

Machine Learning 10-701

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

February 24, 2011

Today:

- Computational Learning Theory
- PAC learning theorem
- VC dimension

Recommended reading:

- Mitchell: Ch. 7
- suggested exercises: 7.1, 7.2, 7.7

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

* see Annual Conference on Learning Theory (COLT)

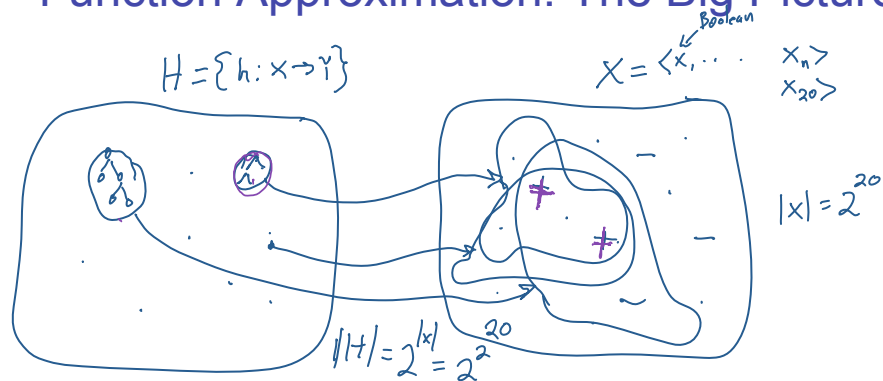
Sample Complexity

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

Target concept is the boolean-valued fn to be learned
 $c: X \rightarrow \{0,1\}$

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 $c(x)$

Function Approximation: The Big Picture



How many labeled examples are needed in order to determine which of the 2^{20} hypotheses is the correct one?

All 2^{20} instances in X must be labeled!

There is no free lunch!

Inductive inference - generalizing beyond the training data is impossible unless we add more assumptions (e.g. priors over H)

Sample Complexity: 3

Given:

- set of instances X
- set of hypotheses H
- set of possible target concepts $C = \{c: X \rightarrow \{0,1\}\}$
- training instances generated by a fixed, unknown probability distribution \mathcal{D} over X $\leftarrow P(x)$

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

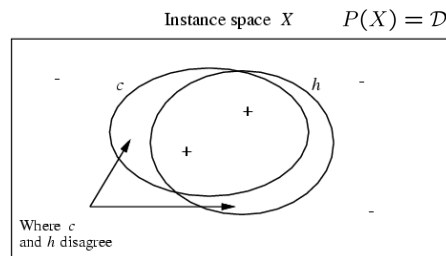
- instances x are drawn from distribution \mathcal{D}
- teacher provides target value $c(x)$ for each

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

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

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}

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

training
examples

Probability
distribution
 $P(x)$

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

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

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}

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

Can we bound

$\text{error}_{\mathcal{D}}(h)$

in terms of

$\text{error}_D(h)$

??

training
examples

Probability
distribution
 $P(x)$

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

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

training examples
 Probability distribution $P(x)$

Can we bound $error_D(h)$ in terms of $error_D(h)$??

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, $error_D(h)$ lies in the interval:

$error_D(h) \pm 1.96 \sqrt{\frac{error_D(h)(1 - error_D(h))}{n}}$

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

Version Spaces

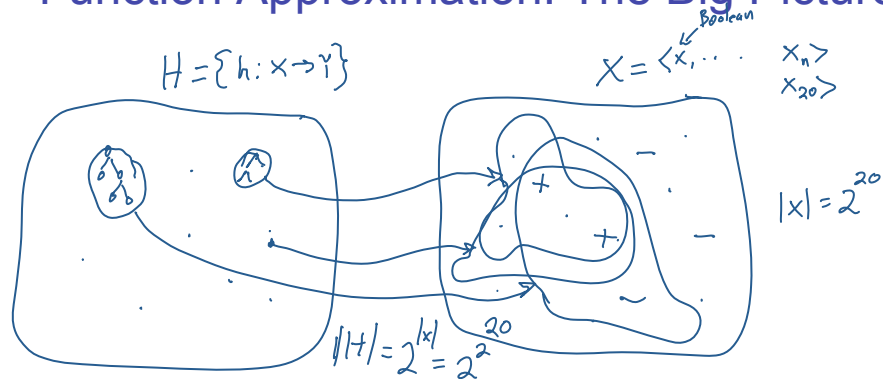
A hypothesis h is **consistent** with a set of training examples D of target concept c if and only if $h(x) = c(x)$ for each training example $\langle x, c(x) \rangle$ in D .

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

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 .

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

Function Approximation: The Big Picture



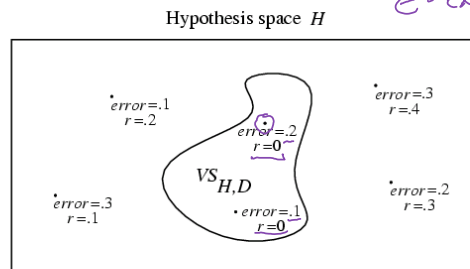
How many labeled examples are needed in order to determine which of the 2^{20} hypotheses is the correct one?

All 2^{20} instances in X must be labeled!

There is no free lunch!

Inductive inference - generalizing beyond the training data is impossible unless we add more assumptions (e.g. priors over H)

Exhausting the Version Space



ϵ -exhausted for $\epsilon > .2$

Definition: 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 true error less than ϵ with respect to c and \mathcal{D} .

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

How many examples will ϵ -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 ϵ -exhausted (with respect to c) is less than

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

How many examples will ϵ -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 ϵ -exhausted (with respect to c) is less than

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

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

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

hyp space H
 instances X
 fn $C: X \rightarrow \{0,1\}$
 m labeled examps
 error tolerance ϵ

let h_1, \dots, h_k be the hyps best with true error $\geq \epsilon$

Prob that h_1 will be consistent with first training example $\leq (1-\epsilon)$
 " h_1 will be cons. w/ m indep drawn examps? $\leq (1-\epsilon)^m$
 " that at least of h_1, \dots, h_k will be consist w/ m ? $\leq k(1-\epsilon)^m$
 $k \leq |H|$ $\leq |H|(1-\epsilon)^m$ if $0 \leq \epsilon \leq 1 - \epsilon$ then $(1-\epsilon) \leq e^{-\epsilon}$
 $\leq |H| e^{-\epsilon m}$

What it means

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

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

↑

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 $(1-\delta)$:

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

Example: H is Conjunction of Boolean Literals

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

Consider classification problem $f: X \rightarrow Y$:

- instances: $X = \langle X_1 X_2 X_3 X_4 \dots X_n \rangle$ where each X_i is boolean
- learned hypotheses are rules of the form:
 - IF $\langle X_1 X_2 X_3 X_4 \dots X_n \rangle = \langle 0, ?, 1, ? \rangle$, 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 0.99, any consistent learner will output a hypothesis with true error at most 0.05?

$$m \geq \frac{1}{.05} \left(\ln |H| + \ln \left(\frac{1}{.01} \right) \right)$$

$n=4 \rightarrow m \geq 180 \leftarrow 3^4 = 81$
 $n=10 \geq 312$
 $n=100 \geq 2290$

$|H| = 3^n$

Example: H is Decision Tree with depth=2

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

Consider classification problem $f: X \rightarrow Y$:

- instances: $X = \langle X_1 \dots X_N \rangle$ where each X_i is boolean
- learned hypotheses are decision trees of depth 2, using only two variables



$$\binom{N}{2} \cdot 16 = \frac{N \cdot (N-1)}{2} \cdot 16$$

How many training examples m suffice to assure that with probability at least 0.99, any consistent learner will output a hypothesis with true error at most 0.05?

$$|H| = 8N^2 - 8N$$

$$m \geq \frac{1}{0.05} \left(\ln (8N^2 - 8N) + \ln \left(\frac{1}{0.01} \right) \right)$$

$N=4 \quad m \geq 184$
 $N=10 \quad m \geq 224$
 $N=100 \quad m \geq 318$

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 , ϵ such that $0 < \epsilon < 1/2$, and δ such that $0 < \delta < 1/2$, learner L will with probability at least $(1 - \delta)$ output a hypothesis $h \in H$ such that $\text{error}_{\mathcal{D}}(h) \leq \epsilon$, in time that is polynomial in $1/\epsilon$, $1/\delta$, n and $\text{size}(c)$.

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 , ϵ such that $0 < \epsilon < 1/2$, and δ such that $0 < \delta < 1/2$, learner L will with probability at least $(1 - \delta)$ output a hypothesis $h \in H$ such that $\text{error}_{\mathcal{D}}(h) \leq \epsilon$, in time that is polynomial in $1/\epsilon$, $1/\delta$, n and $\text{size}(c)$.

Sufficient condition:
Holds if learner 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

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

- What is sample complexity in this case?

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

derived from Hoeffding bounds:

$$\Pr[\text{error}_D(h) > \text{error}_D(h) + \epsilon] \leq e^{-2m\epsilon^2}$$

true error

training error

degree of overfitting

Additive Hoeffding Bounds – Agnostic Learning

- Given m independent coin flips of coin with true $\Pr(\text{heads}) = \theta$ 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

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

- But we must consider all hypotheses in H

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

- So, with probability at least $(1-\delta)$ every h satisfies

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

General Hoeffding Bounds

- When estimating parameter θ 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 θ 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

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

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

Question: If $H = \{h \mid h: X \rightarrow Y\}$ is infinite, what measure of complexity should we use in place of $|H|$?

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

Question: If $H = \{h \mid h: X \rightarrow Y\}$ is infinite, what measure of complexity should we use in place of $|H|$?

Answer: The largest subset of X for which H can guarantee zero training error (regardless of the target function c)