# More Generalization Theorems

Yufei Tao

Department of Computer Science and Engineering
Chinese University of Hong Kong

## Classification

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

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

**Label space** $\mathcal{Y} = \{-1, 1\}$.

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

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

Denote by $\mathcal{D}$ a probabilistic distribution over $\mathcal{X} \times \mathcal{Y}$.

## Classification

**Goal:** Given an object $(\boldsymbol{x}, y)$ drawn from $\mathcal{D}$, we want to predict its label $y$ from its attribute values $\boldsymbol{x}[A_1], ..., \boldsymbol{x}[A_d]$.

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:

$$err_{\mathcal{D}}(h) = \boldsymbol{Pr}_{(\boldsymbol{x},y)\sim\mathcal{D}}[h(\boldsymbol{x}) \neq y].$$

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

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

$$err_S(h) = \frac{\left|(\boldsymbol{x}, y) \in S \mid h(\boldsymbol{x}) \neq y\right|}{|S|}.$$

## Shattering

Let $P$ be a set of points in $\mathbb{R}^d$. Given a classifier $h \in \mathcal{H}$, we define:

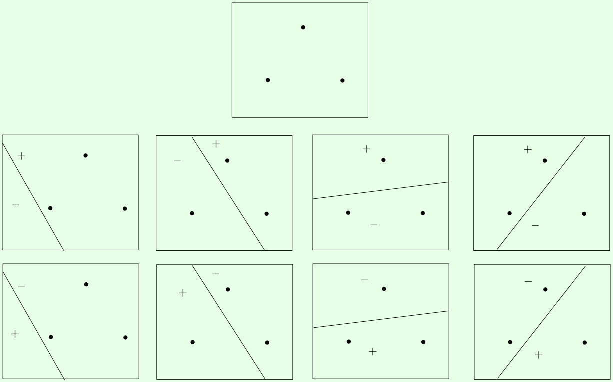
$$P_h = \{p \in P \mid h(p) = 1\}$$

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 $\boldsymbol{w}$ and a threshold $\tau$. Given an instance $\boldsymbol{x} \in \mathbb{R}^d$, $h(\boldsymbol{x}) = 1$ if $\boldsymbol{w} \cdot \boldsymbol{x} \geq \tau$, 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?

## VC Dimension

> Let $\mathcal{P}$ be a subset of $\mathcal{X}$. The **VC-dimension** of $\mathcal{H}$ on $\mathcal{P}$ 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 $\text{VC-dim}(\mathcal{P}, \mathcal{H}) = \lambda$.

## VC Dimension of Extended Linear Classifiers

**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 $\lambda = \mathrm{VC\text{-}dim}(\mathcal{P}, \mathcal{H})$. Fix a value $\delta$ satisfying $0 < \delta \leq 1$. It holds with probability at least $1 - \delta$ that

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

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** $\boldsymbol{w}$ such that

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

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

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

The linear classifier that $\boldsymbol{w}$ 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 $p \in S$:

- $\boldsymbol{w} \cdot \boldsymbol{p} \geq 1$ if $p$ has label 1
- $\boldsymbol{w} \cdot \boldsymbol{p} \leq -1$ 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?

## Margin-Based Generalization Theorem

**Theorem:** Let $\mathcal{H}$ be the set of linear classifiers. Suppose that the training set $S$ is **linearly separable**. Fix a value $\delta$ satisfying $0 < \delta \le 1$. It holds with probability at least $1 - \delta$ that,

$$err_D(h) \le \frac{4R \cdot |\boldsymbol{w}|}{\sqrt{|S|}} + \sqrt{\frac{\ln \frac{2}{\delta} + \ln \lceil \log_2(R|\boldsymbol{w}|) \rceil}{|S|}}.$$

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

$$R = \max_{\boldsymbol{p} \in S} |\boldsymbol{p}|.$$

The proof is outside the syllabus.

The theorem does not depend on the dimensionality $d$.

## Margin-Based Generalization Theorem

Why is the theorem "margin-based"?
The margin of the separation plane defined by $\boldsymbol{w}$ equals $1/|\boldsymbol{w}|$.

When the training set $S$ is linearly separable, we should find a separation plane with the **largest** margin.