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:

- Instances X:

    - e.g. x = <0,1,1,0,0,1>

- Hypotheses 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. (such as <0,1,?,?,?,1> → 1)

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

    - $\langle x_1, c(x_1)\rangle$, … $\langle x_m, c(x_m)\rangle$

Determine:

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

Function Approximation

Given:

- Instances X:

  - e.g. x = <0,1,1,0,0,1>

- Hypotheses 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. (such as <0,1,?,?,?,1> → 1)

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

What we want

What we can observe

Function Approximation

Given:

- Instances X:

    - e.g. x = <0,1,1,0,0,1>

- Hypotheses H: set of functions h: X →{0,1}

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

- 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
- A hypothesis h in H that minimizes $\text{lossfunction}_i(h,D)$

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 $c(x)$

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 $c \in C$

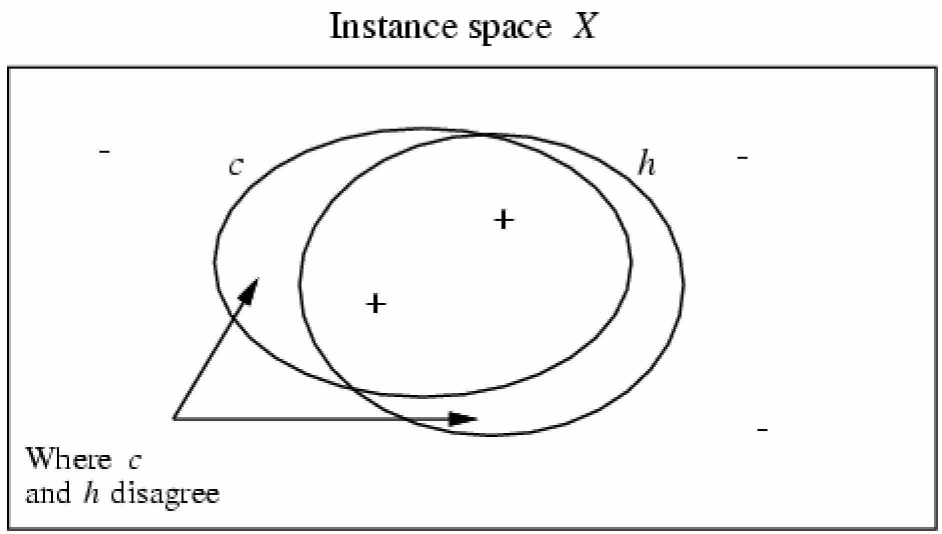
- instances x are drawn from distribution $\mathcal{D}$
- teacher provides target value $c(x)$ for each

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

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

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

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

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

Version Spaces

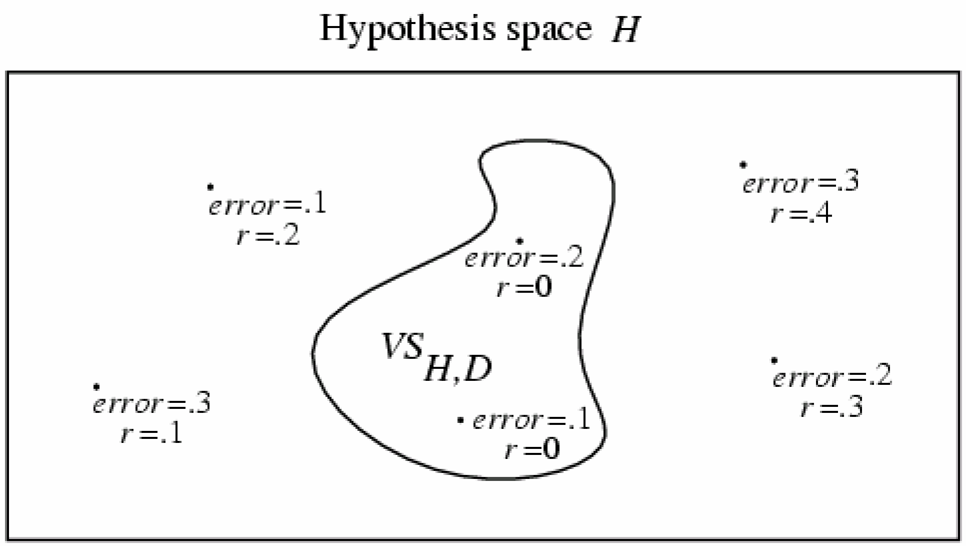
A hypothesis h is **consistent** with a set of training examples D of target concept c if and only if $h(x) = c(x)$ for each training example $\langle x, c(x) \rangle$ in D.

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

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

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

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 <u>time that is polynomial</u> 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 $size(c)$.

Holds if L requires only 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:

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

true error training error degree of overfitting

General Hoeffding Bound

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

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

What if H is not finite?

- Can’t use our result for finite H

- Need some other measure of complexity for H
  - Vapnik-Chervonenkis (VC) dimension!

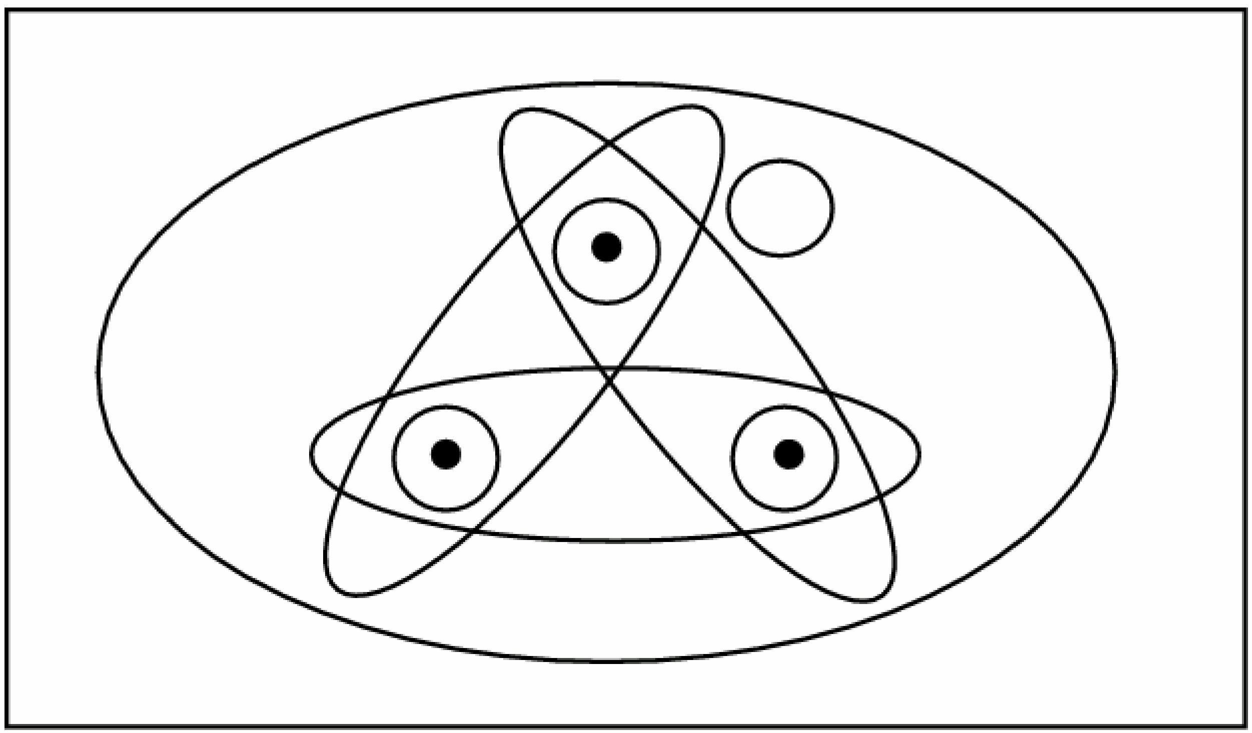
Shattering a Set of Instances

Definition: a **dichotomy** of a set S is a partition of S into two disjoint subsets.

Definition: a set of instances S is **shattered** by hypothesis space H if and only if for every dichotomy of S there exists some hypothesis in H consistent with this dichotomy.

Three Instances Shattered

Instance space X



The Vapnik-Chervonenkis Dimension

Definition: The **Vapnik-Chervonenkis dimension**, $VC(H)$, of hypothesis space H defined over instance space X is the size of the largest finite subset of X shattered by H. If arbitrarily large finite sets of X can be shattered by H, then $VC(H) \equiv \infty$.

Sample Complexity from VC Dimension

How many randomly drawn examples suffice to ϵ-exhaust $VS_{H,D}$ with probability at least $(1-\delta)$?

$$m \geq \frac{1}{\epsilon}(4\log_2(2/\delta) + 8VC(H)\log_2(13/\epsilon))$$

VC dimension: examples

Consider $X = \Re$, want to learn $c: X \rightarrow \{0,1\}$

What is VC dimension of

- H1 $= \{ (x>a \rightarrow y=1) \mid a \in \Re\}$

- H2 $= \{ (x>a \rightarrow y=1) \mid a \in \Re\} + \{ (x<a \rightarrow y=1) \mid a \in \Re\}$

VC dimension: examples

Consider $X = \Re$, want to learn $c: X \rightarrow \{0,1\}$

What is VC dimension of

- H1 $= \{ (x>a \rightarrow y=1) \mid a \in \Re \}$
  - $VC(H1)=1$

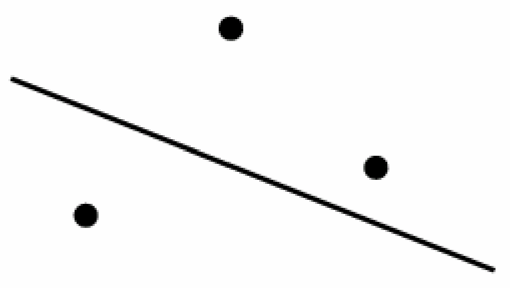
- H2 = $\{ (x>a \rightarrow y=1) \mid a \in \Re \} + \{ (x<a \rightarrow y=1) \mid a \in \Re \}$
  - $VC(H2)=2$

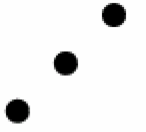
VC dimension: examples

Consider X = $\Re^2$, want to learn c:X→{0,1}

What is VC dimension of

- H1 = $\{ (w \cdot x+b)>0 \rightarrow y=1) \mid w \in \Re^2, b \in \Re\}$





VC dimension: examples

Consider X = $\Re^2$, want to learn c:X→{0,1}

What is VC dimension of

- H1 = { (w·x+b)>0 → y=1) | w $\in \Re^2$, b $\in \Re$}
  - VC(H1)=3
  - For linear separating hyperplanes in n dimensions, VC(H)=n+1

Key Ideas from this lecture

- Sample complexity varies with the learning setting
  - Learner actively queries trainer
  - Examples provided at random
  - …

- Within the PAC learning setting, we can bound the probability that learner will output hypothesis with given error
  - In terms of complexity of H, number of examples
  - For ANY consistent learner (case where $c \in H$)
  - For ANY "best fit" hypothesis (agnostic learning, where $c \notin H$)

- VC dimension is useful measure of complexity of H

- More details: see annual Conference on Learning Theory
  - http://www.learningtheory.org/colt2004/