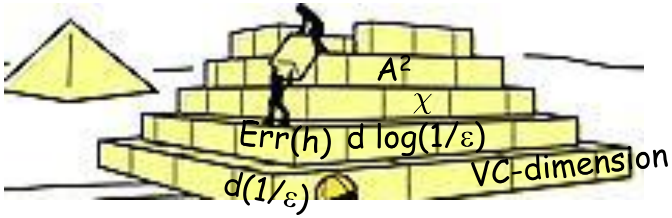
Machine Learning Theory

Maria-Florina (Nina) Balcan

February 9th, 2015

ML:
(7.1. 7.2, 7.3, 7.4.1 – 7.4.3.)

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)

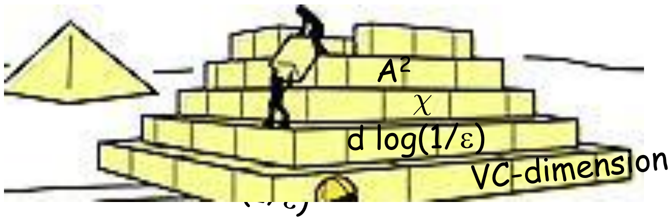
Interesting tools & connections to other areas:

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

Very vibrant field:

- Conference on Learning Theory
- NIPS, ICML

(NeurIPS)



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

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

Probably Approximately Correct

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

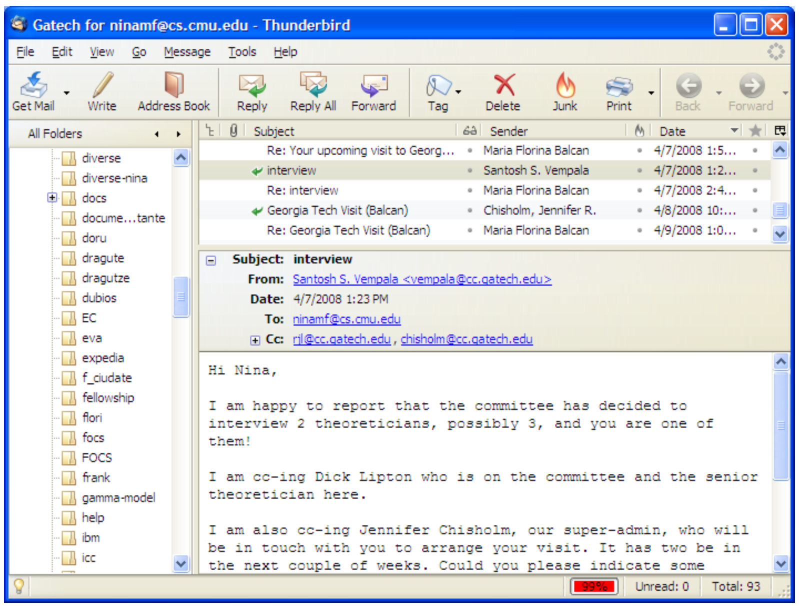
Supervised Classification

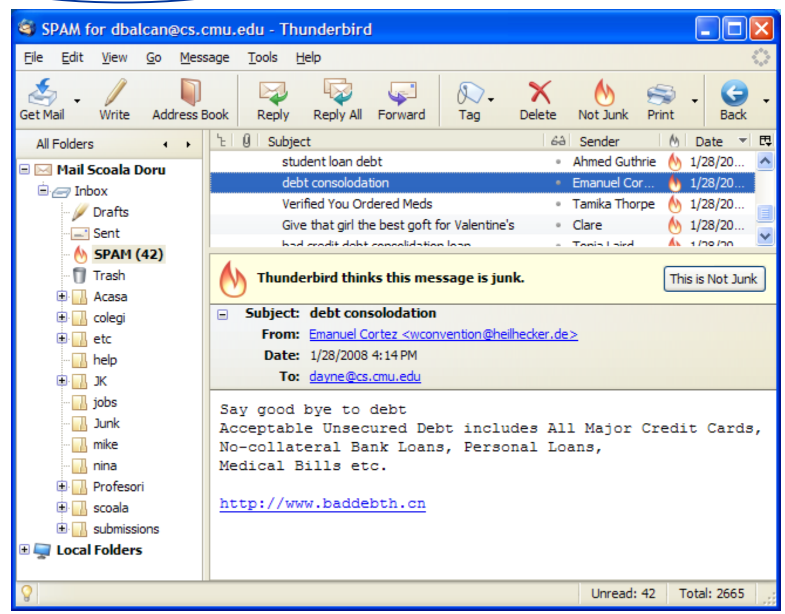
Decide which emails are spam and which are important.

Not spam

Supervised classification

spam





Goal: use emails seen so far to produce good prediction rule for future data. $h(\vec{x}) \to y$

ML

Data: $\{(x_j, y_j)\}_{j=1}^{m}$

Alg: $h: X \to y$

Goal: future incoming emails

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
Y	N	N	N	Y	N
N	N	Y	N	Y	N
Y	N	N	Y	N	Y
N	N	Y	N	N	N

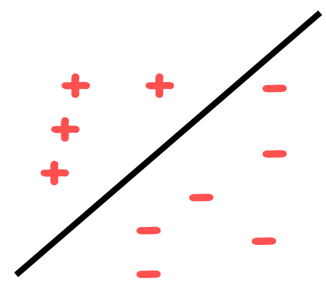
example

label

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.

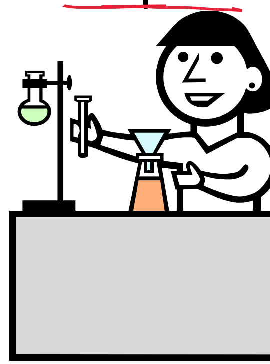
PAC/SLT models for Supervised Learning

Unkown.

Data Source

Distribution D on X

Expert / Oracle



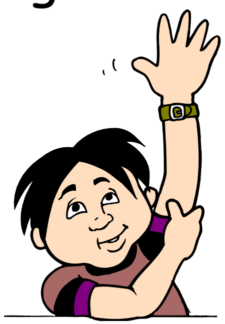
$c^*(x)$:

② noise-free

① binary

$(x_1,\ldots,x_m)$

Learning Algorithm



Labeled Examples

$S = (x_1,c^*(x_1)),\ldots,(x_m,c^*(x_m))$

$c^* : X \to Y \in \{-1,1\}$

Alg.outputs

$h : X \to Y$ $\{-1,1\}$

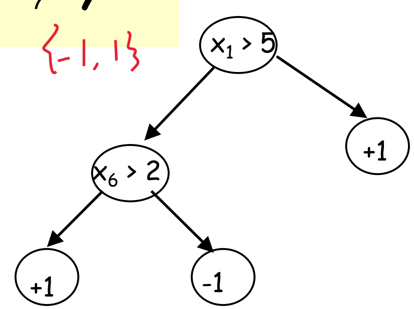
Hypothesis Space

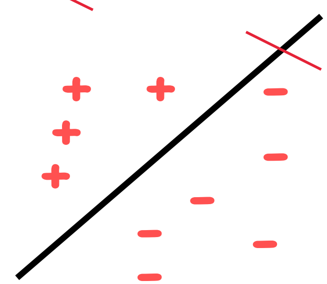
H. $\{h\}$. finite

e.g. $p(y|x) = \sigma(\beta^T x) = \dfrac{1}{1+e^{-\beta^T x}}$

$c^* \in H$: realizable

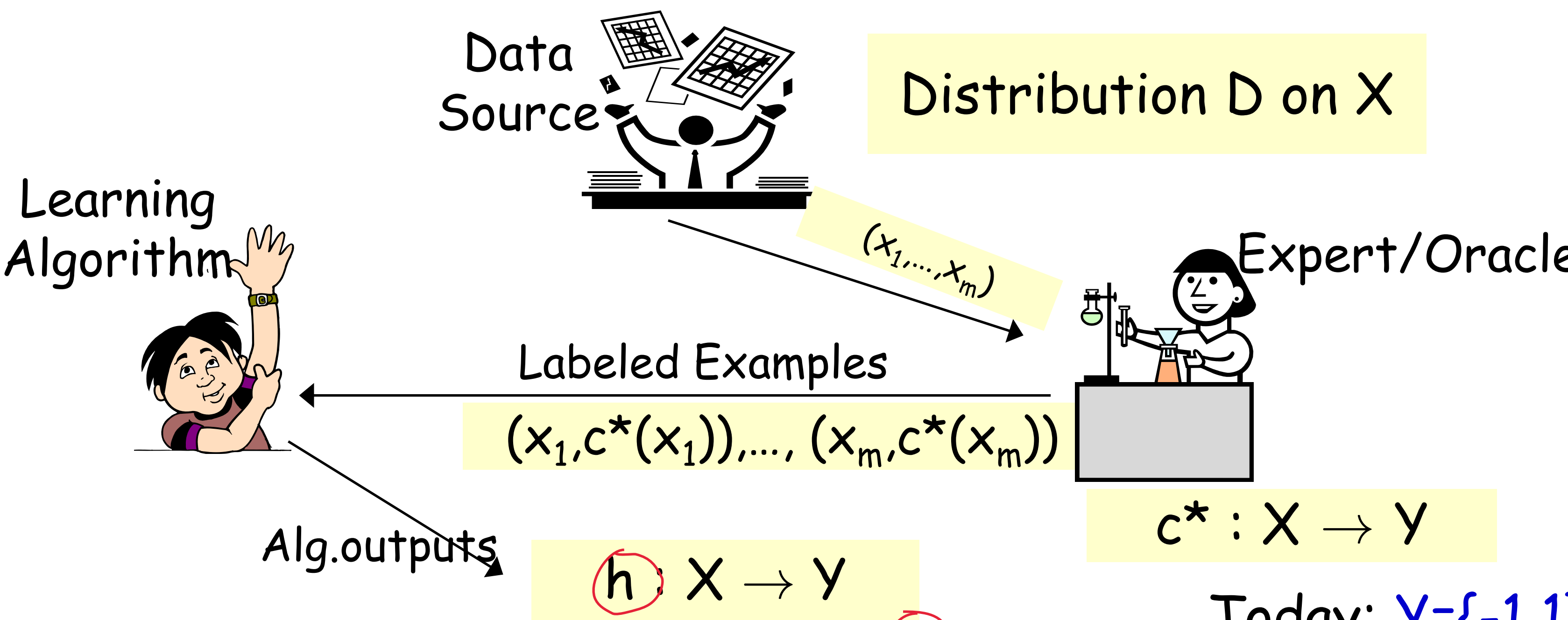
$c^* \notin H$: agonostic





$error_D(h) \le error_S(h) + \varepsilon$

PAC/SLT models for Supervised Learning



Today: $Y=\{-1,1\}$

- 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^* $h \in H$
- Does optimization over S, finds hypothesis h (e.g., a decision tree) $\min_{h \in H} error_S(h)$ $\min_{h \in H} error_S(h)$
- Goal: h has small error over D. $error_D(h) - error_S(h) \to 0$

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

$error_D(h) \neq 0$

- Algo does optimization over S, find hypothesis h.

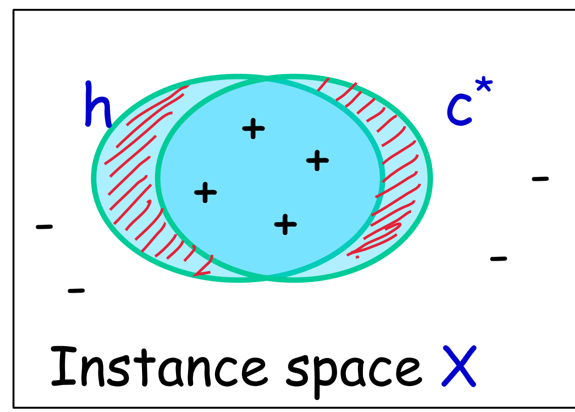
$error_S(h) = 0$

- Goal: h has small error over D.

pointwise

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

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



Need a bias: no free lunch.

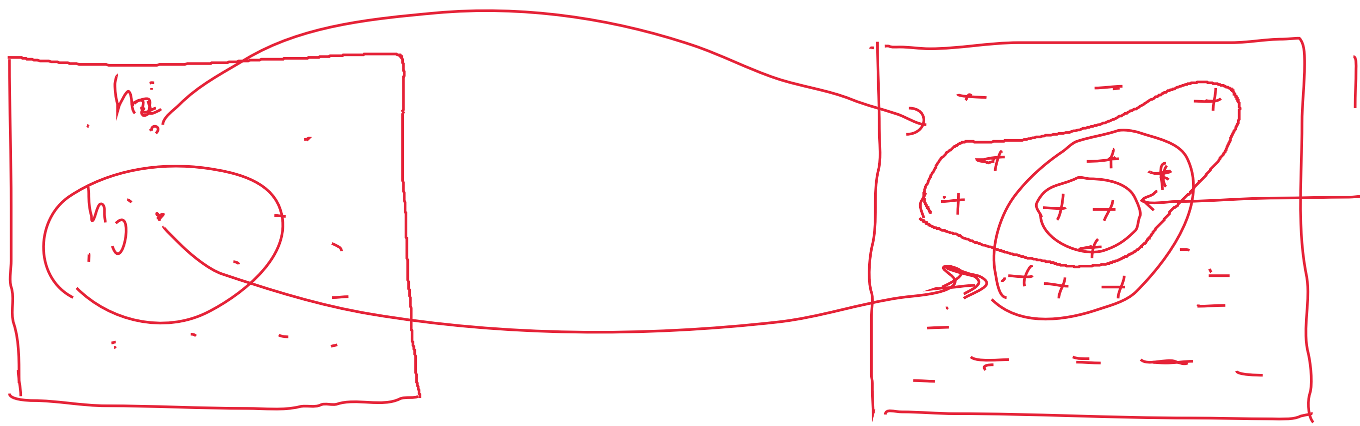
Function Approximation: The Big Picture

$(c^* \in H)$

$H: h \; X \to \{-1, 1\}$ $\qquad$ $\mathcal{X} = \{0, 1\}^n$ $\qquad$ S

$|H|$ finite $\qquad$ $|X| = 2^n$

c^*



Hypothesis Space $\qquad$ Feature Space

$h(S) = \{(h(x_1), \ldots, h(x_m))\}$

dichotomy

$|H| \leq 2^{2^n} = 2^{|X|}$

Additional assumption:

(Complexity)

Q. How many labeled examples are needed to determine which of 2^{2^n} hypos is c^*?

A: 2^n labeled examples

$2^n - 1$: $\quad +1 \to h_i$, $\quad -1 \to h_j$

$2^n - 2$: $\quad +1\ +1 \to h_a$, $\quad +1\ -1 \to h_b$, $\quad -1\ +1 \to h_c$, $\quad -1\ -1 \to h_d$

PAC/SLT models for Supervised Learning

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

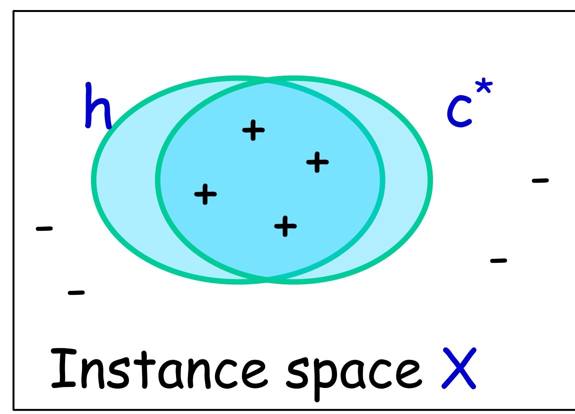
  e.g., $X = R^d$ or $X = \{0,1\}^d$

- 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.

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



Bias: Fix hypotheses space H.

(whose complexity is not too large).

Realizable: $c^* \in H$.

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

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

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. $E_D[h(x) \neq c^*(x)]$

1. True error: $err_D(h) = \Pr_{x\sim D}(h(x) \neq c^*(x))$
2. Expected error
3. Generalization error

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

- But, can only measure:

1. Training error: $err_S(h) = \frac{1}{m}\sum_i I(h(x_i) \neq c^*(x_i)) = \begin{cases} 1, & h(x) \neq c^*(x) \\ 0, & \text{otherwise.} \end{cases}$
2. Empirical error:

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

- Empirical Error Minimization.

$\min_{h\in H} err_S(h)$

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

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

Sample Complexity for Supervised Learning

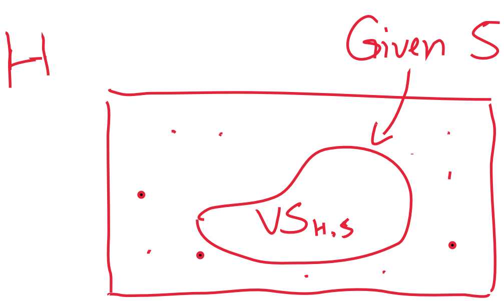
- Consistent Learner
  - outputs hypothesis h that perfectly fits the training data S,

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

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

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

$err_S(h) = 0$



$\forall$: for all

$\exists$: exist

Sample Complexity for Supervised Learning

$(c^* \in H)$

Definition: Consider a hypothesis space H, target concept c, instance distribution $\mathcal{D}$, and set of training examples ~~D~~ S 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 $\mathcal{D}$.

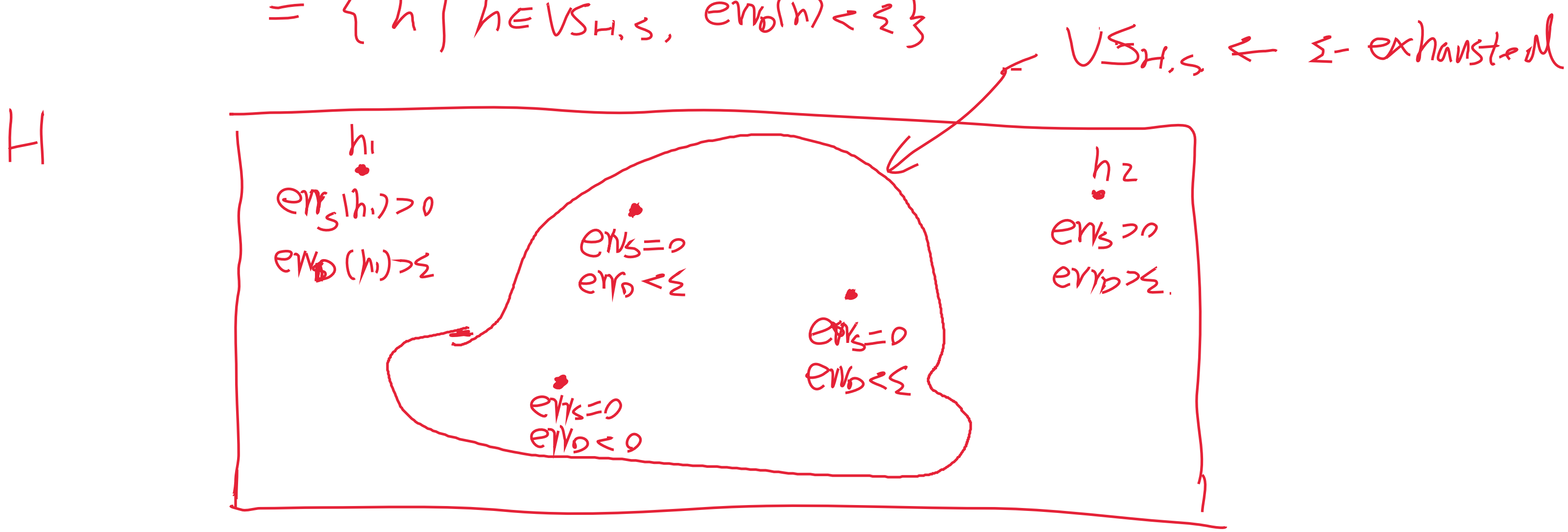
$(0 < \varepsilon < \frac{1}{2})$

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

$$VS_{H,S} = \{ h \mid h \in H,\ err_S(h) = 0 \}$$

$$\varepsilon\text{-exhausted } VS_{H,S} = \{ h \mid h \in H,\ err_S(h) = 0,\ \underline{err_D(h) < \varepsilon} \}$$

$$= \{ h \mid h \in VS_{H,S},\ err_D(h) < \varepsilon \}$$



Sample Complexity for Supervised Learning

① $c^* \in H$

② H finite

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

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

$$P(\overbrace{\exists h \in VS_{H,S},\ err_D(h) \ge \varepsilon}^{\text{bad}}) \le |H| e^{-\varepsilon m}$$

$$1 - P(\forall h \in VS_{H,S},\ err_D(h) < \varepsilon) \le \boxed{|H| e^{-\varepsilon m} \le \delta}, \qquad (0 < \delta < \tfrac{1}{2})$$

$$\underbrace{P(\forall h \in VS_{H,S},\ err_D(h) < \varepsilon)}_{\text{good } (\varepsilon\text{-exhausted})} \ge 1 - \delta$$

with high prob. (whp)

$$\ln|H| - \varepsilon m \ln e \le \ln \delta$$

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

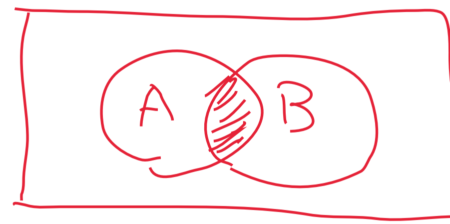
Sample Complexity for Supervised Learning

$$P(\exists h \in VS_{H,S},\ err_D(h) \ge \varepsilon) \le |H| e^{-\varepsilon m}$$

Proof: if $\exists h \in H$, $err_S(h)=0$, $err_D(h) \ge \varepsilon$, then VS is not ε-exhausted.

$(h \in VS_{H,S})$

$\underbrace{\hphantom{xxxxxxxxxxxxxxxxxxxxxxxxxxxx}}_{\text{Calculate the prob.}}$



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

$$err_D(h) \ge \varepsilon \Rightarrow P(h(x) \ne c^*(x)) \ge \varepsilon$$

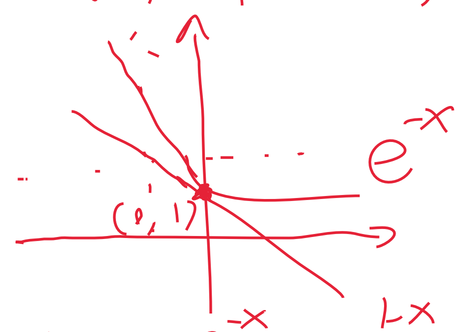
$$\Rightarrow 1 - P(h(x) = c^*(x)) \ge \varepsilon$$

$$\Rightarrow P(h(x) = c^*(x)) \le 1 - \varepsilon$$

$$P(h(S) = c^*(S))$$

$$\Rightarrow P(h(x_1) = c^*(x_1), \ldots, h(x_m) = c^*(x_m))$$

$$\overset{i.i.d.}{=} \prod_{i=1}^{m} P(h(x_i) = c^*(x_i)) \le (1-\varepsilon)^m$$



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

$$P(h_1(S)=c^*(S) \cup h_2(S)=c^*(S) \cup \cdots \cup h_{|H|}(S)=c^*(S))$$

$$\le \sum_{l=1}^{|H|} P(h_l(S)=c^*(S)) \le \sum_{l=1}^{|H|} (1-\varepsilon)^m = |H|(1-\varepsilon)^m \le |H| e^{-\varepsilon m}$$

Sample Complexity for Supervised Learning

($c^* \in H$! $err_S(h)=0$

Consistent Learner

$err_D(h) \le err_S(h) + \varepsilon$

- 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

$\Rightarrow \varepsilon \ge \frac{1}{m}\left[\ln|H| + \ln\frac{1}{\delta}\right]$

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

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

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

$A \Rightarrow B$

$\neg B \Rightarrow \neg A$

$\neg B$ $\neg A$

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 $\le \epsilon$ with prob. $\ge 1-\delta$

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

A

$\neg A$: $err_D(h) \le \varepsilon \le \boxed{\frac{1}{m}\left[\ln|H| + \ln\frac{1}{\delta}\right]}$

SLT

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

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

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

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

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

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

2) Statistical Learning Way:

With probability at least $1-\delta$, for all $\mathrm{h} \in \mathrm{H}$ s.t. $\mathrm{err}_{\mathrm{S}}(\mathrm{h}) = 0$ we have

$$\mathrm{err}_{\mathrm{D}}(\mathrm{h}) \leq \frac{1}{\mathrm{m}}\left(\ln|\mathrm{H}| + \ln\left(\frac{1}{\delta}\right)\right).$$

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 $\leq \varepsilon$.

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.

$err_S(h) - \varepsilon \leq err_D(h) \leq err_S(h) + \varepsilon$

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 $err_S(h) > 0$. $\Rightarrow$ $err_S(h)=0$ $\Rightarrow$ $err_D(h) \leq \varepsilon$

Agnostic Case $c^* \notin H$

What if there is no perfect h?

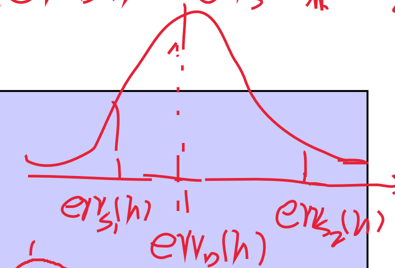
Theorem After m examples, with probab. $\geq 1-\delta$, all $h \in H$ have $|err_D(h) - err_S(h)| < \varepsilon$, for

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

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

Hoeffding bounds

$|err_D(h) - err_S(h)| < \varepsilon$

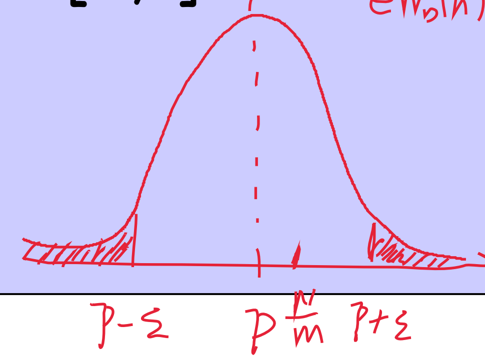


Consider coin of bias p flipped m times.

Let N be the observed # heads. Let $\varepsilon \in [0,1]$.

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

$$\Rightarrow P\left(\left|\frac{N}{m} - p\right| > \varepsilon\right) \leq 2e^{-2m\varepsilon^2}$$

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

$$P(|err_S(h) - err_D(h)| > \varepsilon) \leq 2e^{-2m\varepsilon^2}$$

$$err_D(h) = \Pr_D(h(x) \neq c^*(x))$$
$$= E_D[h(x) \neq c^*(x)]$$
$$err_S(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{h(x_i) \neq c^*(x_i)\}$$

$$P(err_D(h) > err_S(h) + \varepsilon) \leq e^{-2m\varepsilon^2}$$

- if $\exists h \in H$, $err_D(h) > err_S(h) + \varepsilon$

Calculate the prob.

$$P(\exists h \in H, err_D(h) > err_S(h) + \varepsilon) \leq |H| \cdot P(err_D(h) > err_S(h) + \varepsilon)$$
$$\leq |H| e^{-2m\varepsilon^2}$$

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

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

- 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.

What you should know

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