Machine Learning for Intelligent Systems

Lecture 19: Statistical Learning Theory 3

Reading: UML 6

Instructors: Nika Haghtalab (this time) and Thorsten Joachims

Growth Function & VC Dimension

Growth function

The set all m-tuples produced by hypotheses in H on the sample set S

 $H[S] = \left\{ \left(h(x_1), h(x_2), h(x_3), \dots, h(x_m) \right) \right\}_{h \in H}$

Growth function: $H[m] = \max_{|S|=m} |H[S]|$ is the largest number of

unique rows that H can produce on any set of m elements.

Recall: Shattering and VC Dimension -

H shatters a sample set *S* if $|H[S]| = 2^{|S|}$.

VC Dimension of *H* is the size of the largest set *S* that can be

shattered by *H*. \leftarrow VCDim(*H*): Largest *m* for which $H[m] = 2^m$.

To show that VCDim(H) = d we need to show

1. There **exists a set** of **d** points that can be shattered.

2. There is **no set** of d + 1 points that can be shattered.

VC Dimension of Linear Threshold

Theorem: VC Dimension of Linear thresholds in \mathbb{R}^d

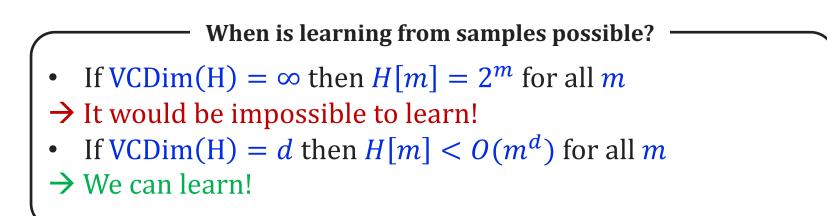
Let *H* be the set of all homogenous linear thresholds in \mathbb{R}^d . We have VCDim(*H*) = *d*.

Let *H* be the set of all linear thresholds (possibly non-homogenous) in \mathbb{R}^d . We have VCDim(H) = d + 1.

- → You can shatter the set $\{\vec{0}, \vec{e_1}, ..., \vec{e_d}\}$, where $\vec{e_i} = (0, ..., 0, 1, 0, ..., 0)$ has 1 only at coordinate i. → VCDim $(H) \ge d + 1$. (Try at home)
- → Showing that we cannot shatter a set of d + 2 points requires more work (we won't cover it).

VC Dimension & Learnability

VC Dimension is roughly the point where the growth function stops being exponential and becomes polynomial.



PAC Learnability

Probably Approximately Correct Learnability

A hypothesis class *H* is **PAC learnable** if there is a function $m_H(\epsilon, \delta)$ and a learning algorithm such that:

For any $\epsilon, \delta \in (0,1)$ and any distribution *P* over $X \times Y$ such that all samples are labeled by one hypothesis $h^* \in H$, running the learning algorithm on $m \ge m_H(\epsilon, \delta)$ i.i.d. samples generated from *P*, the algorithm returns $h \in H$ such that with probability $1 - \delta$ over the choice of the samples, $err_P(h) \le \epsilon$.

Often called "realizable" PAC: There is a hypothesis $err_P(h^*) = 0$

Agnostic PAC Learnability

Probably Approximately Correct Learnability

A hypothesis class *H* is **PAC learnable** if there is a function $m_H(\epsilon, \delta)$ and a learning algorithm such that:

For any $\epsilon, \delta \in (0,1)$ and any distribution *P* over $X \times Y$

running the

learning algorithm on $m \ge m_H(\epsilon, \delta)$ i.i.d. samples generated from

P, the algorithm returns $h \in H$ such that with probability $1 - \delta$

over the choice of the samples $err_P(h) \leq \min_{h \in H} err_P(h) + \epsilon$

Often called "agnostic" PAC: No assumption on $\min_{h \in H} err_P(h)$

Theorem: Sample Complexity Infinite Hypothesis Class (zero empirical error)

Let $m \ge \frac{c_0}{\epsilon} \left(VCDim(H) \ln(\frac{1}{\epsilon}) + \ln(\frac{1}{\delta}) \right)$. For any $X, Y = \{-1, 1\}$, and distribution P on $X \times Y$, with probability $1 - \delta$ over i.i.d draws of set S of m samples, any $h \in H$ such that $\operatorname{err}_S(h) = 0$, also has $\operatorname{err}_P(h) < \epsilon$.

Probably Approximately Correct (PAC) (Belief that $err_P(h^*) = 0$) Agnostic Probably Approximately Correct (No belief about value of $err_P(h^*)$)

Theorem: Sample Complexity Infinite Hypothesis Class (Non-zero empirical error)

Let $m \ge \frac{c_0}{\epsilon^2} \left(VCDim(H) + \ln\left(\frac{1}{\delta}\right) \right)$. For any $X, Y = \{-1, 1\}$, and distribution P on $X \times Y$, with probability $1 - \delta$ over i.i.d draws of set Sof m samples, $h_S = argmin_{h \in H} err_S(H)$ has $err_P(h_S) \le err_P(h^*) + \epsilon$.

Algorithm: Empirical Risk Minimization (ERM)

<u>Empirical Risk Minimization alg</u>: Return $h_S = argmin_{h \in H}err_S(H)$

VC Dimension & Learnability

When is learning from samples possible?

All the following are equivalent:

- *H* has finite VC dimension.
- *H* is (realizable) PAC learnable
- *H* is agnostically PAC learnable
- The Empirical Risk Minimization algorithm PAC learns *H*.
- The Empirical Risk Minimization algorithm agnostically PAC learns *H*.