Lecture 10: VC-dimension (Chapter 5 of Textbook B)

Jinwoo Shin

AI503: Mathematics for AI

Problem Setup and Summary

Consider a set S of labeled *training examples* independently drawn from a probability distribution D over the instance space $\mathcal{X} = \mathbb{R}^d$.

We aim generalization: use the training examples to produce a classification rule that will perform well over new data, i.e., new points that are also drawn from D.

Namely, for a target function $c^* : \mathcal{X} \to \mathcal{Y}$ (where $\mathcal{Y}$ is output space), we find a hypothesis $h : \mathcal{X} \to \mathcal{Y}$ that approximates c^* from some class $\mathcal{H}$ by using S.

VC-dimension is a measurement of complexity for a hypothesis class $\mathcal{H}$.

- One can use it to measure generalization guarantees of a given hypothesis class.

Roadmap

KAIST AI
Graduate School of AI

(1) Generalization

(2) Overfitting and Uniform Convergence

(3) VC-Dimension

(4) VC-Dimension Sample Bound

(5) Other Measures of Complexity

Roadmap

Generalization: Formalizing the problem

Through out the lecture, we consider a binary classification problem of $x \sim D$ where our hypothesis h are $\{-1, 1\}$-valued indicator function:

$$h(x) = \begin{cases} 1, & x \in h \\ -1, & x \notin h \end{cases}$$

Let c^*, called the target concept, we denote each error of h as follows:

- training error: $err_S(h) = \text{Prob}_{x \sim S}[h(x) \neq c^*(x)]$
- true error (i.e., test error): $err_D(h) = \text{Prob}_{x \sim D}[h(x) \neq c^*(x)]$

Generalization: finding a hypothesis h that has a low true error, with the training set.

Generalization: Overfitting

We call the hypothesis h is *overfitting* on the training data when h has a low training error and yet have a high true error, i.e., crucial for generalization.

To analyze the overfitting, we introduce the notion of a *hypothesis class*.

- An hypothesis class $\mathcal{H}$ is a set of candidate formulas of h.

We argue that if the training set S is large enough compared to some property of $\mathcal{H}$, the overfitting is addressed: will introduce *two generalization guarantees*.

Roadmap

KAIST AI Graduate School of AI

Overfitting and Uniform Convergence: PAC learning Guarantee

We assume hypothesis class $\mathcal{H}$ is finite (later we will extend to infinite case).

Theorem 1. Probably approximately correct (PAC) learning Guarantee

Let $\mathcal{H}$ be an hypothesis class and let ϵ and δ be greater than zero. If a training set S of size

$$n \geq \frac{1}{\epsilon}\big(\ln|\mathcal{H}| + \ln(1/\delta)\big),$$

is drawn from distribution D, then with probability greater than or equal to $1-\delta$, every $h \in \mathcal{H}$ with true error $err_D(h) \geq \epsilon$ has training error $err_S(h) > 0$. (equivalently, every $h \in \mathcal{H}$ with $err_S(h) = 0$ has $err_{\mathcal{D}}(h) < \epsilon$).

Overfitting and Uniform Convergence: PAC-learning Guarantee

KAIST AI
Graduate School of AI

Proof. Let $h_1, h_2, \ldots$ be the hypotheses in $\mathcal{H}$ with true error $err_D(h_i) \geq \epsilon$.

Consider drawing the sample S of size n and let A_i be the event that h_i consistent with S, i.e., h_i makes no mistakes on S. Then the probability of event A_i is as:

$$\text{Prob}(A_i) \leq (1-\epsilon)^n.$$

By using two facts (i) $\text{Prob}(\cup_i A_i) \leq \sum_i \text{Prob}(A_i)$, and (ii) $1-\epsilon \leq e^{-\epsilon}$, we obtain the following form:

$$\text{Prob}(\cup_i A_i) \leq |\mathcal{H}| e^{-\epsilon n},$$

One can prove the theorem, by considering δ that satisfies $|\mathcal{H}| e^{-\epsilon n} \leq \delta$.

Overfitting and Uniform Convergence: Uniform Convergence

We assume hypothesis class $\mathcal{H}$ is finite (later we will extend to infinite case).

Theorem 2. Uniform Convergence

Let $\mathcal{H}$ be an hypothesis class and let ϵ and δ be greater than zero. If a training set S of size

$$n \geq \frac{1}{2\epsilon^2}\big(\ln|\mathcal{H}| + \ln(2/\delta)\big),$$

is drawn from distribution D, then with probability greater than or equal to $1-\delta$, every $h \in \mathcal{H}$ satisfies $|err_S(h) - err_D(h)| \leq \epsilon$. (equivalently, every $h \in \mathcal{H}$ with $err_D(h) = 0$ has $err_S(h) < \epsilon$).

Proof. By utilizing Hoeffding bounds guarantee (Theorem 4.3 in the textbook), one can prove the uniform convergence bound (in textbook page 138).

Overfitting and Uniform Convergence: Finite Hypothesis Class

Note that two theorems require $\mathcal{H}$ to be finite in order to be meaningful since they require a sample size of

- $n \geq \frac{1}{\epsilon}\left(\ln|\mathcal{H}| + \ln(1/\delta)\right)$ for theorem 1.
- $n \geq \frac{1}{2\epsilon^2}\left(\ln|\mathcal{H}| + \ln(2/\delta)\right)$ for theorem 2.

In the next section, we will introduce *VC-dimension* (and notion of growth functions) to extend theorems to certain infinite hypothesis classes.

Roadmap

KAIST AI
Graduate School of AI

(1) Generalization

(2) Overfitting and Uniform Convergence

(3) VC-Dimension

(4) VC-Dimension Sample Bound

(5) Other Measures of Complexity

VC-dimension: Definition

Definition 1. Shatter

Given a set S of examples and a hypothesis class $\mathcal{H}$, we say that S is **shattered** by $\mathcal{H}$ if for every $S^+ \subseteq S$ there exists some $h \in \mathcal{H}$ that labels all examples in S^+ as positive (i.e., $+1$) and all examples in $S \backslash S^+$ as negatives (i.e., -1).

In a high level, we say a classifier h can shatter a set $S := S^+ \cup S^-$ if h can achieve zero training error (i.e., classify exactly) on S for all possible partitions.

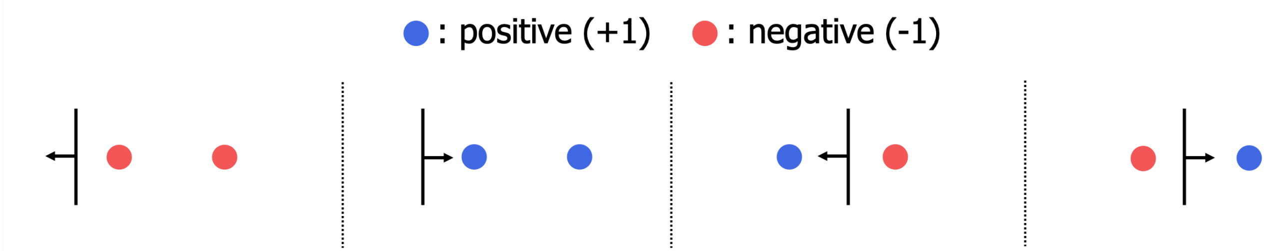
Definition 2. VC-dimension

The **VC-dimension** of $\mathcal{H}$ is the size of the largest set shattered by $\mathcal{H}$.

VC-dimension: Shatter Example (1)

Example. 1-D Case with a linear classifier (i.e., perceptron).

- Can we shatter a set of $|S| = 2$? where $|.|$ denotes the cardinality. Yes

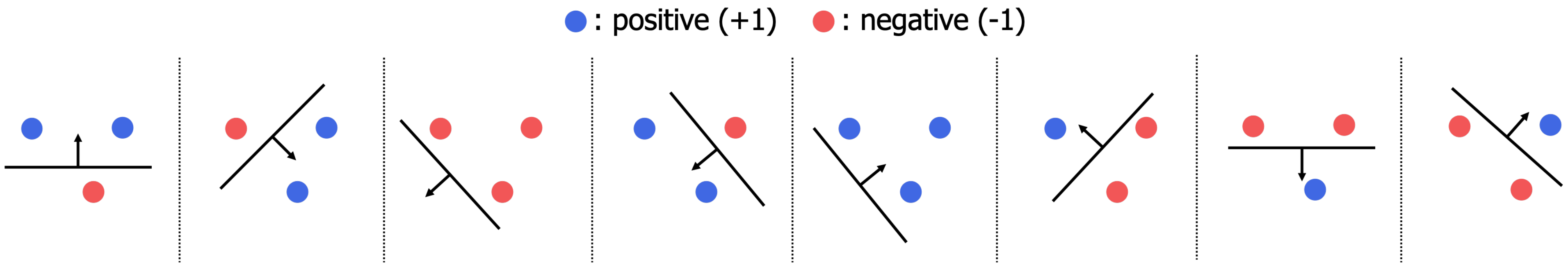


- Can we shatter a set of $|S| = 3$? No

VC-dimension: Shatter Example (2)

Example. 2-D Case with a linear classifier (i.e., perceptron).

- What is the largest set S shattered by $h \in \mathcal{H}$? 3!



More examples (will not handle in class):

- Prove that the largest set S shattered by a linear classifier in d-D is $d+1$.
- Prove that the largest set S shattered by the k-nearest neighbor with $k = 1$ is ∞.

VC-Dimension: Growth Function

KAIST AI
Graduate School of AI

Definition 3. Growth function

Given a set S of examples and a hypothesis class $\mathcal{H}$, let $\mathcal{H}[S] = \{h \cap S : h \in \mathcal{H}\}$. That is, $\mathcal{H}[S]$ is the hypothesis class $\mathcal{H}$ restricted to the set of points S. For integer n and class $\mathcal{H}$, let $\mathcal{H}[n] = \max_{|S|=n} |\mathcal{H}[S]|$; this is called the **growth function** of $\mathcal{H}$.

Connection with VC-dimension: S is shattered by $\mathcal{H}$ if $|\mathcal{H}[S]| = 2^{|S|}$, and then the VC-dimension of $\mathcal{H}$ is the largest n such that $\mathcal{H}[n] = 2^n$.

In a high level, growth function can be thought as a measure of the "size" of $\mathcal{H}$: we will utilize it for the generalization guarantee bound.

Roadmap

KAIST AI
Graduate School of AI

VC-Dimension Sample Bound: Growth Function

KAIST AI
Graduate School of AI

Theorem 3. Growth function sample bound

For any class $\mathcal{H}$ and distribution $\mathcal{D}$, if a training sample S is drawn from $\mathcal{D}$ of size,

$$n \geq \frac{2}{\epsilon}\left[\log_2(2\mathcal{H}[2n]) + \log_2(1/\delta)\right],$$

that with probability $\geq 1-\delta$, every $h \in \mathcal{H}$ with $err_{\mathcal{D}}(h) \geq \epsilon$ has $err_S(h) > 0$ (equivalently, every $h \in \mathcal{H}$ with $err_S(h) = 0$ has $err_{\mathcal{D}}(h) < \epsilon$).

Theorem 4. Growth function uniform convergence

For any class $\mathcal{H}$ and distribution $\mathcal{D}$, if a training sample S is drawn from $\mathcal{D}$ of size,

$$n \geq \frac{8}{\epsilon}\left[\log_2(2\mathcal{H}[2n]) + \log_2(1/\delta)\right],$$

that with probability $\geq 1-\delta$, every $h \in \mathcal{H}$ will have $|err_s(h) - err_D(h)| \leq \epsilon$.

One can extend Theorem 1, and 2 (i.e., generalization bound with finite $\mathcal{H}$) with the growth function $\mathcal{H}[n]$ to obtain the above theorem (see textbook page 154, and 155).

VC-Dimension Sample Bound: Sauer's Lemma

Theorem 5. Sauer's Lemma

If $\mathrm{VCdim}(\mathcal{H}) = d$ then for all $n \in \mathbb{N}$,

$$\mathcal{H}[n] \leq \sum_{i=0}^{d} \binom{n}{i}.$$

Futhermore, for all $n \geq d$, we have

$$\mathcal{H}[n] \leq (\frac{en}{d})^d,$$

where e is Euler's number.

This indicates that if $\mathrm{VCdim}(\mathcal{H})$ is ∞, we always get exponential growth function

However, if $\mathrm{VCdim}(\mathcal{H}) = d$ is finite, growth function increases exponentially up to d and polynomially for $n > d$.

The proof of the theorem is given in the textbook page 155-156.

VC-Dimension Sample Bound

KAIST AI
Graduate School of AI

Corollary 1. VC-dimension sample bound

For any class $\mathcal{H}$ and distribution D , a training sample S of size

$$O\Big(\frac{1}{\epsilon}\big[\mathrm{VCdim}(\mathcal{H})\log(1/\epsilon)+\log(1/\delta)\big]\Big)$$

is sufficient to ensure that with probability $\geq 1-\delta$, every $h \in \mathcal{H}$ with $err_{\mathcal{D}}(h) \geq \epsilon$ has $err_S(h) > 0$ (equivalently, every $h \in \mathcal{H}$ with $err_S(h) = 0$ has $err_{\mathcal{D}}(h) < \epsilon$).

By putting Theorem 3 and 5 together, with a little algebra we get the above corollary (one can obtain similar corollary by combining Theorem 4 and 5).

Note that, Corollary 1 can be much better than Theorem 1, i.e., generalization guarantee with finite hypothesis class $\ln(|\mathcal{H}|)$.

- For any class $\mathcal{H}$, $\mathrm{VCdim}(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$ since $\mathcal{H}$ must have at least 2^k concepts in order to shatter k points.

Roadmap

Other Measures of Complexity: Rademacher Complexity

For your interest; there also exists other measures of complexity for $\mathcal{H}$.

One popular measurement is Rademacher complexity which is as follows:

$$\mathcal{R}_S(\mathcal{H}) := \mathbb{E}_{\sigma_1,\ldots,\sigma_n} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(x_i),$$

where $\sigma_i \in \{-1, 1\}$ is uniformly distributed random variable.

Example. If you assign random labels to the points in S and the best classifier in $\mathcal{H}$ on average gets error 0.45 then $\mathcal{R}_S(\mathcal{H}) = 0.55$ - $0.45 = 0.1$.

One can obtain the true error bound with the Rademacher complexity:

$$err_D(h) \leq err_S(h) + \mathcal{R}_S(\mathcal{H}) + 3\sqrt{\frac{\log(2/\delta)}{2n}},$$

with probability $\geq 1 - \delta$.

For the proof of the Rademacher complexity bound, see the following reference:

- Bartlett and Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results, JMLR 2002.

Questions?