THEORY OF MACHINE LEARNING

LECTURE 7

VC DIMENSION, FUNDAMENTAL THEOREM (CONTD.)

LAST WEEK

- Representative sample: for a hyp class H and distribution D over X, S is called "representative" if for all $h \in H$, | (avg error on S)(h) - risk_D (h) | $\leq \epsilon$
- **Observation.** If training data happened to be a representative sample, ERM is an agnostic learning algorithm
- **Observation 2.** For a finite hypothesis class, a *random* sample of size ~ log |H| is representative
- Chernoff + Union bound

WHAT ABOUT INFINITE CLASSES?

- Note: if sample is representative, we are good!

- “Growth function” of a class: total number of distinct ways in with H can label a set of m points (# distinct “sign patterns”)
- H is the class of l.t.f. in 1D: (m+1)
- H is the class of intervals in 1D: O(m^2)
- H is the class of axis-parallel rectangles? messy, but O(m^4)
- H is the class of “convex polygons” in 2D: 2^m (exponential)

TODAY

- “Small” growth function => hypothesis class is learnable!
- How to bound growth function? VC dimension & Sauer’s lemma (attributed to Sauer and Shelah, VC)

LEARNABILITY IN TERMS OF THE GROWTH FUNCTION

- Theorem: Suppose $\tau_H(m)$ be the growth function of a hypothesis class H. Then for any X, D, if we take a sample S of size m, with prob. 1-δ,

$$\sup_{h \in H} |err(h,S) - err(h,D)| \leq \frac{4 + \sqrt{\log \tau_H(2m)}}{\delta \sqrt{2m}}$$

- In other words, if 'm' is chosen so that RHS < ϵ, theorem implies that random sample of size m is ϵ-representative

EXAMPLES

- H is the class of l.t.f. in 1D: (m+1)
- H is the class of intervals in 1D: O(m^2)
- H is the class of axis-parallel rectangles? messy, but O(m^4)
- H is the class of "convex polygons" in 2D: 2^m (exponential)

HOW TO BOUND GROWTH FUNCTION?

- Shattering & VC dimension.

SAUER-SHELAH LEMMA (VAPNIK-CHERVONENKIS)

- **Lemma.** Let H be a hypothesis class of finite VC dimension d. Then for every m, we have:

$$\tau_H(m) \leq \binom{m}{0} + \binom{m}{1} + \cdots + \binom{m}{d}$$

- Much better than exponential, for m large
- Proof by a clever inductive argument

FUNDAMENTAL THEOREM OF (STAT) LEARNING THEORY

- <u>Theorem:</u> The following statements are *equivalent:*
  - Class H is PAC learnable
  - Class H is *agnostically* PAC learnable
  - Class H has finite VC dimension

- Implies that if H has infinite VC dimension, it is <u>**not**</u> PAC learnable! (same proof as no-free-lunch theorem)

LEARNABILITY IN TERMS OF THE GROWTH FUNCTION

- Theorem: Suppose $\tau_H(m)$ be the growth function of a hypothesis class H. Then for any X, D, if we take a sample S of size m, with prob. 1-δ,

$$\sup_{h \in H} |err(h,S) - err(h,D)| \leq \frac{4 + \sqrt{\log \tau_H(2m)}}{\delta \sqrt{2m}}$$

- In other words, if 'm' is chosen so that RHS < ϵ, theorem implies that random sample of size m is ϵ-representative

OUTLINE -- THE TWO SAMPLE TRICK

- Want to show that a random sample is ϵ-representative
- Take sample S, define event:

  A = Pr [sample is not representative]
- Way to "test" if S is not representative?
  - "Cross validation"
- Define new event S, S'
- "Swapping"