An Introduction to Statistical Theory of Learning

Nakul Verma
Janelia, HHMI
Towards formalizing ‘learning’

What does it mean to **learn** a concept?

- Gain knowledge or experience of the concept.

The basic process of **learning**

- Observe a phenomenon
- Construct a model from observations
- Use that model to make decisions / predictions

How can we make this more precise?
A statistical machinery for learning

Phenomenon of interest:

Input space: \( X \)  
Output space: \( Y \)

There is an unknown distribution \( \mathcal{D} \) over \( (X \times Y) \)

The learner observes \( m \) examples \((x_1, y_1), \ldots, (x_m, y_m)\) drawn from \( \mathcal{D} \)

Construct a model:

Let \( \mathcal{F} \) be a collection of models, where each \( f : X \rightarrow Y \) predicts \( y \) given \( x \)

From \( m \) observations, select a model \( f_m \in \mathcal{F} \) which predicts well.

\[
\text{err}(f) := \mathbb{P}_{(x,y) \sim \mathcal{D}}[f(x) \neq y] \quad \text{(generalization error of } f) 
\]

We can say that we have learned the phenomenon if

\[
\text{err}(f_m) - \text{err}(f^*) \leq \epsilon \quad f^* := \arg \inf_{f \in \mathcal{F}} \text{err}(f)
\]

for any tolerance level \( \epsilon > 0 \) of our choice.
PAC Learning

For all tolerance levels $\epsilon > 0$, and all confidence levels $\delta > 0$, if there exists some model selection algorithm $\mathcal{A}$ that selects $f_m^\mathcal{A} \in \mathcal{F}$ from $m$ observations $\mathcal{A} : (x_i, y_i)_{i=1}^m \mapsto f_m^\mathcal{A}$, and has the property:

with probability at least $1 - \delta$ over the draw of the sample,

$$\text{err}(f_m^\mathcal{A}) - \text{err}(f^*) \leq \epsilon$$

We call

- The model class $\mathcal{F}$ is PAC-learnable.
- If the $m$ is polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$, then $\mathcal{F}$ is efficiently PAC-learnable

A popular algorithm:

Empirical risk minimizer (ERM) algorithm

$$f_m^{\text{ERM}} := \arg\inf_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m 1\{f(x_i) \neq y_i\}$$
PAC learning simple model classes

**Theorem (finite size $\mathcal{F}$):**

Pick any tolerance level $\epsilon > 0$, and any confidence level $\delta > 0$

Let $(x_1, y_1), \ldots, (x_m, y_m)$ be $m$ examples drawn from an unknown $\mathcal{D}$

If $m \geq C \cdot \frac{1}{\epsilon^2} \ln \frac{|\mathcal{F}|}{\delta}$, then with probability at least $1 - \delta$

$$\text{err}(f^\text{ERM}_m) - \text{err}(f^*) \leq \epsilon$$

$\mathcal{F}$ is efficiently PAC learnable

**Occam’s Razor Principle:**

All things being equal, usually the simplest explanation of a phenomenon is a good hypothesis.

Simplicity = representational succinctness
Proof sketch

Define:

\[ \text{err}(f) := \mathbb{E}_{(x,y) \sim D} \left[ 1\{f(x) \neq y\} \right] \]

\[ \text{err}_m(f) := \frac{1}{m} \sum_{i=1}^{m} \left[ 1\{f(x_i) \neq y_i\} \right] \]

(generalization error of \( f \))

(sample error of \( f \))

We need to analyze:

\[
\text{err}(f_{\text{ERM}}^m) - \text{err}(f^*) \\
= \text{err}(f_{\text{ERM}}^m) - \text{err}_m(f_{\text{ERM}}^m) \\
+ \text{err}_m(f_{\text{ERM}}^m) - \text{err}_m(f^*) \\
+ \text{err}_m(f^*) - \text{err}(f^*) \\
\leq 2 \sup_{f \in \mathcal{F}} \left| \text{err}(f) - \text{err}_m(f) \right| \\
\leq 0
\]
Proof sketch

Fix any \( f \in \mathcal{F} \) and a sample \((x_i, y_i)\), define random variable

\[
Z_i^f := 1\{f(x_i) \neq y_i\}
\]

\[
\mathbb{E}[Z_1^f] \quad \quad \quad \quad \frac{1}{m} \sum_{i=1}^{m} [Z_i^f]
\]

(generalization error of \( f \)) (sample error of \( f \))

Lemma (Chernoff-Hoeffding bound ‘63):

Let \( Z_1, \ldots, Z_m \) be \( m \) Bernoulli r.v. drawn independently from \( \mathcal{B}(p) \).

for any tolerance level \( \epsilon > 0 \)

\[
\mathbb{P}_{Z_1} \left[ \left| \frac{1}{m} \sum_{i=1}^{m} [Z_i] - \mathbb{E}[Z_1] \right| > \epsilon \right] \leq 2e^{-2\epsilon^2 m}.
\]
Proof sketch

Need to analyze

\[ \mathbb{P}(x_i, y_i) \left[ \exists f \in \mathcal{F}, \left| \frac{1}{m} \sum_{i=1}^{m} [Z_i^f] - \mathbb{E}[Z_i^f] \right| > \epsilon \right] \]

\[ \leq \sum_{f \in \mathcal{F}} \mathbb{P}(x_i, y_i) \left[ \left| \frac{1}{m} \sum_{i=1}^{m} [Z_i^f] - \mathbb{E}[Z_i^f] \right| > \epsilon \right] \]

\[ \leq 2|\mathcal{F}|e^{-2\epsilon^2 m} \leq \delta \]

Equivalently, by choosing \( m \geq C \cdot \frac{1}{\epsilon^2} \ln \frac{|\mathcal{F}|}{\delta} \) with probability at least \( 1 - \delta \), for all \( f \in \mathcal{F} \)

\[ \left| \frac{1}{m} \sum_{i=1}^{m} [Z_i^f] - \mathbb{E}[Z_i^f] \right| = \left| \text{err}_m(f) - \text{err}(f) \right| \leq \epsilon \]
**Theorem (Occam’s Razor):**

Pick any tolerance level \( \epsilon > 0 \), and any confidence level \( \delta > 0 \) let \((x_1, y_1), \ldots, (x_m, y_m)\) be \( m \) examples drawn from an unknown \( \mathcal{D} \) if \( m \geq C \cdot \frac{1}{\epsilon^2} \ln \frac{|\mathcal{F}|}{\delta} \), then with probability at least \( 1 - \delta \)

\[
\text{err}(f^\text{ERM}_m) - \text{err}(f^*) \leq \epsilon
\]

\( \mathcal{F} \) is efficiently PAC learnable
Learning general concepts

Consider linear classification

$\mathcal{F} = \left\{ \right\} \quad |\mathcal{F}| = \infty$

Occam’s Razor bound is ineffective
Need to capture the true richness of \( \mathcal{F} \)

**Definition (Vapnik-Chervonenkis or VC dimension):**

We say that a model class \( \mathcal{F} \) as VC dimension \( d \), if \( d \) is the largest set of points \( x_1, \ldots, x_d \subset X \) such that for all possible labellings of \( x_1, \ldots, x_d \) there exists some \( f \in \mathcal{F} \) that achieves that labelling.

**Example:** \( \mathcal{F} = \) linear classifiers in \( \mathbb{R}^2 \)

<table>
<thead>
<tr>
<th>Linear classifiers can realize all possible labellings of 3 points</th>
<th>Linear classifiers CANNOT realize all labellings of 4 points</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Diagram 1]</td>
<td>![Diagram 2]</td>
</tr>
</tbody>
</table>

\( \text{VC}(\mathcal{F}) = 3 \)
**Theorem (Vapnik-Chervonenkis ’71):**

Pick any tolerance level $\epsilon > 0$, and any confidence level $\delta > 0$

let $(x_1, y_1), \ldots, (x_m, y_m)$ be $m$ examples drawn from an unknown $\mathcal{D}$

if $m \geq C \cdot \frac{\text{VC}(\mathcal{F}) \ln(1/\delta)}{\epsilon^2}$, then with probability at least $1 - \delta$

$$\text{err}(f_{\text{ERM}}^m) - \text{err}(f^*) \leq \epsilon$$

$\mathcal{F}$ is efficiently PAC learnable

VC Theorem $\Rightarrow$ Occam’s Razor Theorem
Theorem (VC lower bound):

Let $\mathcal{A}$ be any model selection algorithm that given $m$ samples, returns a model from $\mathcal{F}$, that is, $\mathcal{A} : (x_i, y_i)_{i=1}^m \mapsto f^A_m$

For all tolerance levels $0 < \epsilon < 1$, and all confidence levels $0 < \delta < 1/4$, there exists a distribution $\mathcal{D}$ such that if $m \leq C \cdot \frac{VC(\mathcal{F})}{\epsilon^2}$

$$\mathbb{P}_{(x_i, y_i)} \left[ |\text{err}(f^A_m) - \text{err}(f^*)| > \epsilon \right] > \delta$$
Some implications

• VC dimension of a model class **fully characterizes** its learning ability!

• Results are **agnostic** to the underlying distribution.
One algorithm to rule them all?

From our discussion it may seem that ERM algorithm is universally consistent.

This is not the case!

Theorem (no free lunch, Devroye ‘82):

Pick any sample size $m$, any algorithm $\mathcal{A}$ and any $\epsilon > 0$

There exists a distribution $\mathcal{D}$ such that

$$\text{err}(f_m^\mathcal{A}) > \frac{1}{2} - \epsilon$$

while the Bayes optimal error, $\inf_f \text{err}(f) = 0$
Further refinements and extensions

• How to do model class selection? Structural risk results.

• Dealing with kernels – Fat margin theory

• Incorporating priors over the models – PAC-Bayes theory

• Is it possible to get distribution dependent bound? Rademacher complexity

• How about regression? Can derive similar results for nonparametric regression.
Questions / Discussion
Thank You!

References:

