{aligned}
& P_m(R(h_S) > \epsilon) \\
\leq \;& P_m(\cup_{h\in\mathcal{H} \text{ with } R(h)>\epsilon}(\hat{R}_S(h) = 0)) \\
\leq \;& \sum_{h\in\mathcal{H} \text{ with } P(h(\omega)\neq c(\omega))>\epsilon} P_m\left(\frac{1}{m}\sum_{i=1}^{m} 1_{h(\omega_i)\neq c(\omega_i)} = 0\right) \\
= \;& \sum_{h\in\mathcal{H} \text{ with } P(h(\omega)\neq c(\omega))>\epsilon} P_m(\cap_{i=1}^{m}(h(\phi_i(S)) = c(\phi_i(S)))) \\
= \;& \sum_{h\in\mathcal{H} \text{ with } P(h(\omega)\neq c(\omega))>\epsilon} \prod_{i=1}^{m} P_m(h(\phi_i(S)) = c(\phi_i(S))) \\
& \text{ since } \phi_i\text{'s are statistically independent} \\
= \;& \sum_{h\in\mathcal{H} \text{ with } P(h(\omega)\neq c(\omega))>\epsilon} \prod_{i=1}^{m} P(h(\omega) = c(\omega)) \\
& \text{ since } \phi_i(S)\text{'s are identically distributed with } \omega \\
< \;& |\mathcal{H}|(1-\epsilon)^m \leq |\mathcal{H}|e^{-m\epsilon}.
\end{aligned}
$$

By setting

$$\delta \geq |\mathcal{H}| e^{-m\epsilon},$$

we have

$$m \geq \frac{1}{\epsilon}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right).$$

$\square$

Remarks

- The theorem shows that when the hypothesis set $\mathcal{H}$ is finite, a consistent algorithm $\mathbb{A}$ is a PAC-learning algorithm.
- Equivalently, with probability at least $1 - \delta$ and sample size m, the true error of the returned hypothesis h_S is upper bounded as:
$$R(h_S) \leq \frac{1}{m}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right).$$
- True error bound is linear in $1/m$ and only logarithmic in $1/\delta$.
- The price to pay for coming up with a consistent algorithm is the use of a larger hypothesis set H containing target concepts.
- $\log_2|\mathcal{H}|$ is the number of bits used for the representation of $\mathcal{H}$.
- Bound is loose for large $\mathcal{H}$.

The Contents of This Lecture - Part II

- The PAC learning framework.
- Sample complexity, finite $\mathcal{H}$, consistent case.
- Sample complexity, finite $\mathcal{H}$, inconsistent case.

A Relation Between True Error And Empirical Error

Corollary 2.1: Let

- $c : \mathscr{I} \to \{0, 1\}$: a fixed but unknown target concept.
- $h : \mathscr{I} \to \{0, 1\}$: an arbitrary hypothesis.
- $S = (\omega_1, \ldots, \omega_m)$: a sample drawn i.i.d. from the population $\mathscr{I}$.

For any $\epsilon > 0$,

$$\begin{aligned} P_m(\hat{R}_S(h) - R(h) > \epsilon) &< e^{-2m\epsilon^2}, \\ P_m(\hat{R}_S(h) - R(h) < -\epsilon) &< e^{-2m\epsilon^2}. \end{aligned}$$

And by union bound,

$$P_m(|\hat{R}_S(h) - R(h)| > \epsilon) < 2e^{-2m\epsilon^2}.$$

Proof. This result follows immediately from Hoeffding's inequality.
□

Generalization Bound - Single Hypothesis

Corollary 2.2: Let

- $c : \mathscr{I} \to \{0, 1\}$: a fixed but unknown target concept.
- $h : \mathscr{I} \to \{0, 1\}$: an arbitrary hypothesis.
- $S = (\omega_1, \ldots, \omega_m)$: a sample of size m drawn i.i.d. from the population $\mathscr{I}$.

For any $\delta > 0$, with probability at leat $1 - \delta$,

$$R(h) \leq \hat{R}_S(h) + \sqrt{\frac{\ln \frac{2}{\delta}}{2m}}.$$

Proof. Setting $\delta = 2e^{-2m\epsilon^2}$ and solving $\epsilon = \sqrt{\frac{\ln \frac{2}{\delta}}{2m}}$ in Corollary 2.1, we have

$$P_m\left(|R(h) - \hat{R}_S(h)| > \sqrt{\frac{\ln \frac{2}{\delta}}{2m}}\right) < \delta.$$

Thus with probability at least $1 - \delta$,

$$|R(h) - \hat{R}_S(h)| \leq \sqrt{\frac{\ln \frac{2}{\delta}}{2m}},$$

which implies

$$R(h) \leq \hat{R}_S(h) + \sqrt{\frac{\ln \frac{2}{\delta}}{2m}}.$$

□

Applicable to Learning Algorithm?

- Can we apply that bound to the hypothesis h_S returned by a learning algorithm when training on an i.i.d. sample S ?
- No, because h_S is a random hypothesis, depending on the training sample S.
- Note also that the generalization error $R(h_S)$ of the returned hypothesis h_S is a random variable.
- We need a bound that holds simultaneously for all hypotheses, a uniform generalization bound.

Uniform Generalization Bound - Finite Hypothesis Set

Theorem 2.2: Let

- $c : \mathscr{I} \to \{0, 1\}$: a fixed but unknown target concept.
- $\mathcal{H}$: the hypothesis set, consisting of finitely many hypotheses $h : \mathscr{I} \to \{0, 1\}$.
- $S = (\omega_1, \ldots, \omega_m)$: a sample of size m drawn i.i.d. from the population $\mathscr{I}$.

For any $\delta > 0$, with probability at leat $1 - \delta$,

$$\forall\ h \in \mathcal{H}, \quad R(h) \leq \hat{R}_S(h) + \sqrt{\frac{\ln|\mathcal{H}| + \ln\frac{2}{\delta}}{2m}}.$$

Proof. For any $\epsilon > 0$,

$$
\begin{aligned}
& P_m(\max_{h\in\mathcal{H}} |R(h) - \hat{R}_S(h)| > \epsilon) \\
= \quad & P_m(\cup_{h\in\mathcal{H}}(|R(h) - \hat{R}_S(h)| > \epsilon)) \\
\leq \quad & \sum_{h\in\mathcal{H}} P_m(|R(h) - \hat{R}_S(h)| > \epsilon) \text{ by union bound} \\
< \quad & 2|\mathcal{H}|e^{-2m\epsilon^2} \text{ by Corollary 2.1.}
\end{aligned}
$$

Setting $\delta = 2|\mathcal{H}|e^{-2m\epsilon^2}$ and solving $\epsilon = \sqrt{\frac{\ln|\mathcal{H}|+\ln\frac{2}{\delta}}{2m}}$, we have

$$P_m\left(\max_{h\in\mathcal{H}} |R(h) - \hat{R}_S(h)| > \sqrt{\frac{\ln|\mathcal{H}| + \ln\frac{2}{\delta}}{2m}}\right) < \delta.$$

Thus with probability at least $1 - \delta$,

$$\forall\ h \in \mathcal{H},\ \ |R(h) - \hat{R}_S(h)| \leq \sqrt{\frac{\ln|\mathcal{H}| + \ln\frac{2}{\delta}}{2m}},$$

which implies

$$\forall\ h \in \mathcal{H}, \quad R(h) \leq \hat{R}_S(h) + \sqrt{\frac{\ln |\mathcal{H}| + \ln \frac{2}{\delta}}{2m}}.$$

$\square$

Remarks

- Equivalently, for any $\epsilon > 0, \delta > 0$,

$$P_m(\max_{h \in \mathcal{H}} |R(h) - \hat{R}_S(h)| \leq \epsilon) \geq 1 - \delta,$$

  provided that the sample size $m \geq \frac{1}{2\epsilon^2}\left(\ln|\mathcal{H}| + \ln\frac{2}{\delta}\right)$.

- The uniform generalization bound $\hat{R}_S(h) + \sqrt{\frac{\ln|\mathcal{H}| + \ln\frac{2}{\delta}}{2m}}$ suggests seeking a trade-off between reducing the empirical error versus controlling the size of the hypothesis set.
  - A larger hypothesis set is penalized by the second term but could help reduce the empirical error, that is the first term.
  - Occam's Razor principle (law of parsimony): the simplest explanation is best. Thus if all other things being equal (a similar empirical error), a simpler (smaller) hypothesis set is better.

- The uniform generalization bound is in $O(\sqrt{\frac{\ln|\mathcal{H}|}{m}})$, not in $O(\frac{\ln|\mathcal{H}|}{m})$.

Agnostic PAC-Learning

- Definition: A concept class $\mathcal{C}$ is agnostically PAC-learnable if there exists a learning algorithm $\mathbb{A}$, which returns $h_S \in \mathcal{H}$ to approximate an unknown target concept $c \in C$ on a labeled sample S of size m,

$$h_S = \mathbb{A}(S; c, \mathcal{H}),$$

such that for any $\epsilon > 0$, $\delta > 0$, $c \in \mathcal{C}$ and P, we have

$$P_m(R(h_S) - R_H^* \leq \epsilon) \geq 1 - \delta,$$

provided that the sample size m is

$$m \geq \text{poly}(1/\epsilon, 1/\delta, n, \text{size}(c))$$

for a fixed polynomial, where

  - $O(n)$: cost of computational representation of an item ω.

- $O(\text{size}(c))$: cost of computational representation of a c.

- When such an algorithm $\mathbb{A}$ exists, it is called an agnostic PAC-learning algorithm for $\mathcal{C}$.

Efficient Agnostic PAC-Learning

- Definition: A concept class $\mathcal{C}$ is efficiently agnostically PAC-learnable if
  - $\mathcal{C}$ is agnostically PAC-learnable by a learning algorithm $\mathbb{A}$,
  - $\mathbb{A}$ further runs in $\mathrm{poly}(1/\epsilon, 1/\delta, n, \mathrm{size}(c))$.
- When such an algorithm $\mathbb{A}$ exists, it is called an efficient agnostic PAC-learning algorithm for $\mathcal{C}$.

The Empirical Risk Minimization Algorithm $\mathbb{A}^{ERM}$

- $h_S^{ERM} = \mathbb{A}^{ERM}(S; c, \mathcal{H}) = \arg\min_{h\in\mathcal{H}} \hat{R}_S(h)$.

- The estimation error is

$$
\begin{aligned}
R(h_S^{ERM}) - R_H^* &= R(h_S^{ERM}) - \hat{R}_S(h_S^{ERM}) + \hat{R}_S(h_S^{ERM}) - R_H^* \\
&\leq R(h_S^{ERM}) - \hat{R}_S(h_S^{ERM}) + \hat{R}_S(h^*) - R(h^*) \\
&\leq 2 \sup_{h\in\mathcal{H}} |R(h) - \hat{R}_S(h)|.
\end{aligned}
$$

- Application of the uniform generalization bound in Theorem 2.2.

- The ERM algorithm $\mathbb{A}^{ERM}$ with a finite hypothesis set $\mathcal{H}$ is an agnostic PAC-learning algorithm for any concept class $\mathcal{C}$.