THEORY OF MACHINE LEARNING

LECTURE 3

PAC MODEL, GENERALIZATION

RECAP – VALIANT'S THEORY OF (SUPERVISED) LEARNING

- **Formal definition of learning**
  - Formalizing generalization via "distributional assumption"
  - X: space of (all possible) inputs
  - Y: set of labels / outputs
  - "Ground truth label" (**concept**). $\ell: X \mapsto Y$: function mapping inputs to outputs
- Goal of learning
  - "Learn" a hypothesis *h* such that $h(x) = \ell(x)$ for all "inputs of interest"
  - Unknown probability distribution D over X; achieve small "risk" or "generalization error"
  - (Definition of risk): $\Pr_{\{x \sim X\}}[\, h(x) \neq \ell(x)\,]$

Most common formal model to reason about learning

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

- Beyond examples? (technically yes, e.g., teacher/student)

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

- Beyond examples? (technically yes, e.g., teacher/student)

TODAY'S PLAN

- Concept class (or class of hypothesis)
  - Assume that ground-truth label is (at least close to) a function in H
- "No free lunch theorem" (informal). There is no "universal" (concept class agnostic) learning algorithm
- (Agnostic) PAC learning
- Finite classes are PAC learnable

COMMON ML ASSUMPTIONS

- (90s) Data is (approx.) linearly separable
- (these days) There exists 100-layer NN with width < … that achieves low error on task

- "Inductive bias" – assuming specific structure on concept
- What class of models do we use?
- Maybe.. we don't need to start with knowing a concept class

NO FREE LUNCH THEOREM

- Informal: there is no "universal" learner, even if it's allowed to be inefficient (even for binary classification)
- **Theorem.** Let D be the uniform distribution on some input space X. Consider any (possibly randomized) algorithm A that uses < |X|/2 i.i.d. examples and produces $h : X \to \{0,1\}$. There exists a hypothesis h for which A incurs risk > 1/10, with probability > 1/10.
- *(Recall def of "learnable" – fails with $\epsilon = 1/10$ and failure prob. 0.1)*

INFORMAL PROOF

- Extra assumption: suppose A is deterministic; will show theorem with weaker constants
- **Main idea:** 'too many Boolean functions' on X

PROOF

PAC LEARNING

- Moral: must suppose H is a *known* class of hypotheses (concept class)

- **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, takes $m(\epsilon, \delta)$ samples and produces *h* with risk at most ϵ with probability at least $1 - \delta$.
- (The sample size must not depend on D)

EVERY FINITE CLASS IS PAC LEARNABLE