CS60050 Machine Learning
IIT Kharagpur

PAC Learning

Somak Aditya, Sudeshna Sarkar

Goal of Learning Theory

To understand

- What kinds of tasks are ***learnable***?
- What kind of data is required for learnability?
- What are the (space, time) requirements of the learning algorithm.?

To develop and analyze models

- Develop algorithms that provably meet desired criteria
- Prove guarantees for successful algorithms

Goal of Learning Theory

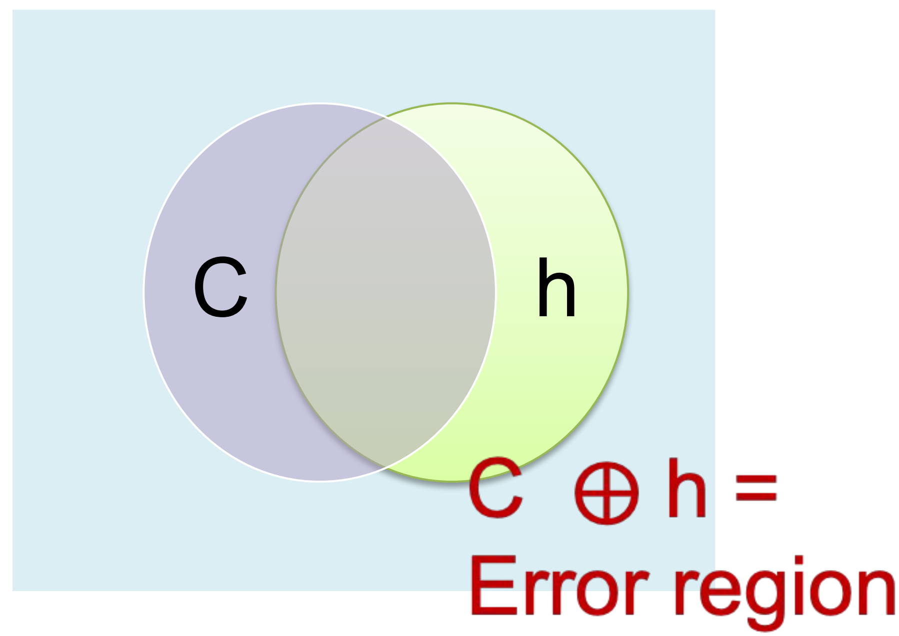
Two core aspects of ML

- Algorithm Design. How to optimize?
- Confidence for rule effectiveness on future data.

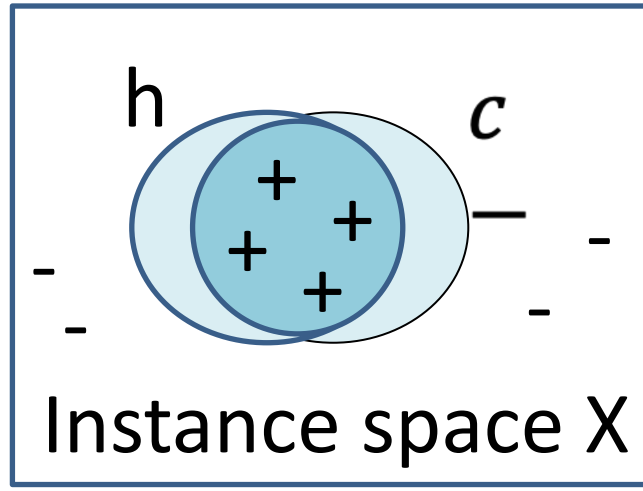
We need particular settings (models)

- Probably Approximately Correct (PAC)

Approximately correct $(P(c \oplus h) \leq \epsilon)$



Prototypical Concept Learning Task



Given

- Instances X (e.g., $X = R^d$ or $X = \{0,1\}^d$
- Distribution $\mathcal{D}$ over X
- Target function c
- Hypothesis Space $\mathcal{H}$
- Training Examples S = $\{(x_i, c(x_i))\}$ x_i i.i.d. from $\mathcal{D}$

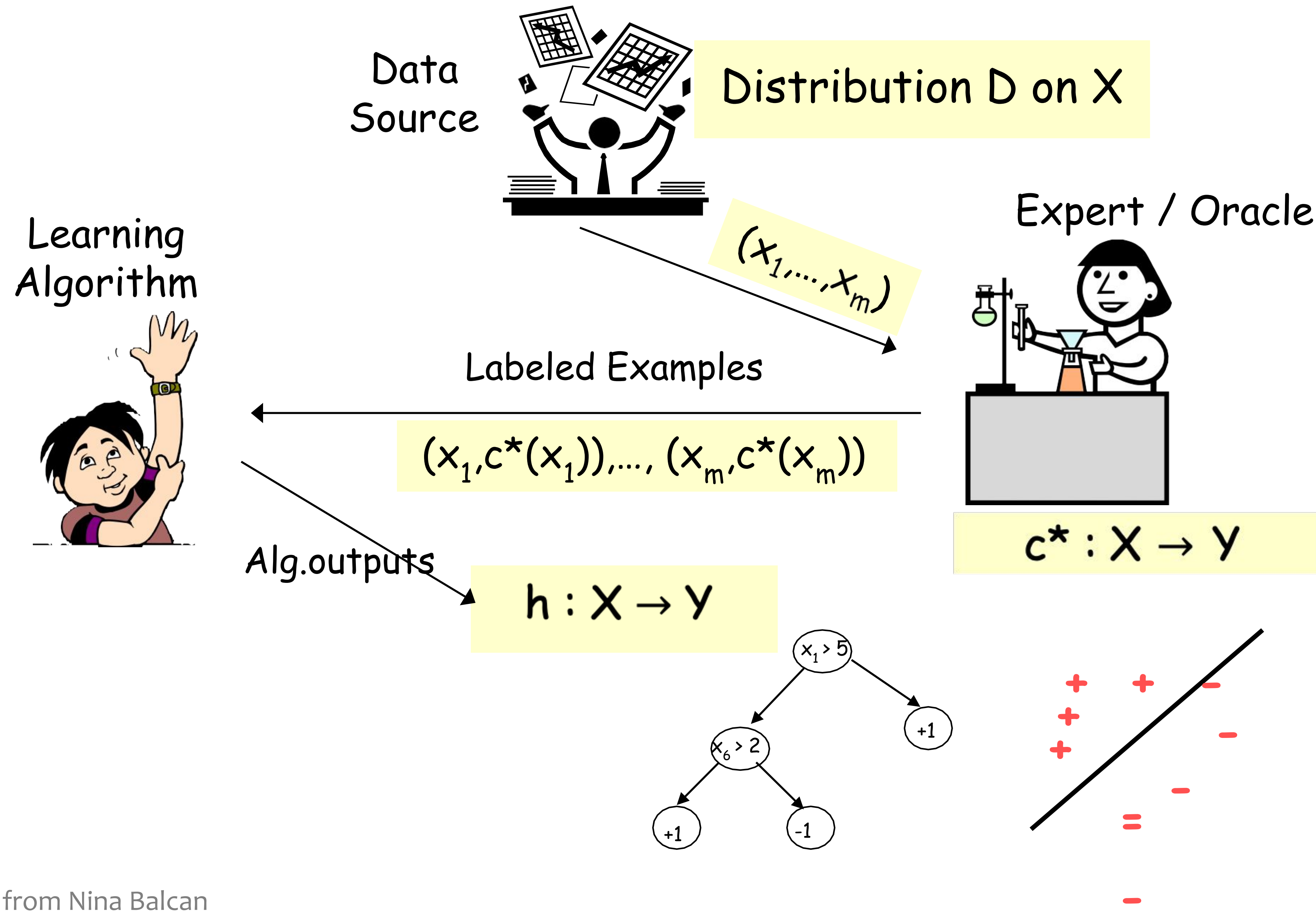
Determine

- A hypothesis h $\in \mathcal{H}$ s.t. $h(x) = c(x)$ for all x in S? consistent
- A hypothesis h $\in \mathcal{H}$ s.t. $h(x) = c(x)$ for all x in X?

ML--> An algorithm does optimization over S, find hypothesis h.

Goal: Find h which has small error over $\mathcal{D}$

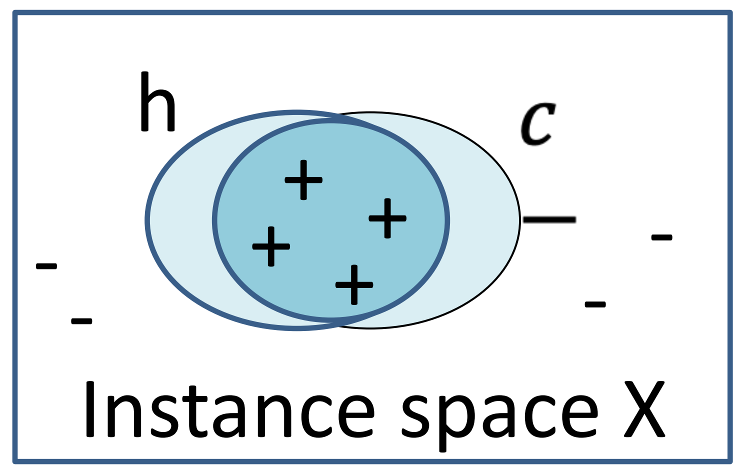
PAC/SLT models for Supervised Learning



Computational Learning Theory

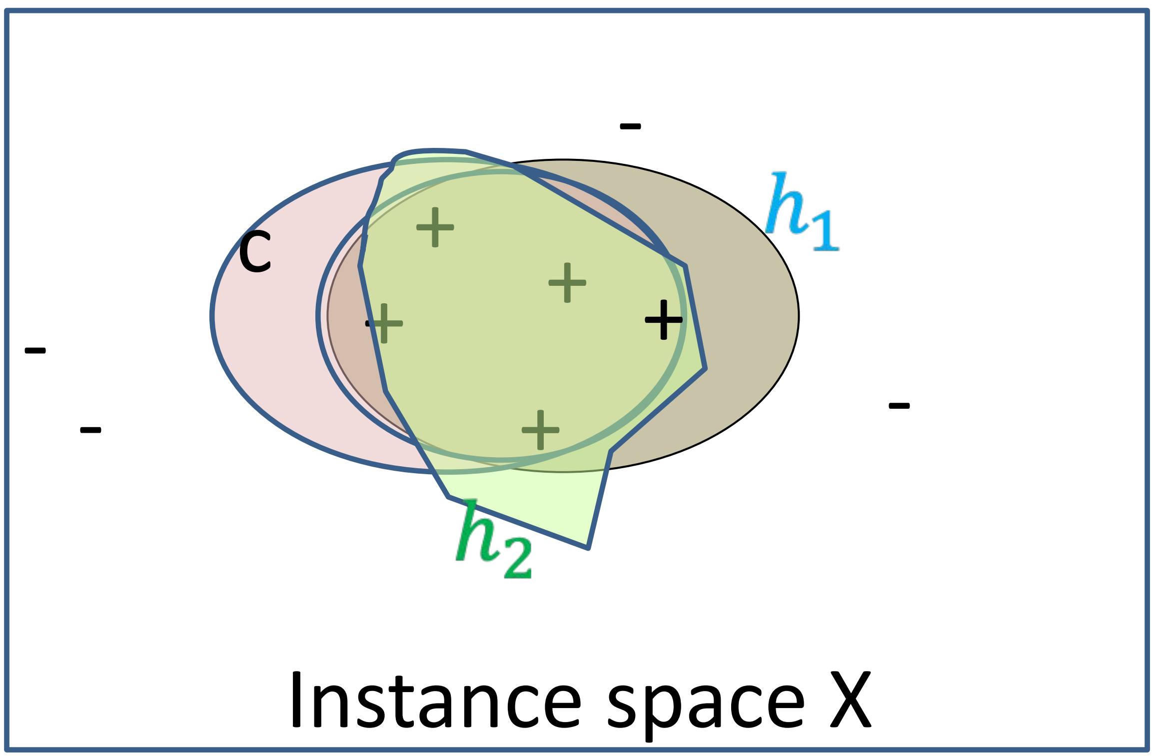
- *Can we be certain about how the learning algorithm generalizes?*
- *We would have to see all the examples.*

- *Inductive inference – generalizing beyond the training data is impossible unless we add more assumptions (e.g., priors over H)*
  - We need a bias!



Function Approximation

- *How many labeled examples in order to determine which of the 2^{2^N} hypothesis is the correct one?*
- *All 2^N instances in X must be labeled!*
- *Inductive inference: generalizing beyond the training data is impossible unless we add more assumptions (e.g., bias)*



$$H = \{h: X \rightarrow Y\}$$

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

Error of a hypothesis

*The **true error** of hypothesis h, with respect to the target concept c and observation distribution $\mathcal{D}$ is the probability that h will misclassify an instance drawn according to $\mathcal{D}$*

$$error_{\mathcal{D}}(h) = Pr_{x\sim\mathcal{D}}[c(x) \neq h(x)]$$

In a perfect world, we'd like the true error to be 0.

Bias: Fix hypothesis space H

c may not be in H => Find h close to c

A hypothesis h is approximately correct if

$$error_{\mathcal{D}}(h) \leq \varepsilon$$

PAC model

Goal: h has small error over D.

$$\text{True error: } error_D(h) = \Pr_{x \sim D}\big(h(x) \neq c^*(x)\big)$$

How often $h(x) \neq c^*(x)$ over future instances drawn at random from D

But, can only measure:

$$\text{Training error: } error_S(h) = \frac{1}{m}\sum_i I(h(x_i) \neq c^*(x))$$

How often $h(x) \neq c^*(x)$ over training Instances

Sample Complexity: bound $error_D(h)$ in terms of $error_S(h)$

Probably Approximately Correct Learning

PAC Learning concerns efficient learning

We would like to prove that

With high ***probability*** an (efficient) learning algorithm will find a hypothesis that is ***approximately*** identical to the hidden target concept.

We specify two parameters, ε and δ and require that

- *with probability at least $(1-\delta)$*
- *a system learn a concept with error at most ε.*

Sample Complexity for Supervised Learning

Theorem

$$m \geq \frac{1}{\epsilon}\left[In(|H|) + In\left(\frac{1}{\delta}\right)\right]$$

labeled examples are sufficient so that with prob. $1-\delta$*, all* $h \in H$ *with* $error_{\mathcal{D}}(h) \geq \epsilon$ *have* $error_S(h) > 0$.

- *inversely linear in* ϵ
- *logarithmic in* $|H|$
- ϵ *error parameter: D might place low weight on certain parts of the space*
- δ *confidence parameter: there is a small chance the examples we get are not representative of the distribution*

Sample Complexity for Supervised Learning

Theorem: $m \geq \frac{1}{\epsilon}\left[In(|H|) + In\left(\frac{1}{\delta}\right)\right]$ *labeled examples are sufficient so that with prob.* $1-\delta$, *all* $h \in H$ *with* $error_{\mathcal{D}}(h) \geq \epsilon$ *have* $error_S(h) > 0$.

Proof: Assume k bad hypotheses H_{bad}=$\{h_1, h_2, \ldots, h_k\}$ with

$$err_{\mathcal{D}}(h_i) \geq \epsilon$$

- *Fix* h_i. *Prob.* h_i *consistent with first training example is* $\leq 1-\epsilon$. *Prob.* h_i *consistent with first m training examples is* $\leq (1-\epsilon)^m$.
- *Prob. that at least one* h_i *consistent with first m training examples is* $\leq k(1-\epsilon)^m \leq |H|(1-\epsilon)^m$.
- *Calculate value of m so that* $|H|(1-\epsilon)^m \leq \delta$
- *Use the fact that* $1-x \leq e^{-x}$, *sufficient to set* $|H|e^{-\epsilon m} \leq \delta$

$$P(\text{consist}(H_{bad}, D)) \le |H| e^{-\varepsilon m} \le \delta$$

$$e^{-\varepsilon m} \le \frac{\delta}{|H|}$$

$$-\varepsilon m \le \ln\left(\frac{\delta}{|H|}\right)$$

$$m \ge \left(-\ln \frac{\delta}{|H|}\right) / \varepsilon \quad \text{(flip inequality)}$$

$$m \ge \left(\ln \frac{|H|}{\delta}\right) / \varepsilon$$

$$m \ge \left(\ln \frac{1}{\delta} + \ln |H|\right) / \varepsilon$$

Sample Complexity: Finite Hypothesis Spaces Realizable Case

PAC: How many examples suffice to guarantee small error whp.
Theorem

$$m \geq \frac{1}{\epsilon}\left[In(|H|) + In\left(\frac{1}{\delta}\right)\right]$$

labeled examples are sufficient so that with prob. $1-\delta$, *all* $h \in H$ *with* $err_{\mathcal{D}}(h) \geq \epsilon$ *have* $err_S(h) > 0$.

<u>Statistical Learning Way:</u>
With probability at least $1-\delta$, *all* $h \in H$ *s.t.* $err_S(h) = 0$ *we have*

$$err_D(h) \leq \frac{1}{m}\left[In(|H|) + In\left(\frac{1}{\delta}\right)\right]$$

Sample complexity: inconsistent finite $|\mathcal{H}|$

For a single hypothesis to have misleading training error

$$\Pr[error_{\mathcal{D}}(f) \leq \varepsilon + error_S(f)] \leq e^{-2m\varepsilon^2}$$

We want to ensure that the best hypothesis has error bounded in this way

So consider that any one of them could have a large error

$$\Pr[(\exists f \in \mathcal{H}) error_{\mathcal{D}}(f) \leq \varepsilon + error_S(f)] \leq |\mathcal{H}| e^{-2m\varepsilon^2}$$

From this we can derive the bound for the number of samples needed.

$$m \geq \frac{1}{2\varepsilon^2}(\ln|\mathcal{H}| + \ln(\frac{1}{\delta}))$$

Sample Complexity: Finite Hypothesis Spaces

Consistent Case

Theorem

$$m \geq \frac{1}{\epsilon}\left[In(|H|) + In\left(\frac{1}{\delta}\right)\right]$$

labeled examples are sufficient so that with prob. $1-\delta$, *all* $h \in H$ *with* $err_{\mathcal{D}}(h) \geq \epsilon$ *have* $err_S(h) > 0$.

Inconsistent Case

What if there is no perfect h?

Theorem: After m examples, with probability $\geq 1-\delta$, *all* $h \in H$ *have* $|err_{\mathcal{D}}(h) - err_S(h)| < \epsilon$, *for*

$$m \geq \frac{2}{2\,\epsilon^2}\left[In(|H|) + In\left(\frac{2}{\delta}\right)\right]$$

Sample complexity: example

$\mathcal{C}$: Conjunction of n Boolean literals. Is $\mathcal{C}$ PAC-learnable?

$$|\mathcal{H}| = 3^n$$

$$m \geq \frac{1}{\varepsilon}(n \ \ln 3 + \ln(\frac{1}{\delta}))$$

Concrete examples:

- $\delta=\varepsilon=0.05$, $n=10$ gives 280 examples
- $\delta=0.01$, $\varepsilon=0.05$, $n=10$ gives 312 examples
- $\delta=\varepsilon=0.01$, $n=10$ gives 1,560 examples
- $\delta=\varepsilon=0.01$, $n=50$ gives 5,954 examples

Result holds for any consistent learner, such as Find-S.

Sample Complexity of Learning Arbitrary Boolean Functions

Consider any boolean function over n boolean features such as the hypothesis space of DNF or decision trees. There are 2^{2^n} *of these, so a sufficient number of examples to learn a PAC concept is:*

$$m \geq \frac{1}{\varepsilon}(\ln 2^{2^n} + \ln(\frac{1}{\delta})) = \frac{1}{\varepsilon}(2^n \ln 2 + \ln(\frac{1}{\delta}))$$

δ=ε=0.05, n=10 gives 14,256 examples
δ=ε=0.05, n=20 gives 14,536,410 examples
δ=ε=0.05, n=50 gives $1.561 x 10^{16}$ *examples*

Questions For Today

1. Given a classifier with zero training error, what can we say about generalization error? (Sample Complexity, Realizable Case)

2. Given a classifier with low training error, what can we say about generalization error? (Sample Complexity, Agnostic Case)

3. Is there a theoretical justification for regularization to avoid overfitting? (Structural Risk Minimization)

Sample Complexity Results

Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four Cases we care about…

We started with the finite case…

	Realizable	Agnostic
Finite $\lvert\mathcal{H}\rvert$		
Infinite $\lvert\mathcal{H}\rvert$		

Sample Complexity Results

Definition 0.1. The **sample complexity** of a learning algorithm is the number of examples required to achieve arbitrarily small error (with respect to the optimal hypothesis) with high probability (i.e. close to 1).

Four Cases we care about…

	Realizable	Agnostic
Finite $\lvert\mathcal{H}\rvert$	$N \geq \frac{1}{\epsilon}\left[\log(\lvert\mathcal{H}\rvert) + \log(\frac{1}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $R(h) \geq \epsilon$ have $\hat{R}(h) > 0$.	
Infinite $\lvert\mathcal{H}\rvert$		

Learning Theory Objectives

You should be able to…

- Identify the properties of a learning setting and assumptions required to ensure low generalization error
- Distinguish true error, train error, test error
- Define PAC and explain what it means to be approximately correct and what occurs with high probability
- Apply sample complexity bounds to real-world learning examples
- Distinguish between a large sample and a finite sample analysis
- Theoretically motivate regularization

Example: Conjunctions

In-Class Quiz:

Suppose H = class of conjunctions over **x** in $\{0,1\}^M$

If M = 10, ε = 0.1, δ = 0.01, how many examples suffice?

	Realizable	Agnostic
Finite $\|\mathcal{H}\|$	$N \geq \frac{1}{\epsilon}\left[\log(\|\mathcal{H}\|) + \log(\frac{1}{\delta})\right]$ labeled examples are sufficient so that with probability $(1-\delta)$ all $h \in \mathcal{H}$ with $R(h) \geq \epsilon$ have $\hat{R}(h) > 0$.	
Infinite $\|\mathcal{H}\|$		

Concept Learning Task

"Days in which Aldo enjoys swimming"

Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

- Hypothesis Representation: Conjunction of constraints on the 6 instance attributes
    - "?" : any value is acceptable
    - specify a single required value for the attribute
    - "∅" : that no value is acceptable

Concept Learning

h = (?, Cold, High, ?, ?, ?)

indicates that Aldo enjoys his favorite sport on cold days with high humidity

Most general hypothesis: (?, ?, ?, ?, ?, ?)

Most specific hypothesis: (∅, ∅, ∅, ∅, ∅, ∅)

Find-S Algorithm

1. *Initialize h to the most specific hypothesis in $\mathcal{H}$*
2. *For each positive training instance x*

   For each attribute constraint a_i in h

   IF the constraint a_i in h is satisfied by x

   THEN do nothing

   ELSE replace a_i in h by next more general constraint satisfied by x
3. *Output hypothesis h*

Concept Learning

Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

<u>Finding a Maximally Specific Hypothesis</u>

Find-S Algorithm

$h_1 \leftarrow (\varnothing, \varnothing, \varnothing, \varnothing, \varnothing, \varnothing)$

$h_2 \leftarrow$ (Sunny, Warm, Normal, Strong, Warm, Same)

$h_3 \leftarrow$ (Sunny, Warm, ?, Strong, Warm, Same)

$h_4 \leftarrow$ (Sunny, Warm, ?, Strong, ?, ?)

Back