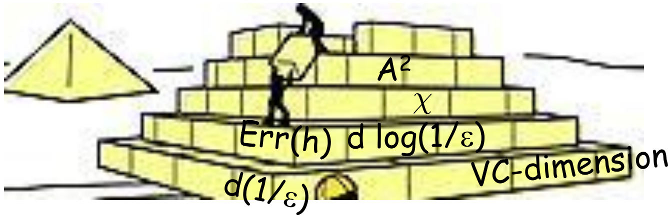
Machine Learning Theory

Maria-Florina (Nina) Balcan

February 9th, 2015

1. ML: 7.1, 7.2, 7.3.1, 7.4.1–7.4.3

2. 林轩田: MLF



Goals of Machine Learning Theory

Develop & analyze models to understand:

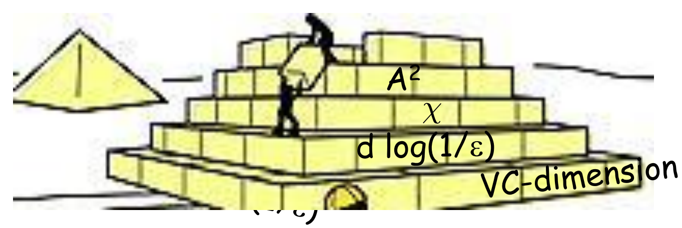
- what kinds of tasks we can hope to learn, and from what kind of data; what are key resources involved (e.g., data, running time)
- prove guarantees for practically successful algs (when will they succeed, how long will they take?)
- develop new algs that provably meet desired criteria (within new learning paradigms)

Interesting tools & connections to other areas:

- Algorithms, Probability & Statistics, Optimization, Complexity Theory, Information Theory, Game Theory.

Very vibrant field:

- Conference on Learning Theory
- NIPS, ICML



Today's focus: Sample Complexity for Supervised Classification (Function Approximation)

- Statistical Learning Theory (Vapnik)
- PAC (Valiant)

Probably Approximately Correct (PAC)

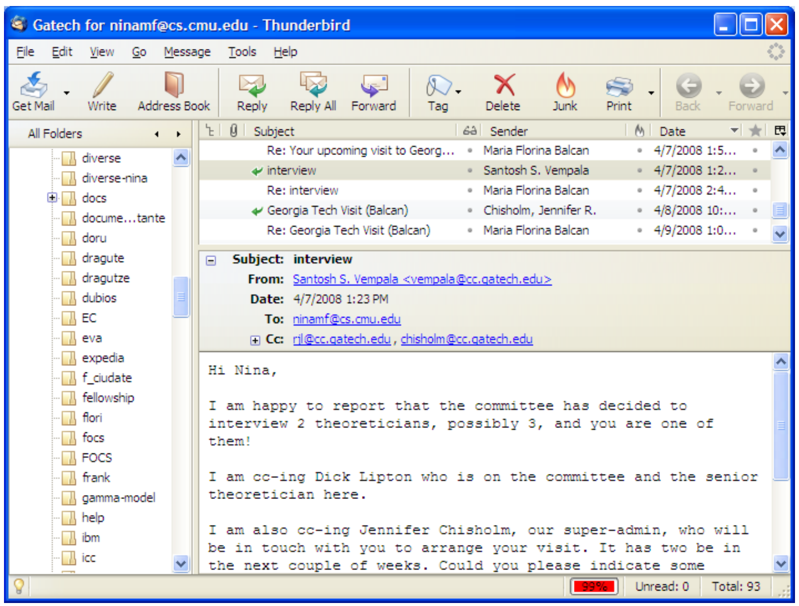
- Recommended reading: Mitchell: Ch. 7
  - Suggested exercises: 7.1, 7.2, 7.7
- Additional resources: my learning theory course!

Supervised Classification

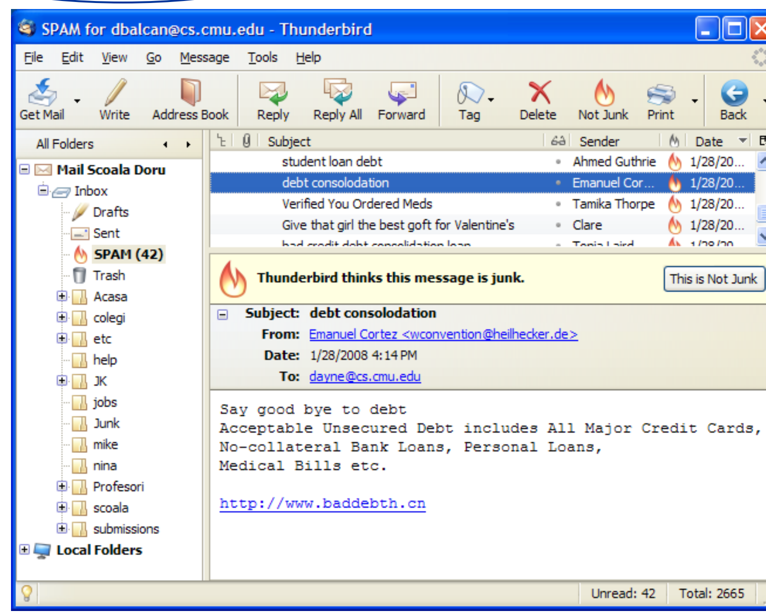
Decide which emails are spam and which are important.

Supervised classification

Not spam



spam



Goal: use emails seen so far to produce good prediction rule for future data.

Example: Supervised Classification

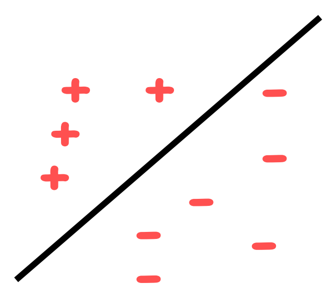
Represent each message by features. (e.g., keywords, spelling, etc.)

	"money"	"pills"	"Mr."	bad spelling	known-sender	spam?	
	Y	N	Y	Y	N	Y	
	N	N	N	Y	Y	N	
	N	Y	N	N	N	Y	
example	Y	N	N	N	Y	N	label
	N	N	Y	N	Y	N	
	Y	N	N	Y	N	Y	
	N	N	Y	N	N	N	

Reasonable RULES:

Predict SPAM if unknown AND (money OR pills)

Predict SPAM if 2money + 3pills –5 known > 0



Linearly separable

Two Core Aspects of Machine Learning

Algorithm Design. How to optimize?

Computation

Automatically generate rules that do well on observed data.

- E.g.: logistic regression, SVM, Adaboost, etc.

Confidence Bounds, Generalization

(Labeled) Data

Confidence for rule effectiveness on future data.

- Very well understood: Occam's bound, VC theory, etc.

- Note: to talk about these we need a precise model.

PAC/SLT models for Supervised Learning

Unknown : D, c^*

$\{(x_i, y_i)\}_{i=1}^m$

$y = c^*(x)$, $c^* \in C$

$\hat{y} = h(x)$; $h \in H$

① $c^* \in H$

② $c^* \notin H$

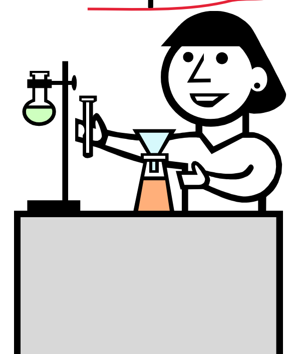
Data Source

Distribution D on X

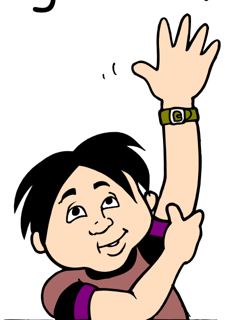
② noise-free

i.i.d. $(x_1, \dots, x_m)$

Expert / Oracle



Learning Algorithm



Labeled Examples

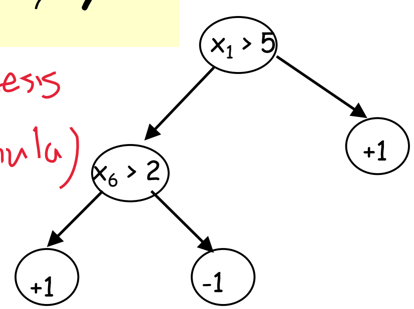
$S = (x_1, \underbrace{c^*(x_1)}_{y_1}), \dots, (x_m, \underbrace{c^*(x_m)}_{y_m})$

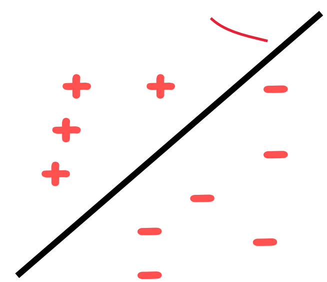
① $c^* : X \to Y = \{0, 1\}$

Alg.outputs

$h : X \to Y$

final hypothesis (learned formula)





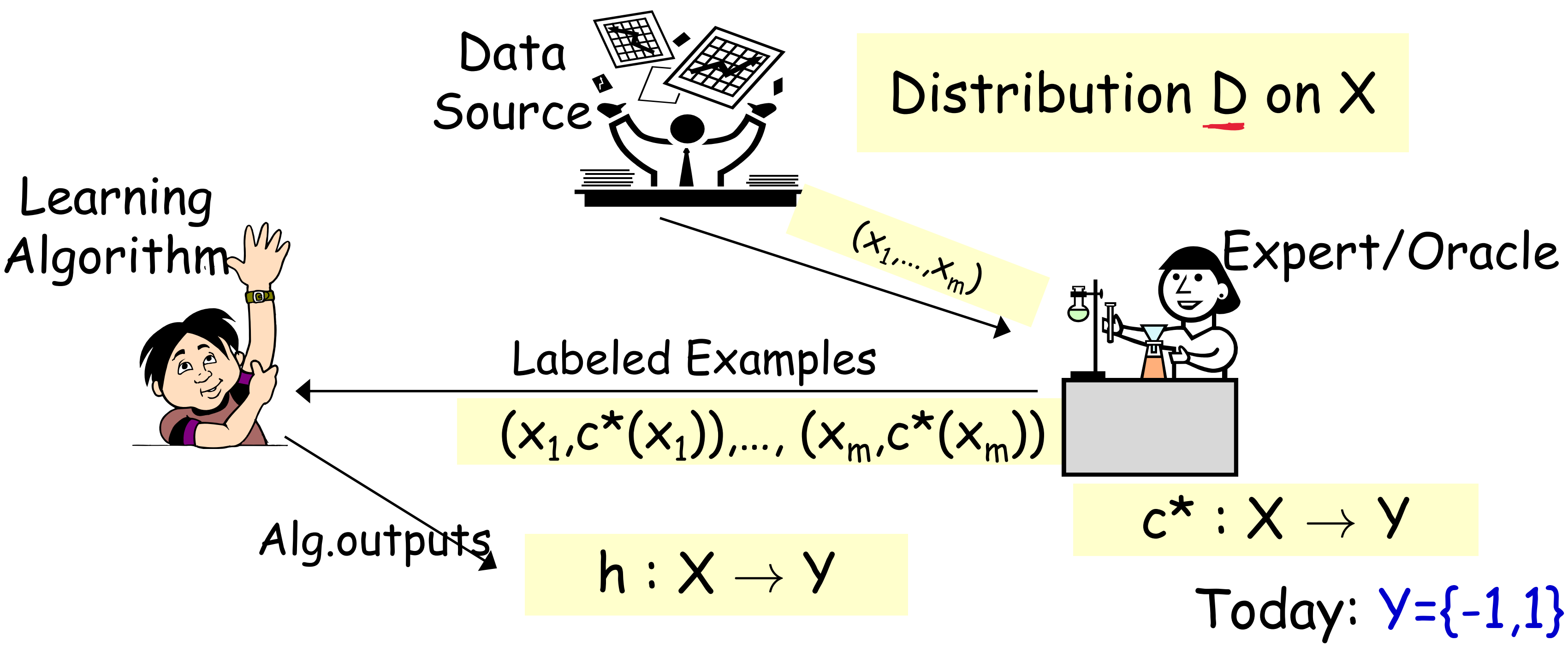
Hypothesis set H

(candidate formulas)

$X = \mathbb{B}^d$

$|H| = 2^d$

PAC/SLT models for Supervised Learning



- Algo sees training sample S: $(x_1,c^*(x_1)),\ldots, (x_m,c^*(x_m))$, x_i independently and identically distributed (i.i.d.) from D; labeled by c^*
- Does optimization over S, finds hypothesis h (e.g., a decision tree).
- Goal: h has small error over D.

$$err_D(h) \underbrace{\approx}_{\text{Testing}} \underbrace{err_S(h) \approx 0}_{\text{training}}$$

PAC/SLT models for Supervised Learning

- X – feature or instance space; distribution D over X

  e.g., $X = R^d$ or $X = \{0,1\}^d$

- Algo sees training sample S: $(x_1,c^*(x_1)),\ldots,(x_m,c^*(x_m))$, x_i i.i.d. from D
  - labeled examples - assumed to be drawn i.i.d. from some distr. D over X and labeled by some target concept c^*
  - labels $\in \{-1,1\}$ - binary classification

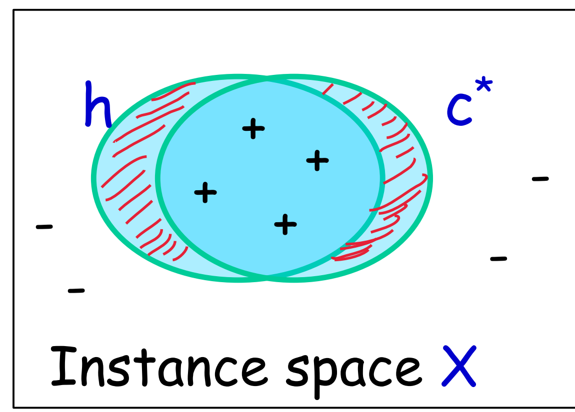
$c^*: X \to \{0,1\}$

- Algo does optimization over S, find hypothesis h.

- Goal: h has small error over D.

$$= E_D\big(h(x) \neq c^*(x)\big)$$

$$err_D(h) = \Pr_{x\sim D}\big(h(x) \neq c^*(x)\big)$$



$$L(h(x), c^*(x)) = \begin{cases} 1, & h(x) \neq c^*(x) \\ 0, & h(x) = c^*(x) \end{cases}$$

Need a bias: no free lunch.

$c^* \in H$

Function Approximation: The Big Picture

$\boxed{c^* \in H}$

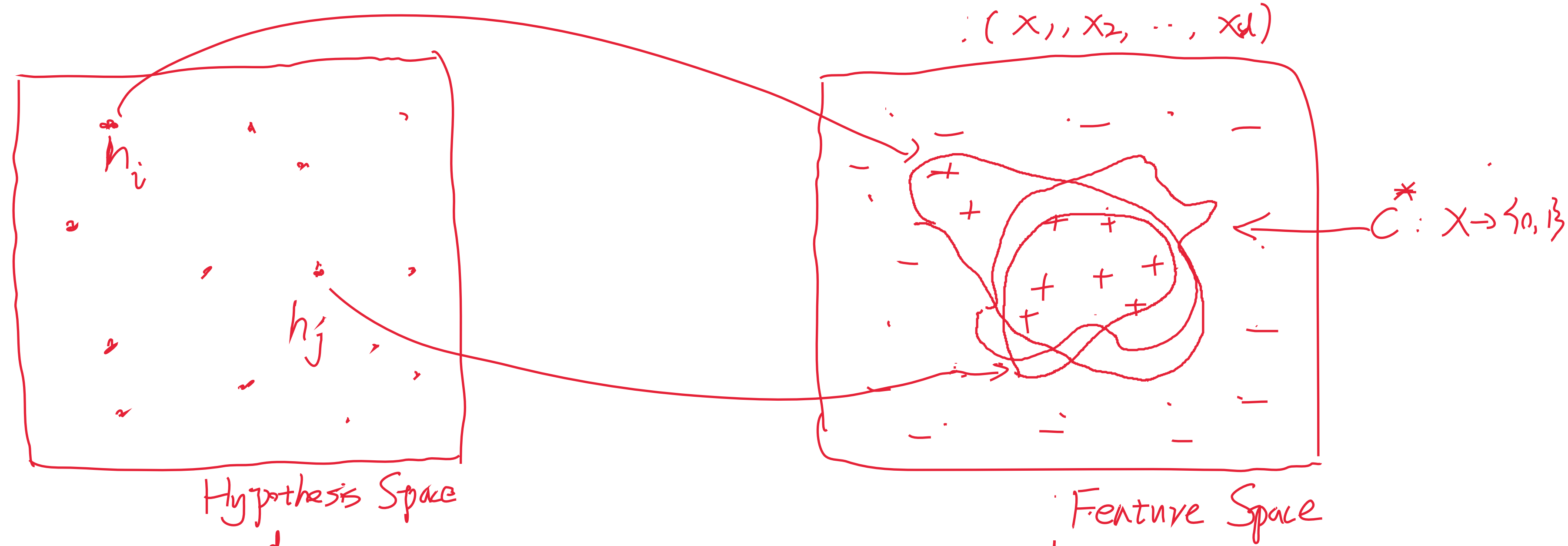
$H: \{h: X \to \{0,1\}\}$ $\qquad$ $X = \mathbb{B}^d$

$(x_1, x_2, \ldots, x_d)$

1-nn

k-nn

$\left(\frac{N}{k}\right)$



$|H| = 2^{2^d} = 2^{|X|}$ $\qquad$ $|X| = 2^d$

Q: How many labeled examples are needed in order to determine which of 2^{2^d} hypotheses is correct?

A: 2^d labeled examples

$2^d - 1$: 2 hypos. $\left\langle \begin{matrix} + \\ - \end{matrix} \right.$

$2^d - 2$: 2^2 hypos. $\left\langle \begin{matrix} + & + \\ + & - \\ - & + \\ - & - \end{matrix} \right.$

Additional assumption (Complexity)

PAC/SLT models for Supervised Learning

- X – feature or instance space; distribution D over X

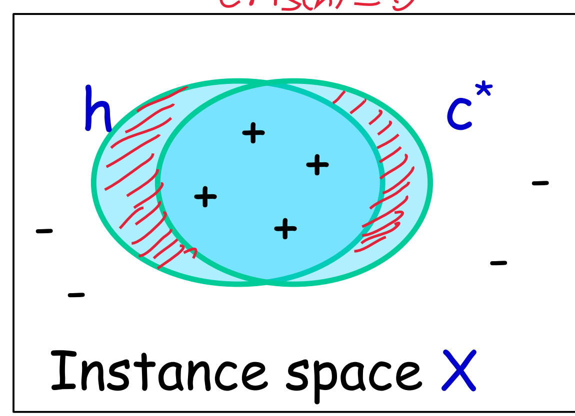
  e.g., $X = R^d$ or $X = \{0,1\}^d$

- Algo sees training sample S: $(x_1,c^*(x_1)),\ldots,(x_m,c^*(x_m))$, x_i i.i.d. from D
  - labeled examples - assumed to be drawn i.i.d. from some distr. D over X and labeled by some target concept c^*
  - labels $\in \{-1,1\}$ - binary classification
- Algo does optimization over S, find hypothesis h.
- Goal: h has small error over D.

$err_D(h) > 0$

$err_S(h) = 0$

$$err_D(h) = \Pr_{x\sim D}(h(x) \neq c^*(x))$$



Bias: Fix hypotheses space H .

(whose complexity is not too large).

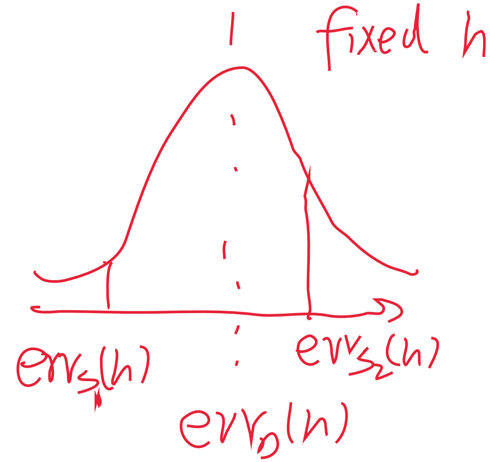
Realizable: $c^* \in H$.

Agnostic: c^* "close to" H. $c^* \notin H$

H : finite

PAC/SLT models for Supervised Learning

- Algo sees training sample S: $(x_1,c^*(x_1)),\ldots,(x_m,c^*(x_m))$, x_i i.i.d. from D
- Does optimization over S, find hypothesis $h \in H$.
- Goal: h has small error over D.



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

Generalization error

Expected risk

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

- But, can only measure:

$err_D(h) \simeq err_S(h) \simeq 0$

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

Empirical error

Empirical risk

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

Sample complexity: bound $err_D(h)$ in terms of $err_S(h)$

$\underline{err_D(h) \simeq err_S(h)}\ \underline{\simeq 0}$

Sample Complexity for Supervised Learning

- Consistent Learner
  - outputs hypothesis h that perfectly fits the training data S,

$$h(x) = c^*(x), \qquad \forall x \in S.$$

$err_S(h) = 0$

- Version Space (VS)
  - set of all hypotheses $h \in H$ that correctly classify the training data S,

$$VS_{H,S} = \{h \in H | \forall x \in S, h(x) = c^*(x)\}.$$

$\forall h, \; err_S(h) = 0$

$\forall$: \forall

$VS \subseteq H$

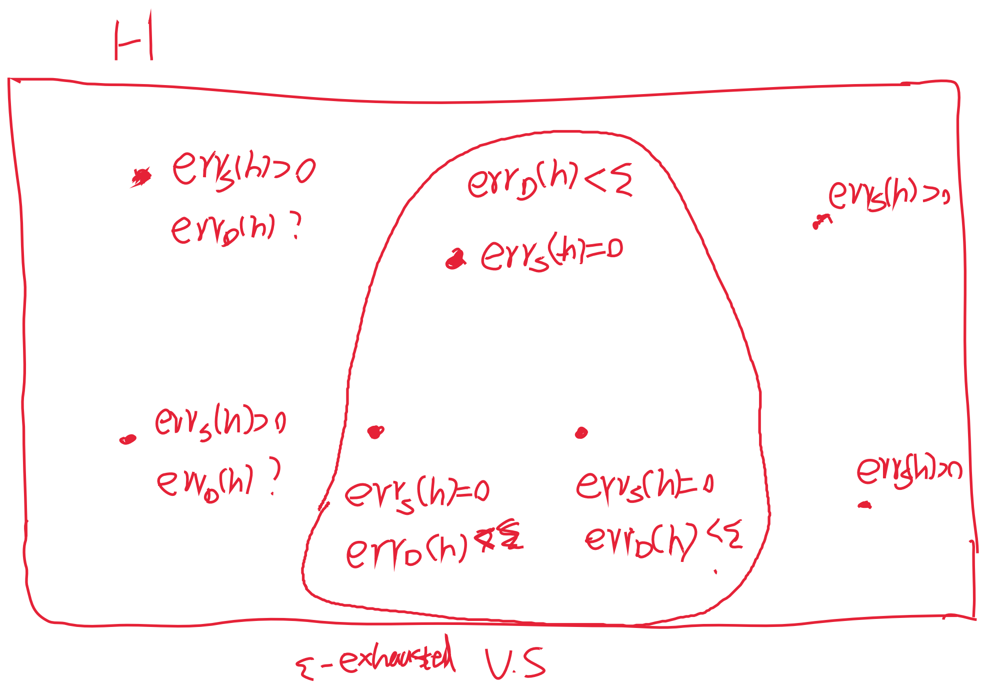
$\exists$: \exists

Sample Complexity for Supervised Learning

Definition: Consider a hypothesis space H, target concept c, instance distribution $\mathcal{D}$, and set of training examples D of c. The version space $VS_{H,D}$ is said to be **ϵ-exhausted** with respect to c and $\mathcal{D}$, if every hypothesis h in $VS_{H,D}$ has error less than ϵ with respect to c and $\mathcal{D}$.

$$(\forall h \in VS_{H,D})\ error_{\mathcal{D}}(h) < \epsilon$$

$err_S(h) = 0$



Sample Complexity for Supervised Learning

Theorem 7.1. **ϵ-exhausting the version space.** If the hypothesis space H is finite, and D is a sequence of $m \geq 1$ independent randomly drawn examples of some target concept c, then for any $0 \leq \epsilon \leq 1$, the probability that the version space $VS_{H,D}$ is not ϵ-exhausted (with respect to c) is less than or equal to

$$|H|e^{-\epsilon m}$$

$(\forall h, err_S(h) = 0)$

bad

$$Pr(\underbrace{\exists h \in VS, err_D(h) \geq \varepsilon}_{\text{bad}}) \leq |H| e^{-\varepsilon m} \leq \delta$$

$$1 - Pr(\underbrace{\forall h \in VS, err_D(h) < \varepsilon}_{\text{good}}) \leq \boxed{|H| e^{-\varepsilon m} \leq \delta}$$

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

whp

(with high probability)

$$Pr(\forall h \in VS, err_D(h) < \varepsilon) \geq 1 - \delta$$

Sample Complexity for Supervised Learning

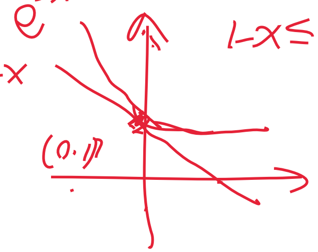
Proof: $h_1, h_2, \dots, h_k \in H$: bad hypotheses

$(err_D(h) > \varepsilon) \iff P_r(h(x) \neq c(x)) > \varepsilon$

if one of $\{h_k\}_{k=1}^{K}$ satisfies $err_S(h) = 0$,

then the V.S. is not ε-exhausted

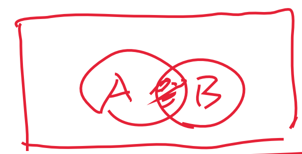
$1-x \le e^{-x}$



bad h: $P_r(h(x) = c^*(x)) \le 1-\varepsilon, \ \forall x \text{ in } D$

$P_r(h(S) = c^*(S)) \le (1-\varepsilon)^m \quad \forall S \sim D, \ |S| = m$

$P(A \cup B) \le P(A) + P(B)$



$\{h_k\}_{k=1}^{K}$:

$$P_r\left(h_1(S) = c^*(S) \cup h_2(S) = c^*(S) \cup \cdots \cup h_k(S) = c^*(S)\right)$$

$$\le \sum_{k=1}^{K} P_r(h(S) = c^*(S)) \le k(1-\varepsilon)^m \le |H|(1-\varepsilon)^m$$

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

Sample Complexity for Supervised Learning

Consistent Learner

- Input: S: $(x_1,c^*(x_1)),\ldots,(x_m,c^*(x_m))$
- Output: Find h in H consistent with the sample (if one exits).

Theorem

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

Quiz

labeled examples are sufficient so that with prob. $1-\delta$, all $h \in H$ with $\underbrace{err_D(h) \geq \varepsilon}_{A}$ have $\underbrace{err_S(h) > 0}_{B}$.

$err_S(h)=0 \Rightarrow err_D(h) < \varepsilon$

$A \Longrightarrow B$

$\neg B \Longrightarrow \neg A$

Contrapositive: if the target is in H, and we have an algo that can find consistent fns, then we only need this many examples to get generalization error $\leq \epsilon$ with prob. $\geq 1-\delta$

Sample Complexity for Supervised Learning

H finite

Consistent Learner

- Input: S: $(x_1,c^*(x_1)),\ldots,(x_m,c^*(x_m))$
- Output: Find h in H consistent with the sample (if one exits).

Theorem

Bound inversely linear in ϵ

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

labeled examples are sufficient so that with prob. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

Bound only logarithmic in |H|

- ϵ is called error parameter
  - D might place low weight on certain parts of the space
- δ is called confidence parameter
  - there is a small chance the examples we get are not representative of the distribution D

$S_1, S_2, \ldots, S_k$

$err_D(h) > \varepsilon$

Sample Complexity for Supervised Learning

Consistent Learner

- Input: S: $(x_1,c^*(x_1)),\ldots,(x_m,c^*(x_m))$
- Output: Find h in H consistent with the sample (if one exits).

Theorem

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

labeled examples are sufficient so that with prob. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

Example: H is the class of conjunctions over $X = \{0,1\}^n$. $|H| = 3^n$

E.g., $h = x_1 \overline{x_3} x_5$ or $h = x_1 \overline{x_2} x_4 x_9$

Then $m \geq \frac{1}{\epsilon}\left[n \ln 3 + \ln\left(\frac{1}{\delta}\right)\right]$ suffice

$n = 10, \epsilon = 0.1, \delta = 0.01$ then $m \geq 156$ suffice

Sample Complexity for Supervised Learning

$O(\ln n) < O(n) < O(n^k)$

Consistent Learner

- Input: S: $(x_1,c^*(x_1)),\ldots,(x_m,c^*(x_m))$
- Output: Find h in H consistent with the sample (if one exits).

Theorem

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

labeled examples are sufficient so that with prob. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

Example: H is the class of conjunctions over $X = \{0,1\}^n$.

Side HWK question: show that any conjunction can be represented by a small decision tree; also by a linear separator.

Sample Complexity for Supervised Learning

Theorem

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

labeled examples are sufficient so that with prob. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

Proof Assume k bad hypotheses $h_1, h_2, \ldots, h_k$ with $err_D(h_i) \geq \epsilon$

1) 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 $\leq (1-\epsilon)^m$.

2) Prob. that at least one h_i consistent with first m training examples is $\leq k\,(1-\epsilon)^m \leq |H|(1-\epsilon)^m$.

3) Calculate value of m so that $|H|(1-\epsilon)^m \leq \delta$

3) Use the fact that $1-x \leq e^{-x}$, sufficient to set $|H|\,e^{-\epsilon m} \leq \delta$

Sample Complexity: Finite Hypothesis Spaces

Realizable Case

Theorem

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

labeled examples are sufficient so that with prob. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

Probability over different samples of m training examples

Sample Complexity: Finite Hypothesis Spaces Realizable Case

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

Theorem

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

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

labeled examples are sufficient so that with prob. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

$A \Rightarrow B$

$\neg B \Rightarrow \neg A$

$C^* = VS$

$\Uparrow$

$C^* \in H$ finite

2) Statistical Learning Way:

With probability at least $1-\delta$, for all $h \in H$ s.t. $err_S(h) = 0$ we have

1. $C^* \in H$
2. H finite

$$err_D(h) \leq \frac{1}{m}\left(\ln|H| + \ln\left(\frac{1}{\delta}\right)\right).$$

$$err_D(h) \leq \boxed{err_S(h)} + \varepsilon$$

3. over estimate

Supervised Learning: PAC model (Valiant)

- X - instance space, e.g., $X = \{0,1\}^n$ or $X = R^n$
- $S_l=\{(x_i, y_i)\}$ - labeled examples drawn i.i.d. from some distr. D over X and labeled by some target concept c^*
  - labels $\in \{-1,1\}$ - binary classification

- Algorithm A PAC-learns concept class H if for any target c^* in H, any distrib. D over X, any $\varepsilon, \delta > 0$:
  - A uses at most poly(n,$1/\varepsilon$,$1/\delta$,size(c^*)) examples and running time.
  - With probab. $1-\delta$, A produces h in H of error at $\leq \varepsilon$.

$err_D(h) \leq \varepsilon$

probably

approximately

What if $c^* \notin H$?

Uniform Convergence

Theorem

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

labeled examples are sufficient so that with prob. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

- This basic result only bounds the chance that a bad hypothesis looks perfect on the data. What if there is no perfect $h \in H$ (agnostic case)?

- What can we say if $c^* \notin H$?
- Can we say that whp all $h \in H$ satisfy $|err_D(h) - err_S(h)| \leq \varepsilon$?
  - Called "uniform convergence".
  - Motivates optimizing over S, even if we can't find a perfect function.

$err_S(h) - \varepsilon \leq \boxed{err_D(h) \leq err_S(h) + \varepsilon}$

Sample Complexity: Finite Hypothesis Spaces

Realizable Case

$c^* \in H$

Theorem

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

labeled examples are sufficient so that with prob. $1-\delta$, all $h \in H$ with $err_D(h) \geq \varepsilon$ have $err_S(h) > 0$.

Agnostic Case

$c^* \notin H$

What if there is no perfect h?

Theorem After m examples, with probab. $\geq 1-\delta$, all $h \in H$ have $|err_D(h) - err_S(h)| < \varepsilon$, for

$$m \geq \frac{1}{2\varepsilon^2}\left[\ln(|H|) + \ln\left(\frac{2}{\delta}\right)\right]$$

To prove bounds like this, need some good tail inequalities.

Hoeffding bounds

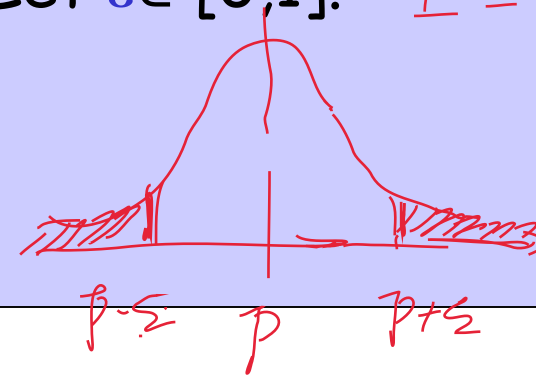
$\neq 1$

Consider coin of bias p flipped m times. $E(A) = p$

Let N be the observed # heads. Let $\varepsilon \in [0,1]$. $E \simeq \frac{N}{m}$

Hoeffding bounds:

- $\Pr[N/m > p + \varepsilon] \leq e^{-2m\varepsilon^2}$, and
- $\Pr[N/m < p - \varepsilon] \leq e^{-2m\varepsilon^2}$.



Exponentially decreasing tails

- Tail inequality: bound probability mass in tail of distribution (how concentrated is a random variable around its expectation).

$$err_D(h) = E_D(h(x) \neq c^*(x))$$

$$err_S(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}(h(x_i) \neq c^*(x_i))$$

$$Pr(|p - \frac{N}{m}| \geq \varepsilon) \leq 2e^{-2m\varepsilon^2}$$

$$Pr(p \geq \frac{N}{m} + \varepsilon) \leq e^{-2m\varepsilon^2}$$

$$\boxed{Pr(err_D(h) \geq err_S(h) + \varepsilon) \leq e^{-2m\varepsilon^2}}$$

$$\Pr(\underbrace{err_D(h) \geq err_S(h) + \varepsilon}_{\text{bad}}) \leq e^{-2m\varepsilon^2}$$

assume there is $\boxed{\text{at least one } h \in H \text{ satisfying the inequality}}$ Pr

$\Delta err(h) = err_D(h) - err_S(h)$

$$\Pr(\underbrace{\Delta err(h_1)}_{>\varepsilon} \cup \underbrace{\Delta err(h_2)}_{>\varepsilon} \cup \cdots \cup \underbrace{\Delta err(h_{|H|})}_{>\varepsilon})$$

$$\leq \sum_{j=1}^{|H|} \Pr(\Delta err(h_j) > \varepsilon) \leq \boxed{|H|\, e^{-2m\varepsilon^2} \leq \delta}$$

$$1 - \Pr(\underbrace{\forall h \in H,\ err_D(h) < err_S(h) + \varepsilon}_{\text{good}}) \leq \delta$$

$$\Pr(\text{good}) \geq 1 - \delta$$

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

Sample Complexity: Finite Hypothesis Spaces Agnostic Case

Theorem After m examples, with probab. $\geq 1-\delta$, all $h \in H$ have $|err_D(h) - err_S(h)| < \varepsilon$, for

$$\boxed{m \geq \frac{1}{2\varepsilon^2}\left[\ln(|H|) + \ln\left(\frac{2}{\delta}\right)\right] \qquad c^* \notin H}$$

- **Proof:** Just apply Hoeffding.
  - Chance of failure at most $2|H|e^{-2|S|\varepsilon^2}$.
  - Set to δ. Solve.
- So, whp, best on sample is ε-best over D.
  - Note: this is worse than previous bound ($1/\varepsilon$ has become $1/\varepsilon^2$), because we are asking for something stronger.
  - Can also get bounds "between" these two.

$$\boxed{\begin{gathered} c^* \in H \\ m \geq \frac{1}{\varepsilon}\left(\ln|H| + \ln\frac{1}{\delta}\right) \end{gathered}}$$

What you should know

- Notion of sample complexity.
- Understand reasoning behind the simple sample complexity bound for finite H.

H finite

$c^* \in H$

$c^* \notin H$

$err_D(h)$ $\quad$ $err_S(h)$

$c^* \in US$

$err_S(h)=0$

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

$err_D(h) \leq$ ______

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

ε ||

$err_D(h) \leq err_S(h) + \sqrt{\frac{1}{2m}\ln\frac{2|H|}{\delta}}$