More Generalization Theorems

Yufei Tao

Department of Computer Science and Engineering
Chinese University of Hong Kong

Classification

For today's lecture, let us consider a slightly more general version of the classification problem by allowing a don't-know option for the classifiers.

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

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

Label space $\mathcal{Y} = \{-1, 1, *\}$, where * means **"don't know"**.

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

- we use $\boldsymbol{x}[A_i]$ to represent the value of $\boldsymbol{x}$ on A_i $(1 \leq i \leq d)$.

Classification

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

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 $\mathcal{D}$.

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:

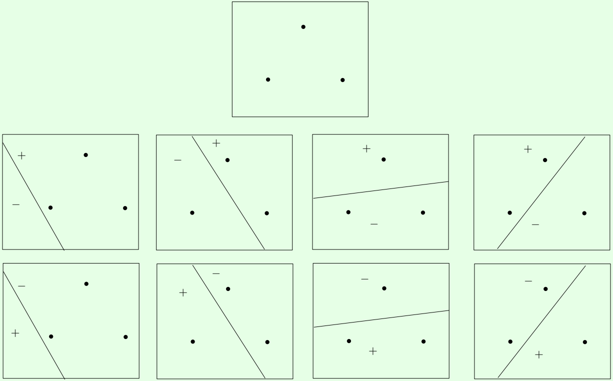
$$P_h = \{p \in P \mid h(p) = 1\}$$

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 **generic linear classifier** h is described by a d-dimensional weight vector $\boldsymbol{w}$ and a threshold τ. 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 generic 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 Generic Linear Classifiers

Theorem: Let $\mathcal{H}$ be the set of generic 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 $\lambda = \mathrm{VC\text{-}dim}(\mathcal{P}, \mathcal{H})$. Fix a value δ 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.

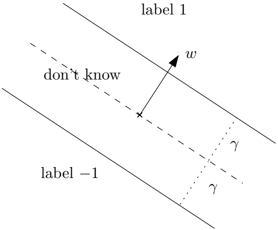
The new generalization theorem places **no constraints** on the size of $\mathcal{H}$.

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 (generic) 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, later in the course we may even want to work with $d = \infty$.

Next, we will introduce another generalization theorem to address the problem.



A **margin classifier** is a function $h : \mathcal{X} \to \mathcal{Y}$ where h is defined by a d-dimensional unit vector $\boldsymbol{w}$ (called the **weight vector**) and a non-negative real value γ (called the **margin**) such that

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

Theorem: Let $\mathcal{P}$ be a set of points whose distances to the origin are bounded by R. Let $\mathcal{H}_\gamma$ be the set of margin classifiers with margin at least γ. Then, $\mathrm{VC\text{-}dim}(\mathcal{P}, \mathcal{H}_\gamma) \le (R/\gamma)^2$.

The proof is outside the syllabus.

For the linear classification problem, the theorem provides strong justification on choosing a linear classifier whose separation plane is as far away from the sample points as possible.

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.

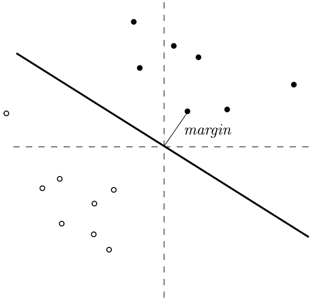
A set $S \subseteq \mathbb{R}^d$ is **linearly separable** if there is a d-dimensional vector $\boldsymbol{w}$ such that for each $\boldsymbol{p} \in S$:

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

The linear classifier defined by $\boldsymbol{w}$ is said to **separate** S.

Let h be a linear classifier defined by a d-dimensional vector $\boldsymbol{w}$. Its **separation plane**, denoted as π, is the plane defined by equation $\boldsymbol{x} \cdot \boldsymbol{w} = 0$.

Suppose that h separates a linearly separable set S. Then, the **margin** of h on S is the smallest distance of the points in S to π.



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

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

for **every** $h \in \mathcal{H}$ on S, where γ is the margin of h on S and

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

The proof is outside the syllabus.

The theorem does not depend on the dimensionality d.