# Computational Learning Theory

Reading:

- Mitchell chapter 7

Suggested exercises:

- 7.1, 7.2, 7.5, 7.7

Machine Learning 10-701

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

November 1, 2010

# Computational Learning Theory

What general laws constrain inductive learning?

We seek theory to relate:

- Probability of successful learning
- Number of training examples
- Complexity of hypothesis space
- Accuracy to which target function is approximated
- Manner in which training examples presented

# Sample Complexity

Want to learn $F: X \to Y$
$C: X \to Y$

How many training examples are sufficient to learn the target concept?

1. If learner proposes instances, as queries to teacher
   - Learner proposes instance $x$, teacher provides $c(x)$
2. If teacher (who knows $c$) provides training examples
   - teacher provides sequence of examples of form $\langle x, c(x) \rangle$
3. If some random process (e.g., nature) proposes instances
   - instance $x$ generated randomly, teacher provides $c(x)$

$P(X)$

# Sample Complexity: 3

Given:

- set of instances $X$
- set of hypotheses $H$
- set of possible target concepts $C$
- training instances generated by a fixed, unknown probability distribution $\mathcal{D}$ over $X$

$c : X \to \{0,1\}$, $c \in C$

$\mathcal{D} = P(X)$

Learner observes a sequence $D$ of training examples of form $\langle x, c(x) \rangle$, for some target concept $c \in C$

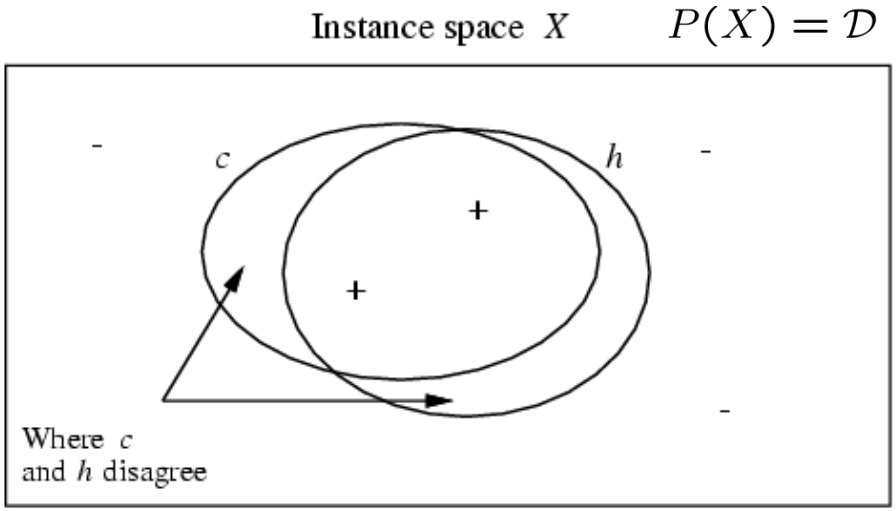
- instances $x$ are drawn from distribution $\mathcal{D}$
- teacher provides target value $c(x)$ for each

Learner must output a hypothesis $h$ estimating $c$

- $h$ is evaluated by its performance on subsequent instances drawn according to $\mathcal{D}$

Note: randomly drawn instances, noise-free classifications

# True Error of a Hypothesis




**Definition:** The **true error** (denoted $error_{\mathcal{D}}(h)$) of hypothesis $h$ with respect to target concept $c$ and distribution $\mathcal{D}$ is the probability that $h$ will misclassify an instance drawn at random according to $\mathcal{D}$.

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

# Two Notions of Error

*Training error* of hypothesis $h$ with respect to target concept $c$

- How often $h(x) \neq c(x)$ over training instances D

$$error_{\mathsf{D}}(h) \equiv \Pr_{x \in \mathsf{D}}[c(x) \neq h(x)] \equiv \frac{\sum_{x \in \mathsf{D}} \delta(c(x) \neq h(x))}{|\mathsf{D}|}$$

training examples

*True error* of hypothesis $h$ with respect to $c$

- How often $h(x) \neq c(x)$ over future instances drawn at random from $\mathcal{D}$

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

Probability distribution P(x)

# Two Notions of Error

Can we bound

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

in terms of

$error_{\mathrm{D}}(h)$

??

*Training error* of hypothesis $h$ with respect to target concept $c$

- How often $h(x) \neq c(x)$ over training instances D

$$error_{\mathrm{D}}(h) \equiv \Pr_{x \in \mathrm{D}}[c(x) \neq h(x)] \equiv \frac{\sum_{x \in \mathrm{D}} \delta(c(x) \neq h(x))}{|\mathrm{D}|}$$

training examples

*True error* of hypothesis $h$ with respect to $c$

- How often $h(x) \neq c(x)$ over future instances drawn at random from $\mathcal{D}$

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

Probability distribution P(x)

$$error_{\mathrm{D}}(h) \equiv \Pr_{x \in \mathrm{D}}[c(x) \neq h(x)] \equiv \frac{\sum_{x \in \mathrm{D}} \delta(c(x) \neq h(x))}{|\mathrm{D}|}$$

training examples

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

Probability distribution P(x)

Can we bound

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

in terms of

$error_{\mathrm{D}}(h)$

??

if D was a set of examples drawn from $\mathcal{D}$ and ***independent*** of *h*, then we could use standard statistical confidence intervals to determine that with 95% probability, $error_{\mathcal{D}}(h)$ lies in the interval:

$$error_{\mathrm{D}}(h) \pm 1.96\sqrt{\frac{error_{\mathrm{D}}(h)\,(1 - error_{\mathrm{D}}(h))}{n}}$$

but D is the ***training data*** for *h* ….

# Version Spaces

Target concept is the (usually unknown) boolean fn to be learned

c: X → {0,1}

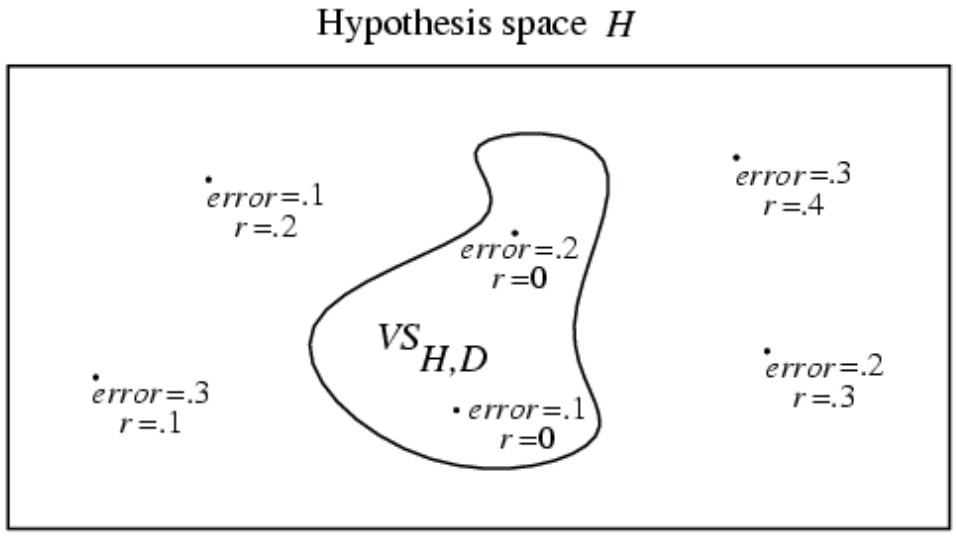
A hypothesis $h$ is **consistent** with a set of training examples $D$ of target concept $c$ if and only if $h(x) = c(x)$ for each training example $\langle x, c(x)\rangle$ in $D$.

$$Consistent(h, D) \equiv (\forall \langle x, c(x)\rangle \in D)\ h(x) = c(x)$$

The **version space**, $VS_{H,D}$, with respect to hypothesis space $H$ and training examples $D$, is the subset of hypotheses from $H$ consistent with all training examples in $D$.

$$VS_{H,D} \equiv \{h \in H | Consistent(h, D)\}$$

# Exhausting the Version Space




($r$ = training error, $error$ = true error)

**Definition:** The version space $VS_{H,D}$ is said to be $\epsilon$-**exhausted** with respect to $c$ and $\mathcal{D}$, if every hypothesis $h$ in $VS_{H,D}$ has true error less than $\epsilon$ with respect to $c$ and $\mathcal{D}$.

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

How many examples will $\epsilon$-exhaust the VS?

---

**Theorem:** [Haussler, 1988].

If the hypothesis space $H$ is finite, and $D$ is a sequence of $m \geq 1$ independent random examples of some target concept $c$, then for any $0 \leq \epsilon \leq 1$, the probability that the version space with respect to $H$ and $D$ is not $\epsilon$-exhausted (with respect to $c$) is less than

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

How many examples will $\epsilon$-exhaust the VS?

---

**Theorem:** [Haussler, 1988].

If the hypothesis space $H$ is finite, and $D$ is a sequence of $m \geq 1$ independent random examples of some target concept $c$, then for any $0 \leq \epsilon \leq 1$, the probability that the version space with respect to $H$ and $D$ is not $\epsilon$-exhausted (with respect to $c$) is less than

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

Interesting! This bounds the probability that <u>any consistent learner</u> will output a hypothesis $h$ with $error(h) \geq \epsilon$

<u>Any(!)</u> learner that outputs a hypothesis consistent with all training examples (i.e., an h contained in $VS_{H,D}$)

hyp space $H$,
target fn $c: X \to \{0,1\}$

Let $h_1 \cdots h_k$ be hyps from $H$ with true error $\geq \epsilon$

Prob that one of these, $h_i$, will be consistent w/ first train examp?

$$\leq (1-\epsilon)$$

Pr that $h_i$ will be consistent w/ $m$ indep train examps?

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

Pr that at least one of $h_1 \cdots h_k$ will be consistent w/ $m$ examps?

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

$k \leq |H|$

$$\leq |H| (1-\epsilon)^m$$

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

if $0 \leq \epsilon \leq 1$
then $(1-\epsilon) \leq e^{-\epsilon}$

# What it means

[Haussler, 1988]: probability that the version space is not ε-exhausted after *m* training examples is at most $|H|e^{-\epsilon m}$

$$\Pr[(\exists h \in H) s.t. (error_{train}(h) = 0) \wedge (error_{true}(h) > \epsilon)] \leq |H|e^{-\epsilon m}$$

Suppose we want this probability to be at most δ

1. How many training examples suffice?

$$m \geq \frac{1}{\epsilon}(\ln |H| + \ln(1/\delta))$$

2. If $error_{train}(h) = 0$ then with probability at least (1-δ):

$$error_{true}(h) \leq \frac{1}{m}(\ln |H| + \ln(1/\delta))$$

# Example: H is Conjunction of Boolean Literals

$$m \geq \frac{1}{\epsilon}(\ln |H| + \ln(1/\delta))$$

Consider classification problem f:X→Y:

- instances: <$X_1\ X_2\ X_3\ X_4$> where each $X_i$ is boolean
- learned hypotheses are rules of the form:
  - IF <$X_1\ X_2\ X_3\ X_4$> = <$0,?,1,?$> , THEN Y=1, ELSE Y=0
  - i.e., rules constrain any subset of the $X_i$

How many training examples *m* suffice to assure that with probability at least 0.9, *any* consistent learner will output a hypothesis with true error at most 0.05?

$|H| = 3^4$

$|H| = 3^n$

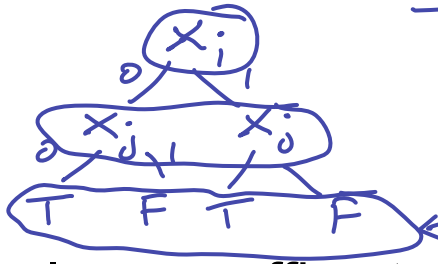
$m \geq \frac{1}{.05}(n \ln 3 + \ln(\frac{1}{.01}))$

# Example: H is Decision Tree with depth=2

$$m \geq \frac{1}{\epsilon}(\ln|H| + \ln(1/\delta))$$

Consider classification problem f:X→Y:

- instances: <$X_1$ ... $X_N$> where each $X_i$ is boolean
- learned hypotheses are decision trees of depth 2, using only two variables




$\binom{N}{2} \cdot 16 = |H| =$

$(N \cdot N-1)/2 \cdot 16$

← any labels

1-δ, δ=0.1

How many training examples *m* suffice to assure that with probability at least 0.9, *any* consistent learner will output a hypothesis with true error at most 0.05?

ε

$$m \geq \frac{1}{0.05}\left(\ln\left(8N^2 - 8N\right) + \ln\left(\frac{1}{0.1}\right)\right)$$

$2 \ln N = \ln N^2$

for N = 2, m ≥ 101

for N = 10, m ≥ 164

# PAC Learning

Consider a class $C$ of possible target concepts defined over a set of instances $X$ of length $n$, and a learner $L$ using hypothesis space $H$.

> *Definition:* $C$ is **PAC-learnable** by $L$ using $H$ if for all $c \in C$, distributions $\mathcal{D}$ over $X$, $\epsilon$ such that $0 < \epsilon < 1/2$, and $\delta$ such that $0 < \delta < 1/2$,
>
> learner $L$ will with probability at least $(1 - \delta)$ output a hypothesis $h \in H$ such that $error_{\mathcal{D}}(h) \leq \epsilon$, in time that is polynomial in $1/\epsilon$, $1/\delta$, $n$ and $size(c)$.

# PAC Learning

Consider a class $C$ of possible target concepts defined over a set of instances $X$ of length $n$, and a learner $L$ using hypothesis space $H$.

*Definition:* $C$ is **PAC-learnable** by $L$ using $H$ if for all $c \in C$, distributions $\mathcal{D}$ over $X$, $\epsilon$ such that $0 < \epsilon < 1/2$, and $\delta$ such that $0 < \delta < 1/2$,

learner $L$ will with probability at least $(1 - \delta)$ output a hypothesis $h \in H$ such that $error_{\mathcal{D}}(h) \leq \epsilon$, in time that is polynomial in $1/\epsilon$, $1/\delta$, $n$ and $size(c)$.

Sufficient condition:

Holds if L requires only a polynomial number of training examples, and processing per example is polynomial

# Agnostic Learning

So far, assumed $c \in H$

Agnostic learning setting: don't assume $c \in H$

- What do we want then?
  - The hypothesis $h$ that makes fewest errors on training data
- What is sample complexity in this case?

note ε here is the difference between the training error and true error

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

derived from Hoeffding bounds:

$$Pr[error_{\mathcal{D}}(h) > error_{D}(h) + \epsilon] \leq e^{-2m\epsilon^2}$$

true error  training error  degree of overfitting

# Additive Hoeffding Bounds – Agnostic Learning

- Given *m* independent coin flips of coin with Pr(heads) = $\theta$
  bound the error in the maximum likelihood estimate $\hat{\theta}$

$$\Pr[\theta > \hat{\theta} + \epsilon] \leq e^{-2m\epsilon^2}$$

- Relevance to agnostic learning: for any *single* hypothesis *h*

$$\Pr[error_{true}(h) > error_{train}(h) + \epsilon] \leq e^{-2m\epsilon^2}$$

- But we must consider all hypotheses in H

$$\Pr[(\exists h \in H) error_{true}(h) > error_{train}(h) + \epsilon] \leq |H| e^{-2m\epsilon^2}$$

- So, with probability at least (1-δ) every h satisfies

$$error_{true}(h) \leq error_{train}(h) + \sqrt{\frac{\ln |H| + \ln \frac{1}{\delta}}{2m}}$$

# General Hoeffding Bounds

- When estimating parameter $\theta$ inside [a,b] from *m* examples

$$P(|\hat{\theta} - E[\hat{\theta}]| > \epsilon) \leq 2e^{\frac{-2m\epsilon^2}{(b-a)^2}}$$

- When estimating a probability $\theta$ is inside [0,1], so

$$P(|\hat{\theta} - E[\hat{\theta}]| > \epsilon) \leq 2e^{-2m\epsilon^2}$$

- And if we're interested in only one-sided error, then

$$P((E[\hat{\theta}] - \hat{\theta}) > \epsilon) \leq e^{-2m\epsilon^2}$$