THEORY OF MACHINE LEARNING

LECTURE 3

PAC MODEL, GENERALIZATION

Homework 1:

RECAP – VALIANT'S THEORY OF (SUPERVISED) LEARNING

Naive: $h(x) = \ell(x) \ \forall x$

- **Formal definition of learning**
  - Formalizing generalization via "distributional assumption"
  - X: space of (all possible) inputs — D: prob. distribution over X
  - Y: set of labels / outputs (Classification)
  - "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): $\Pr_{\{x \sim \mathcal{D}\}} [\, h(x) \neq \ell(x) \,]$

Most common formal model to reason about learning

RECAP – VALIANT'S THEORY OF (SUPERVISED) LEARNING

input sample $\underline{S} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$.
(unknown to learner)

- **Learnability (from examples).** [Suppose D is fixed.] We say that a concept ~~class~~ 'l' 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)

learning algo. $\underline{A} : (\mathcal{X} \times \mathcal{Y})^m \longrightarrow \mathcal{H} \rightsquigarrow$ possible output hypotheses.

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

$\mathcal{H}$: { collection of functions $h: \mathcal{X} \to \mathcal{Y}$. }.

ℓ

known hyp. class.

- 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

simple structure on class of concepts/hypotheses: as long as you can choose the right set of features, ~~you ca~~ you can predict label via linear separators.

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

  (more complex ~~i~~ but still known — structure on hypothesis class.)

- "Inductive bias" – assuming specific structure on concept
- What class of models do we use? (for a given task.)
- 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* -> {0,1}. There exists a hypothesis *h* for which *A* incurs risk > 1/10, with probability > 1/10.

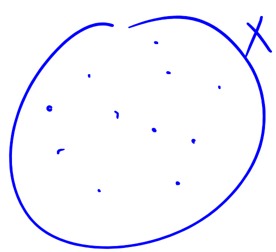
  0.9|X|

  |X|/16

  0.05

- *(Recall def of "learnable" – fails with* $\epsilon = 1/10$ *and failure prob. 0.1)*

  1



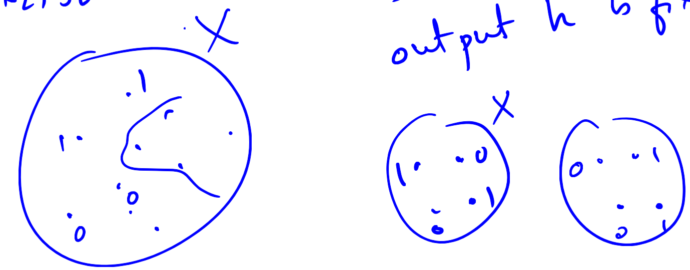
(∴ h produced by the theorem is not learnable..).

ε~1/20

|X|/16

INFORMAL PROOF

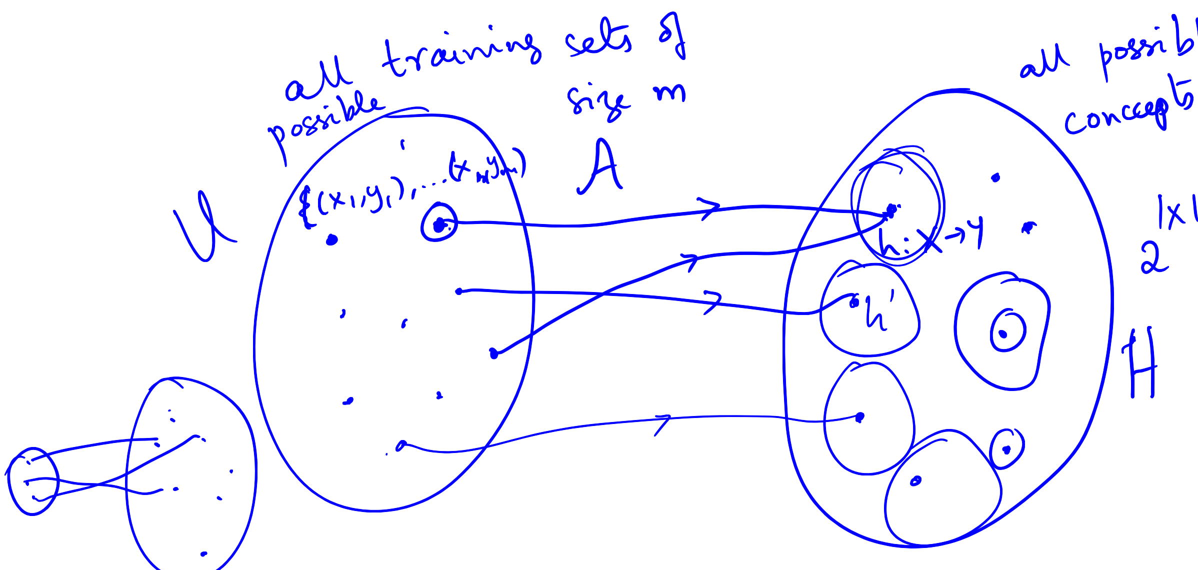
if input to A {i.e. $\{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\}$ is given, output h is fixed.



- Extra assumption: suppose A is deterministic; will show theorem with weaker constants

- **Main idea:** 'too many Boolean functions' on X

$2^{|X|}$ Boolean fns.



$|H| = 2^{|X|}$

$|U| \le \boxed{\binom{|X|}{m} \cdot 2^m}$

if $m < |X|/16$, it turns out that $|U| \ll |H|$

$\exists$ some h that cannot be learned exactly.

PROOF

→ [Gilbert - Varshamov bound].

- But ... the goal is to obtain an h' that agrees with h on 90% of inputs.. ($\Leftrightarrow$ risk is < 0.1)

h is a $|x|$-bit string.

(exercise) how many $|x|$-bit strings differ from h in $\leq |x|/10$ places?

$$\underbrace{\binom{|x|}{1} + \binom{|x|}{2} + \dots + \binom{|x|}{|x|/10}} \sim \binom{|x|}{|x|/10} \ll 2^{|x|}$$

Can verify: $\binom{|x|}{m} \cdot 2^m \cdot \binom{|x|}{|x|/10} < 2^{|x|}$

$\Rightarrow \exists h$ that differs in $> \frac{|x|}{10}$ places from ALL possible outputs of algo..

$\Rightarrow \exists h$ s.t. risk > 0.1 (with prob 1.)

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