

More Generalization Theorems

Yufei Tao

Department of Computer Science and Engineering
Chinese University of Hong Kong

Classification

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

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

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

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

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

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

Classification

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

A **classifier** is a function

$$h: \mathcal{X} \rightarrow \mathcal{Y}.$$

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

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

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

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 $err_{\mathcal{D}}(h)$ 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:

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

Shattering

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

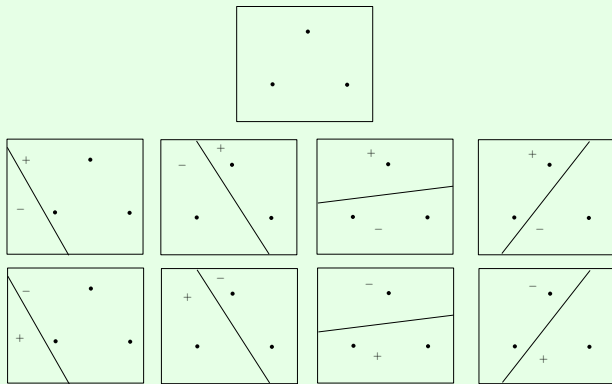
$$P_h = \{p \in P \mid h(p) = 1\}$$

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 \mathbf{w} and a threshold τ . Given an instance $\mathbf{x} \in \mathbb{R}^d$, $h(\mathbf{x}) = 1$ if $\mathbf{w} \cdot \mathbf{x} \geq \tau$, 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 λ , 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 τ 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 δ satisfying $0 < \delta \leq 1$. It holds with probability at least $1 - \delta$ that

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

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 $h : \mathcal{X} \rightarrow \mathcal{Y}$ where h is defined by a d -dimensional **weight vector** \mathbf{w} such that

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

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

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

The linear classifier that \mathbf{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?

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 δ satisfying $0 < \delta \leq 1$. It holds with probability at least $1 - \delta$ that,

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

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

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

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 \mathbf{w} equals $1/|\mathbf{w}|$ (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.