THEORY OF MACHINE LEARNING

LECTURE 3

PAC MODEL, VC DIMENSION

RECAP – VALIANT'S THEORY OF (SUPERVISED) LEARNING

- **Learnability (from examples).** [Suppose D is fixed.] We say that a concept class is "learnable" if there exists an [efficient] algorithm ***A*** with the property: for all $\epsilon > 0$, there exists *m* (number of samples) such that when *given m* *i.i.d.* samples from D along with their labels, ***A*** produces a hypothesis *h* with risk less than ϵ, with prob. >= 0.9

- (Recall, risk = expected error on sample from distribution)
- Beyond examples? (technically yes, e.g., teacher/student)

RECAP: NO FREE LUNCH THEOREM

- Motivation: do we really need to restrict the hypothesis/concept class *before starting* learning? – yes!
- No free lunch: (informal) there is no "universal" learner, even if it's allowed to be inefficient (even for binary classification under a uniform distribution, unless it "sees most of the labels")
- Proof via a counting argument – too many hypotheses

TODAY'S PLAN

- Definition. (Agnostic) PAC learning
- Finite classes are PAC learnable
- Dealing with infinite classes: 'growth function' and VC dimension

PAC LEARNING (REALIZABLE CASE)

- **Learnability of a concept class.** *A* concept class H is *PAC learnable* (over domain X) if there exists an algorithm ***A*** that for all $\epsilon, \delta > 0$ and distributions D, has the following property:
  - given $m(\epsilon, \delta)$ samples (x, f(x)), where x ~ D and f is a (unknown) function in H, it outputs *h* with risk at most ϵ with probability at least $1 - \delta$.
- (The sample size must not depend on D)
- *A*s such *h* need not belong to H (improper learning)

PAC LEARNING (NON-REALIZABLE CASE)

- **Learnability of a concept class.** *A* concept class H is **agnostically** *PAC learnable* (over domain X) if there exists an algorithm ***A*** that for all $\epsilon, \delta > 0$ and distributions D, has the following property:
  - given $m(\epsilon, \delta)$ samples (x, f(x)), where x ~ D and f is a (unknown) function not necessarily in H, it outputs *h* with risk at most ϵ **more than the risk of the** h′ in H that is **"closest" to f**, with probability at least $1 - \delta$.
- (The sample size must not depend on D)
- Again, *h* need not belong to H (improper learning)

EVERY FINITE CLASS IS PAC LEARNABLE (EVEN AGNOSTIC)

- Suppose H has only finitely many hypotheses
  (input space X may still be infinite)
- **Generic algorithm:** <u>empirical risk minimization</u> (ERM)
- Key idea: "representative sample"

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) $| < \epsilon$

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 <u>support</u> [a, b]. Then we have

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

WHAT ABOUT INFINITE CLASSES?

- Note: as long as sample is representative, we are good!
- 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”