THEORY OF MACHINE LEARNING

LECTURE 8

FUNDAMENTAL THEOREM OF STATISTICAL ML, INTRO TO OPTIMIZATION

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) | $\leq \epsilon$
- How to show that random sample is representative whp, for an *infinite* hypothesis class (Chernoff + Union bound fails)
- Growth function $\tau_H(m)$; small growth function => random sample is representative
- Polynomial vs exponential!
- Shattering, VC dimension

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,

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

- If $\tau_H(m) \approx m^d$ for some parameter d then m ~ $\frac{d \log\left(\frac{d}{\epsilon}\right)}{\epsilon^2}$ makes the RHS < ϵ

LAST LECTURE – SHATTERING AND VC DIMENSION

- A hypothesis class H is said to shatter a set S if all possible classifications (all 2^|S| of them) can be obtained using hypotheses
- Intuitively for such a hyp class, giving the labels of a subset of S doesn't give any information about labels of other points!
- VC dimension: is the size of the largest set in X that can be shattered by H
- Examples: VC dimension of 1-D LTFs, etc.

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 – homework)

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 – homework)

SOME IMPLICATIONS

- If H has infinite VC dimension, it is **not** PAC learnable! (same proof as no-free-lunch theorem – homework)
- ERM is all you need, assuming you have enough samples
  - Doing ERM efficiently is a challenge (next section)
- Agnostic case usually as hard as *realizable* case
- **Caveat**. Learnability guarantees only apply to ERM, not (say) to an improper learner

OPTIMIZATION

HOW TO SOLVE ERM EFFICIENTLY?

BASICS

- **Linear classification**
- Linear classification – non realizable

- Loss functions
- Convexity and convex optimization