Computational Learning Theory

Read Chapter 7 of Machine Learning
[Suggested exercises: 7.1, 7.2, 7.5, 7.7]

- Computational learning theory
- Setting 1: learner poses queries to teacher
- Setting 2: teacher chooses examples
- Setting 3: randomly generated instances, labeled by teacher
- Probably approximately correct (PAC) learning
- Vapnik-Chervonenkis Dimension

Function Approximation

Given:

• Instance space X:

- e.g. X is set of boolean vectors of length n; x = <0,1,1,0,0,1>

• Hypothesis space H: set of functions h: X → Y

- e.g., H is the set of boolean functions (Y={0,1}) defined by conjunction of constraints on the features of x.

• Training Examples D: sequence of positive and negative examples of an unknown target function c: X →{0,1}

- $<x_1, c(x_1)>, \ldots <x_m, c(x_m)>$

Determine:

• A hypothesis h in H such that h(x)=c(x) for all x in X

Function Approximation

Given:

• Instance space X:

- e.g. X is set of boolean vectors of length n; x = <0,1,1,0,0,1>

• Hypothesis space H: set of functions h: X → Y

- e.g., H is the set of boolean functions (Y={0,1}) defined by conjunctions of constraints on the features of x.

• Training Examples D: sequence of positive and negative examples of an unknown target function c: X →{0,1}

- $<x_1, c(x_1)>, \ldots <x_m, c(x_m)>$

Determine:

What we want

• A ~~hypothesis h in H such that h(x)=c(x) for all x in X~~

• A hypothesis h in H such that h(x)=c(x) for all x in D

What we can observe

Computational Learning Theory

What general laws constrain inductive learning?

We seek theory to relate:

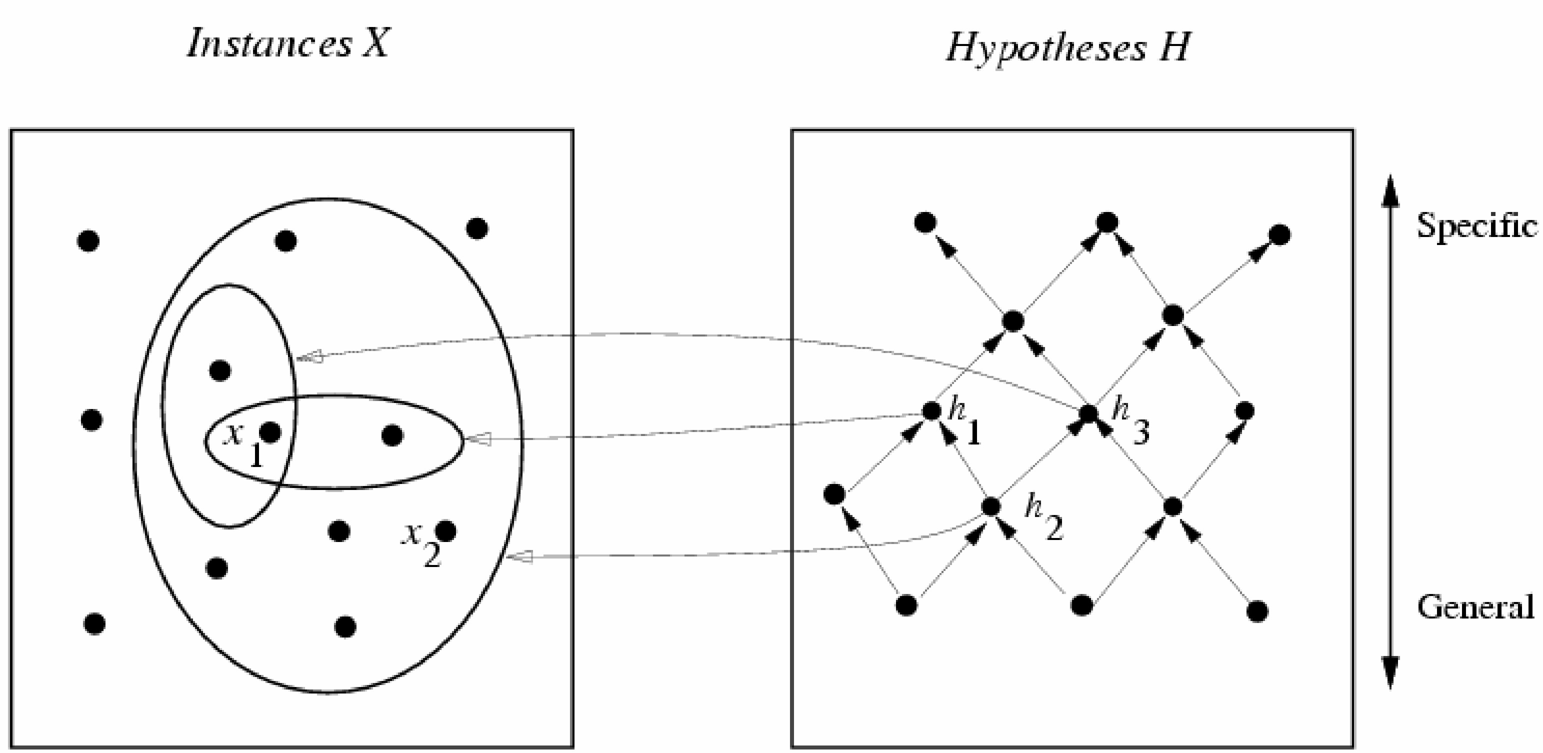
- Probability of successful learning
- Number of training examples
- Complexity of hypothesis space
- Accuracy to which target function is approximated
- Manner in which training examples presented

Sample Complexity

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 $c(x)$

Instances, Hypotheses, and More-General-Than



x_1 = <Sunny, Warm, High, Strong, Cool, Same>
x_2 = <Sunny, Warm, High, Light, Warm, Same>

h_1 = <Sunny, ?, ?, Strong, ?, ?>
h_2 = <Sunny, ?, ?, ?, ?, ?>
h_3 = <Sunny, ?, ?, ?, Cool, ?>

Sample Complexity: 1

Learner proposes instance x, teacher provides $c(x)$ (assume c is in learner's hypothesis space H)

Optimal query strategy: play 20 questions

i.e., minimizes the number of queries needed to converge to the correct hypothesis.

- pick instance x such that half of hypotheses in VS classify x positive, half classify x negative
- When this is possible, need $\lceil \log_2 |H| \rceil$ queries to learn c
- when not possible, need even more

Sample Complexity: 2

Teacher (who knows c) provides training examples (assume c is in learner's hypothesis space H)

Optimal teaching strategy: depends on H used by learner

Consider the case $H =$ conjunctions of up to n boolean literals and their negations

e.g., $(AirTemp = Warm) \wedge (Wind = Strong)$, where $AirTemp, Wind, \ldots$ each have 2 possible values.

Sample Complexity: 2

Teacher (who knows c) provides training examples (assume c is in learner's hypothesis space H)

Optimal teaching strategy: depends on H used by learner

Consider the case $H =$ conjunctions of up to n boolean literals and their negations

e.g., $(AirTemp = Warm) \wedge (Wind = Strong)$, where $AirTemp, Wind, \ldots$ each have 2 possible values.

- if n possible boolean attributes in H, $n+1$ examples suffice
- why?

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

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

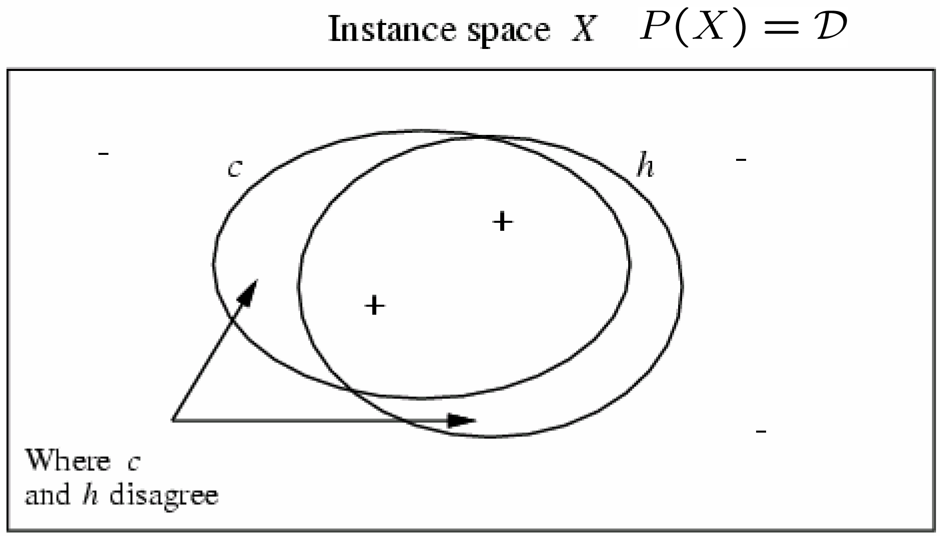
- instances x are drawn from distribution $\mathcal{D}$
- teacher provides target value $c(x)$ for each

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

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

Set of 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

Training error of hypothesis h with respect to target concept c

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

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

Set of 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)

Can we bound $error_{\mathcal{D}}(h)$ in terms of $error_{\mathsf{D}}(h)$??

Version Spaces

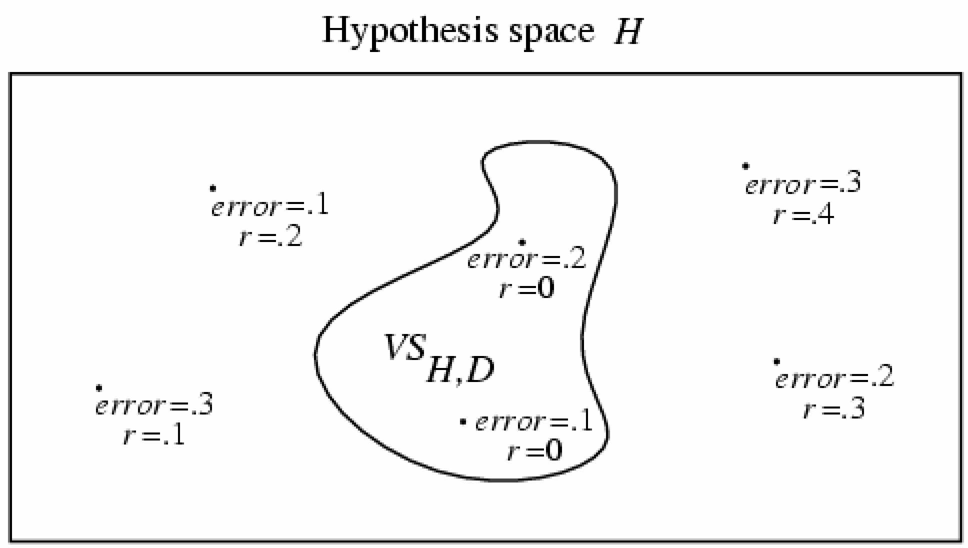
A hypothesis h is **consistent** with a set of training examples D of target concept c if and only if $h(x) = c(x)$ for each training example $\langle x, c(x) \rangle$ in D.

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

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.

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

Exhausting the Version Space



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

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 $\mathcal{D}$.

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

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

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

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 $VS_{H,D}$)

If we want to this probability to be below δ

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

then

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

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):

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

Learning Conjunctions of Boolean Literals

How many examples are sufficient to assure with probability at least $(1-\delta)$ that

every h in $VS_{H,D}$ satisfies $error_{\mathcal{D}}(h) \leq \epsilon$

Use our theorem:

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

Suppose H contains conjunctions of constraints on up to n boolean attributes (i.e., n boolean literals). Then $|H| = 3^n$, and

$$m \geq \frac{1}{\epsilon}(\ln 3^n + \ln(1/\delta))$$

or

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

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 $error_{\mathcal{D}}(h) \leq \epsilon$, 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, 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 $error_{\mathcal{D}}(h) \leq \epsilon$, in <u>time that is polynomial</u> 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?

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

derived from Hoeffding bounds:

$$Pr[\underbrace{error_{\mathcal{D}}(h)}_{\text{true error}} > \underbrace{error_D(h)}_{\text{training error}} + \underbrace{\epsilon}_{\text{degree of overfitting}}] \leq e^{-2m\epsilon^2}$$

Additive Hoeffding Bounds – Agnostic Learning

- Given *m* independent coin flips of coin with Pr(heads) = θ bound the error in the 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 $\theta \in [a,b]$ from *m* examples

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

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

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

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

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