THEORY OF MACHINE LEARNING

LECTURE 5

PAC MODEL, VC DIMENSION

LAST WEEK

- Motivation: do we really need to restrict the hypothesis/concept class *before starting* learning? – yes! (No free lunch theorem)
- (PAC Learning): Learnability of a concept class H over domain X
  - Informally, for any $\mathrm{f} \in H$ and any distribution D over X, given examples of the form (x, f(x)), we can learn a hypothesis 'h' such that Risk_D (h) is < ϵ, with high prob. ($(1-\delta)$, for some parameter δ)
  - Sample size only function of H, ϵ, δ (not distribution)
  - Learned hypothesis need not belong to H (improper learning)
- (Agnostic): f need not belong to H

GENERIC ALGORITHM

- <u>Empirical risk minimization:</u> given the samples (x, f(x)), find hypothesis $h \in H$ that minimizes the total error on the samples
  - most natural algorithm == minimize training error
- How to do it efficiently?
  - Don't care for now… maybe brute force over hypothesis class?
- When does it work?
  - If sample is "representative" of distribution --- for <u>*every*</u> hypothesis in class, error on sample ~= error on distribution (i.e. risk)

REPRESENTATIVE SAMPLE

- Let H be a hypothesis class and X be an input space with a distribution D on it, and let f be a target function. Sample $S \subseteq X$ is said to be ϵ –"representative" if **for all** h in H, we have:

  $| \frac{1}{|S|}$ error (S, h) - risk_D (h, f) $| < \epsilon$

- If we happen to get a representative sample, we have desired bound on risk!
- Is a sample representative "with high probability"?

RANDOM SAMPLE IS REPRESENTATIVE WHP!

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

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

- Note: 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 idea: first start with a single hypothesis $h \in H$; what is the probability that error on sample ~= error on D?

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

- For a class H and an input space X, we can define a notion of "growth function"

LEARNABILITY IN TERMS OF THE GROWTH FUNCTION

- Theorem: Suppose $\tau_H(m)$ is an upper bound on the total number of "distinct 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: