More Generalization Theorems

Yufei Tao

Department of Computer Science and Engineering
Chinese University of Hong Kong
Classification

Let \( A_1, \ldots, A_d \) be \( d \) attributes, where \( A_i \ (i \in [1, d]) \) has domain \( \text{dom}(A_i) = \mathbb{R} \).

Instance space \( \mathcal{X} = \text{dom}(A_1) \times \text{dom}(A_2) \times \ldots \times \text{dom}(A_d) = \mathbb{R}^d \).

Label space \( \mathcal{Y} = \{-1, 1\} \).

Each instance-label pair (a.k.a. object) is a pair \((x, y)\) in \( \mathcal{X} \times \mathcal{Y} \).

- \( x \) is a vector; we use \( x[A_i] \) to represent the vector's value on \( A_i \) \((1 \leq i \leq d)\).

Denote by \( \mathcal{D} \) a probabilistic distribution over \( \mathcal{X} \times \mathcal{Y} \).
**Goal:** Given an object \((x, y)\) drawn from \(\mathcal{D}\), we want to predict its label \(y\) from its attribute values \(x[A_1],...,x[A_d]\).

A **classifier** is a function 

\[ h : \mathcal{X} \to \mathcal{Y}. \]

Denote by \(\mathcal{H}\) a collection of classifiers.

The **error of \(h\) on \(\mathcal{D}\)** (i.e., generalization error) is defined as:

\[ \text{err}_{\mathcal{D}}(h) = \Pr_{(x,y) \sim \mathcal{D}}[h(x) \neq y]. \]

We want to learn a classifier \(h \in \mathcal{H}\) with small \(\text{err}_{\mathcal{D}}(h)\) from a **training set** \(S\) where each object is drawn independently from \(\mathcal{D}\).
We want to learn a classifier \( h \in \mathcal{H} \) with small \( \text{err}_D(h) \) from a training set \( S \) where each object is drawn independently from \( D \).

The error of \( h \) on \( S \) (i.e., empirical error) is defined as:

\[
\text{err}_S(h) = \frac{|(x, y) \in S \mid h(x) \neq y|}{|S|}.
\]
Let $P$ be a set of points in $\mathbb{R}^d$. Given a classifier $h \in \mathcal{H}$, we define:

$$P_h = \{ p \in P \mid h(p) = 1 \}$$

namely, the set of points in $P$ that $h$ classifies as 1.

$\mathcal{H}$ shatters $P$ if, for any subset $P' \subseteq P$, there exists a classifier $h \in \mathcal{H}$ satisfying $P' = P_h$.
**Example:** An extended linear classifier $h$ is described by a $d$-dimensional weight vector $w$ and a threshold $\tau$. Given an instance $x \in \mathbb{R}^d$, $h(x) = 1$ if $w \cdot x \geq \tau$, or $-1$ otherwise. Let $\mathcal{H}$ be the set of all extended linear classifiers.

In 2D space, $\mathcal{H}$ shatters the set $P$ of points shown below.
Example (cont.): Can you find 4 points in $\mathbb{R}^2$ that can be shattered by $\mathcal{H}$?

The answer is no. Can you prove this?
Let \( P \) be a subset of \( \mathcal{X} \). The **VC-dimension** of \( \mathcal{H} \) on \( P \) is the size of the largest subset \( P \subseteq \mathcal{P} \) that can be shattered by \( \mathcal{H} \).

If the VC-dimension is \( \lambda \), we write \( \text{VC-dim}(\mathcal{P}, \mathcal{H}) = \lambda \).
**Theorem:** Let $\mathcal{H}$ be the set of extended linear classifiers. $\text{VC-dim}(\mathbb{R}^d, \mathcal{H}) = d + 1$.

The proof is outside the syllabus.

**Example:** We have seen earlier that when $d = 2$, $\mathcal{H}$ can shatter at least one set of 3 points but cannot shatter any set of 4 points. Hence, $\text{VC-dim}(\mathbb{R}^2, \mathcal{H}) = 3$.

**Think:** Now consider $\mathcal{H}$ as the set of linear classifiers (where the threshold $\tau$ is fixed to 0). What can you say about $\text{VC-dim}(\mathbb{R}^d, \mathcal{H})$?
VC-Based Generalization Theorem

The support set of $\mathcal{D}$ is the set of points in $\mathbb{R}^d$ that have a positive probability to be drawn according to $\mathcal{D}$.

**Theorem:** Let $\mathcal{P}$ be the support set of $\mathcal{D}$ and set $\lambda = \text{VC-dim}(\mathcal{P}, \mathcal{H})$. Fix a value $\delta$ satisfying $0 < \delta \leq 1$. It holds with probability at least $1 - \delta$ that

$$
err_D(h) \leq err_S(h) + \sqrt{\frac{8 \ln \frac{4}{\delta} + 8 \lambda \cdot \ln \frac{2e|S|}{\lambda}}{|S|}}.
$$

for every $h \in \mathcal{H}$, where $S$ is the set of training points.

The proof is outside the syllabus.
The new generalization theorem places no constraints on the size of $\mathcal{H}$.

**Think:** What implications can you draw about the Perceptron algorithm?
If a set $\mathcal{H}$ of classifiers is “more powerful” — namely, having a greater VC dimension — it is more difficult to learn because a larger training set is needed.

For the set $\mathcal{H}$ of (extended) linear classifiers, the training set size needs to be $\Omega(d)$ to ensure a small generalization error. This becomes a problem when $d$ is large. In fact, in some situations we may even want to work with $d = \infty$.

Next, we will introduce another generalization theorem for the linear classification problem.
Recall:

**Linear classifier**: A function $h : \mathcal{X} \to \mathcal{Y}$ where $h$ is defined by a $d$-dimensional **weight vector** $w$ such that

- $h(x) = 1$ if $x \cdot w \geq 0$;
- $h(x) = -1$ otherwise.

$S$ is **linearly separable** if there is a $d$-dimensional vector $w$ such that for each $p \in S$:

- $w \cdot p > 0$ if $p$ has label 1;
- $w \cdot p < 0$ if $p$ has label $-1$.

The linear classifier that $w$ defines is said to **separate** $S$. 
Let $h$ be a linear classifier defined by a $d$-dimensional vector $w$. We say that $h$ is **canonical** if for every point $p \in S$:

- $w \cdot p \geq 1$ if $p$ has label 1
- $w \cdot p \leq -1$ if $p$ has label $-1$;

and the equality holds on **at least one point** in $S$.

**Think:** If $h$ separates $S$, it always has a canonical form. Why?
Theorem: Let $\mathcal{H}$ be the set of linear classifiers. Suppose that the training set $S$ is linearly separable. Fix a value $\delta$ satisfying $0 < \delta \leq 1$. It holds with probability at least $1 - \delta$ that,

$$\text{err}_D(h) \leq \frac{4R \cdot |w|}{\sqrt{|S|}} + \sqrt{\ln \frac{2}{\delta} + \ln \lceil \log_2 (R |w|) \rceil} \cdot \frac{1}{|S|},$$

for every canonical $h \in \mathcal{H}$, where $w$ is the $d$-dimensional vector defining $h$ and

$$R = \max_{p \in S} |p|.$$
Margin-Based Generalization Theorem

Why is the theorem “margin-based”? The margin of the separation plane defined by \( w \) equals \( \frac{1}{|w|} \) -(we will derive this later in the course).

When the training set \( S \) is linearly separable, we should find a separation plane with the largest margin.