THEORY OF MACHINE LEARNING

LECTURE 6

VC DIMENSION, FUNDAMENTAL THEOREM

LAST LECTURE

- 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 gives a hypothesis with good generalization.
  (Question of ERM being efficient is orthogonal…)
- **Observation 2.** For a finite hypothesis class, a *random* sample of size ~ log |H| is representative
- Proof using "concentration" inequality (Chernoff/Hoeffding)

CONCENTRATION BOUND

- **Chernoff bound (Hoeffding).** Suppose $X_1, X_2, \ldots X_n$ are n iid samples from a distribution with mean μ and <u>support</u> [a, b]. Then we have

$$\Pr\left[\left|\frac{1}{n}(X_1 + \cdots + X_n) - \mu\right| > \epsilon\right] \leq 2\exp\left(-\frac{\epsilon^2 n}{(a-b)^2}\right)$$

- <u>Note</u>: exponential dependence on n

FINITE CLASSES ARE LEARNABLE

- Claim: for any X and distribution D over it, a sample of size $O\left(\frac{1}{\epsilon^2} \log \frac{|H|}{\delta}\right)$ is representative with prob. at least $1 - \delta$
- Proof: write 'm' for the sample size
  - First look at a single $h \in H$
  - Prob. that |sample error(h) – risk(h)| > ϵ can be viewed as an application of Chernoff bound!
  - Gives a bound $2e^{-\epsilon^2 m} < \frac{\delta}{|H|}$
  - Union bound to prove that Pr[diff > ϵ for *some* h] < δ

WHAT ABOUT INFINITE CLASSES?

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

  (modulo inefficiency of ERM)

- What if we can divide hypotheses into finitely many "classes"?

- Example of threshold functions on a line

GROWTH FUNCTION OF A CLASS

- Maximum number of “possible classifications” of an input of size m

LEARNABILITY IN TERMS OF THE GROWTH FUNCTION

- Theorem: Suppose $\tau_H(m)$ is an upper bound on the total number of distinct "classifications" (or "sign patterns") possible for any sample of size m. Then for any X, D, if we take a sample S of size m, we have, with prob. 1-δ,

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

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

- Theorem: 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 **not** PAC learnable! (same proof as no-free-lunch theorem)

LEARNABILITY IN TERMS OF THE GROWTH FUNCTION

- Theorem: Suppose $\tau_H(m)$ is an upper bound on the total number of distinct "classifications" (or "sign patterns") possible for any sample of size m. Then for any X, D, if we take a sample S of size m, we have, with prob. 1-δ,

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

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"