CS485/685
Lecture 18: March 8, 2012

VC Dimension

[BDSS] Chapter 5

PAC Learnability Recap

- A hypothesis class is PAC learnable when there exists an algorithm and a sample size $N = m\left(\frac{1}{\epsilon}, \frac{1}{\delta}\right)$ that achieves a loss within ϵ of optimal with probability greater than $1 - \delta$.
- Key: sample size N is finite for any ϵ and δ
- All finite hypothesis classes are PAC learnable
  - What about infinite hypothesis classes?

A non-learnable infinite class

- Let X be an infinite input space and $Y = \{0,1\}$
- Let H be the class of all binary partitions of X
  - This class is infinite since X is infinite
- **Theorem:** The class of all binary partitions H of an infinite input space X is not PAC learnable
  - i.e., there does not exist any algorithm that can achieve a loss within ϵ of optimal with probability greater than $1 - \delta$ based on a finite training set.

Non-learnability proof

- Two steps
  1. Use no-free-lunch theorem to characterize the expected loss of any algorithm
  2. Use Markov's inequality to show that it is impossible to achieve certain ϵ, δ pairs with a finite training set.

No-free-lunch theorem recap

- **No-free-lunch theorem:** for the class of all binary partitions, the expected loss of any algorithm is at least ¼ for some D when the training set is ½ the input space
  - i.e. $\mathrm{E}_{S\sim D^N}\left[L_D\big(A(S)\big)\right] \geq \frac{1}{4} \quad \forall A$
- Consequences:
  - Cannot guarantee to do better than random on new data
  - No generalization unless we assume a learning bias

Generalized no-free-lunch

- Let $N \leq \frac{|X|}{k}$ (i.e., sample size is at most one k^{th} the input size). For all algorithms A, there exists D, f such that $L_D(f) = 0$, but

$$E_{S\sim D^N}\left[L_D\big(A(S)\big)\right] \geq \frac{1}{2} - \frac{1}{2k}$$

- Intuition: picture

Infinite Hypothesis Class

- Let sample size N be finite and the input space X be infinite (hence H is also infinite). For all algorithms A, there exists D, f such that $L_D(f) = 0$, but
$$E_{S\sim D^N}\left[L_D\big(A(S)\big)\right] \geq \frac{1}{2}$$
- Derivation: as $k \rightarrow \infty$, then $\frac{1}{2} - \frac{1}{2k} \rightarrow \frac{1}{2}$

Markov Inequality

- Recall that $\forall a \ \Pr[Z > a] \leq \frac{E[Z]}{a}$
- Hence $\Pr\left[L_D\big(A(S)\big) \leq \epsilon\right]$

$= \Pr\left[1 - L_D\big(A(S)\big) > 1 - \epsilon\right]$

$\leq \frac{1 - E_{S\sim D^N}\left[L_D\left(A(S)\right)\right]}{1-\epsilon}$ (by Markov inequality)

$\leq \frac{1-1/2}{1-\epsilon}$ (by no-free-lunch theorem)

- If we pick $\epsilon = 0.1$, then $\Pr\left[L_D\big(A(S)\big) \leq 0.1\right] \leq 5/9$
- Hence, for any $1 - \delta > 5/9$, the training set must be infinite

A Learnable Infinite Class

- Let $X = \Re$, $Y = \{0,1\}$
- Consider threshold functions

$$h_a(x) = \begin{cases} 1 & x < a \\ 0 & \text{otherwise} \end{cases}$$

- **Lemma:** The class of threshold functions is PAC learnable
  - i.e. a finite training set is sufficient to achieve a loss of at most ϵ with probability greater than $1 - \delta$ even though the class is uncountably infinite

Intuition

- Picture

Proof

- Algorithm A: pick $h_a = \max\{x_n | y_n = 1\}$
- Sample size: $N \geq \log(\frac{1}{\delta})/\epsilon$
- Probability of a bad sample

$$\Pr_{S\sim D^N}[L_D(A(S)) > \epsilon]$$
$$= \Pr_{S\sim D^N}[\forall(x,y) \in S, x \notin (a_0, a^*)]$$
$$= (1-\epsilon)^N$$
$$\leq e^{-\epsilon N}$$
$$\leq \delta \quad \text{since } N \geq \log(\tfrac{1}{\delta})/\epsilon$$

PAC Learnability for Infinite Classes

- Idea: Uncountably infinite hypothesis classes can still be PAC learnable when the hypothesis space restricted to a finite number of data points does not grow exponentially.
- Threshold functions   Binary partitions

Restricting Hypothesis Classes

- Let $H: X \rightarrow \{0,1\}$
- Let $C = \{c_1, c_2, \dots, c_N\} \subset X$
  - i.e., C is a finite subset of the input space

- **Definition:** The **restriction** of H to C is the set of different functions from C to $\{0,1\}$ derived from H

$$H_C = \{(h(c_1), h(c_2), \dots, h(c_N)) | h \in H\}$$

Examples

- Threshold functions

- Binary partitions

Learnability of Restricted Classes

- **Lemma:** Let C be a subset of the input space X of size $|C| = 2N$. Let H_C be a hypothesis space restricted to C of size $|H_C| = 2^{2N}$. For a training set of size N and any algorithm A, there exists D, f such that $L_D(f) = 0$, but
$$E_{S\sim D^N}\left[L_D\big(A(S)\big)\right] \geq 1/4$$

- This follows from the no-free-lunch theorem
- As $|C| \to \infty$ then $N \to \infty$ and H is not PAC learnable

Shattering

- To determine the learnability of infinite classes, we would like to know whether $|H_C|$ grows exponentially with $|C|$

- **Definition (Shattering)**: A hypothesis class H shatters a finite set $C \subset X$ if $|H_C| = 2^{|C|}$

Vapnik-Chervonenkis (VC) Dimension

- **Definition (VC dimension):** The VC dimension of H, denoted $VCdim(H)$ is the size of the largest $C \subset X$ that can be shattered by H
- If H can shatter arbitrarily large C's then $VCdim(H) = \infty$

Non-learnability of Infinite Classes

- **Theorem:** If $VCdim(H) = \infty$ then H is not PAC learnable
  - i.e., it is impossible to achieve certain ϵ, δ pairs with a finite training set.
  - This follows from the definitions of VC dimension, shattering and the non-learnability of the class of all binary partitions.

Fundamental Theorem of PAC Learning

- **Theorem**: Let H be a hypothesis class of functions from a domain X to $\{0,1\}$. The following statements are equivalent:
  1. H has the uniform convergence property
  2. H is agnostic PAC learnable
  3. H is PAC learnable
  4. H has finite VC dimension

Sample Complexity

- **Theorem:** Let H be a hypothesis class with $VCdim(H) = d < \infty$. There exist constants C_1 and C_2 such that
  1. H is PAC learnable with sample complexity

$$C_1 \frac{d + \ln\left(\frac{1}{\delta}\right)}{\epsilon} \leq N \leq C_2 \frac{d \ln(\frac{1}{\epsilon}) + \ln\left(\frac{1}{\delta}\right)}{\epsilon}$$

  2. H has the uniform convergence property and is agnostic PAC learnable with sample complexity

$$C_1 \frac{d + \ln\left(\frac{1}{\delta}\right)}{\epsilon^2} \leq N \leq C_2 \frac{d + \ln\left(\frac{1}{\delta}\right)}{\epsilon^2}$$

VC Dimension Calculation

- Since the key to learnability is the VC dimension, we need to estimate the VC dimension of hypothesis classes
- To show that $VCdim(H) = d$, we need to show that
  1. There exists a set C of size d that is shattered by H
  2. Every set C of size $d + 1$ is not shattered by H

Example 1: threshold functions

Example 2: Intervals

Example 3: Axis aligned rectangles