Learning Theory Introduction

Statistical Learning Theory

Chris Mesterharm

Department of Computer Science
Rutgers, the State University of New Jersey

Feb 1, 2012

Problem Definition

Background

Statistical Learning Theory is based on classical statistics.

Notation

Assume I have a fixed but unknown distribution $P(X, Y)$ where X is the set of examples and Y is the set of labels. For notational convenience, let $Z = X \times Y$ where $z = (x, y) \in Z$ is an instance of the learning problem.

Problem

Input A set of independent samples from $P(X, Y)$. (Training data.)

Output A function $h : X \to Y$ that has low error or $P(X, Y)$.

Problem Definition Cont

Goal

Find a function $h(x)$ that has a small error on $P(X, Y)$.

Error

Many possible definitions of error. For this lecture,

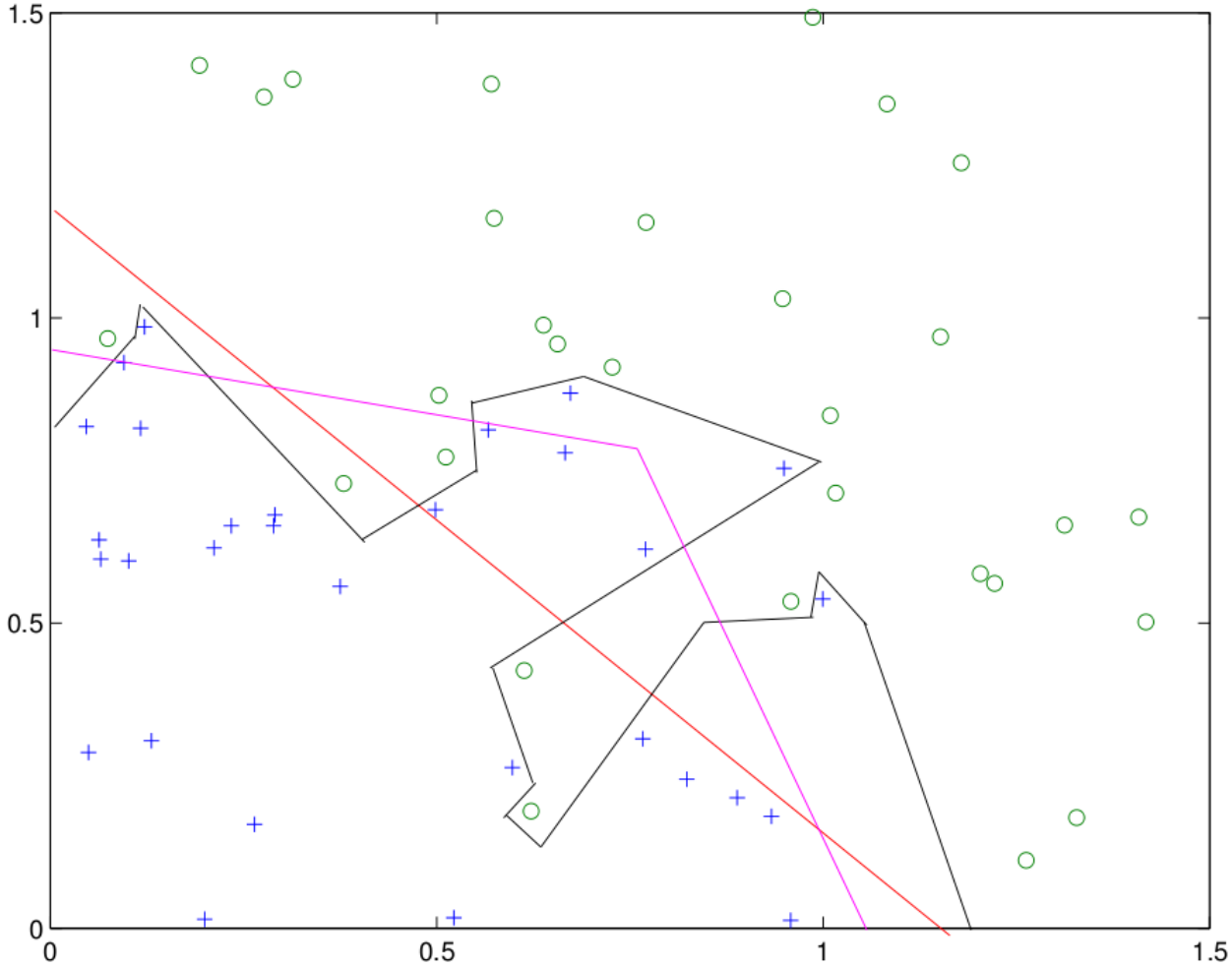
$$\text{Error}(h(x)) = \text{E}_P[|h(x) - y|]$$

where $Y = \{0, 1\}$. This measures the probability of a mistake.

Example

X is the set of all possible webpages.
$Y = \{\text{not a homepage, a homepage}\}$.

2-D Example



Fair Use. (Introduction to Statistical Learning Theory.)

$P(X, Y)$

Possible assumptions on $P(X)$

- Fixed distribution. Example: Uniform.
- Parametrized distribution. Example: Gaussian.
- Arbitrary distribution.

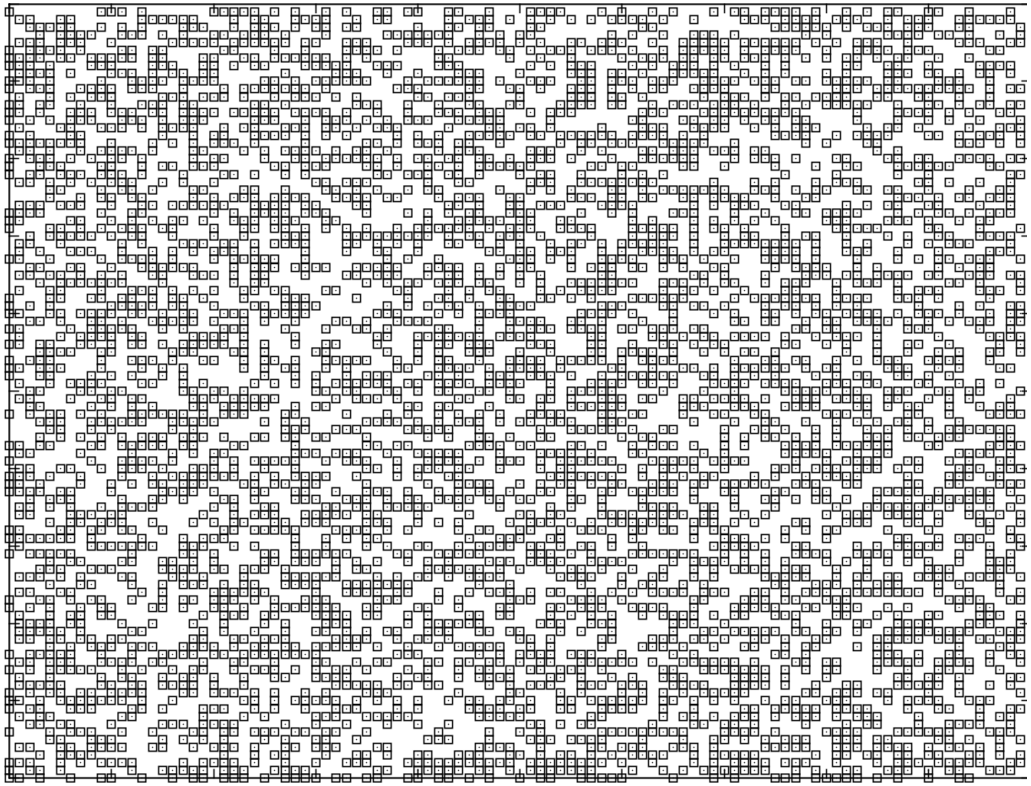
Possible assumptions on $P(Y|X)$

- There exist a function $f \in F$ that determines the labels.
- The exists a function $f \in F$ that determines the labels but the label is modified by a noise function.
- Arbitrary distribution.

We focus on arbitrary $P(X, Y)$.

Arbitrary $P(X, Y)$

How can we learn with arbitrary $P(X, Y)$?



We redefine the problem.

Find Best Hypothesis

Statistical Learning Theory

Given a set of concepts H, find a concept in H that gives close to the lowest error when sampling from $P(X, Y)$.

PAC

Popular alternative model we discuss is PAC. In the PAC model, we assume that the labels are determined by a function $f \in F$ and that $F \subset H$. In the PAC model, the best hypothesis has zero error.

We focus on more general statistical model.

$|H| = 1$

Measure Error

The first step is to estimate the error of a hypothesis on $P(X, Y)$. This is useful in many contexts.

- Let $Z_1, Z_2, \ldots, Z_m$ be m samples from $P(X, Y)$.
- Let $L_i(h) = |h(X_i) - Y_i|$
- Let $\bar{L}(h) = \frac{1}{m} \sum_{i=1}^{m} L_i(h)$.
- $E[\bar{L}(h)] = E[|h(X) - Y|]$ which is the error of h.

How small is $|\bar{L}(h) - E[\bar{L}(h)]|$. Use classical statistics to give a confidence interval. Each test instance L_i is a Bernoulli trial with chance error(h) of success. This gives a binomial distribution. (Give R example.)

Chernoff Bound

Let $L_1, L_2, \ldots, L_m$ be iid random variables such that $L_i \in [0, 1]$.

Theorem (*Bernstein*)

$$\Pr(|\bar{L} - E[\bar{L}]| > \epsilon) \leq 2\exp(-2m\epsilon^2)$$

It is useful to express this bound in different forms. Let δ be the desired maximum probability that the estimate is bad. As long as $2\exp(-2m\epsilon^2) \leq \delta$ then the Chernoff bound fails with at most δ probability. We can rearrange this to give different bounds.

Epsilon Bound and Sample Bound

$$\epsilon \geq \sqrt{\frac{\log(2/\delta)}{2m}} \qquad m \geq \frac{\log(2/\delta)}{2\epsilon^2}$$

The use of ϵ and δ is where PAC gets its name. PAC stands for Probably Approximately Correct.

Finite H

- Let $H = \{h_1, h_2, \ldots, h_k\}$.
- Recall $L_i(h) = |h(X_i) - Y_i|$.
- Recall $\bar{L}(h) = \frac{1}{m}\sum_{i=1}^{m} L_i(h)$.
- Let C_j be the event that $|E[\bar{L}(h_j)] - \bar{L}(h_j)| > \epsilon$.

Union Bound

$$\Pr(C_1 \cup C_2 \cup \cdots \cup C_k) \leq \Pr(C_1) + \Pr(C_2) + \cdots \Pr(C_k).$$

$$\Pr(\exists h \in H \text{ such that } |E[\bar{L}(h)] - \bar{L}(h)| > \epsilon) \leq 2k \exp(-2m\epsilon^2).$$

Main Result

$$m \geq \frac{\log(2k/\delta)}{2\epsilon^2}.$$

Key Assumptions

Key Assumptions

- Independent samples.
- Identical distribution.
- Independent samples.

Independent samples are very powerful to get central limit like results. One can weaken these assumptions and still get strong uniform convergence results, for example martingale sequences.

Empirical Risk Minimization

ERM Algorithm

Find the hypothesis in H with the smallest error on the m training data instances. Call this hypothesis h^b.

Let h^* be the hypothesis with the smallest error on $P(X, Y)$.

Theorem

If the error estimate on every $h \in H$ is within ϵ of the true error of h then

$$E[\bar{L}(h^*)] \leq E[\bar{L}(h^b)] \leq E[\bar{L}(h^*)] + 2\epsilon.$$

We can guarantee that we get ϵ close with $1 - \delta$ probability using previous results.

Example

How many instances of training data do you need to get with .01 accuracy of the best hypothesis in H where $|H| = 10000$ using the ERM algorithm.

Example

How many instances of training data do you need to get with .01 accuracy of the best hypothesis in H where $|H| = 10000$ using the ERM algorithm. Set the confidence interval to 99%.

Example

How many instances of training data do you need to get with .01 accuracy of the best hypothesis in H where $|H| = 10000$ using the ERM algorithm. Set the confidence interval to 99%.

Solution

- $\delta = .01$

Example

How many instances of training data do you need to get with .01 accuracy of the best hypothesis in H where $|H| = 10000$ using the ERM algorithm. Set the confidence interval to 99%.

Solution

- $\delta = .01$
- $k = 10000$

Example

How many instances of training data do you need to get with .01 accuracy of the best hypothesis in H where $|H| = 10000$ using the ERM algorithm. Set the confidence interval to 99%.

Solution

- $\delta = .01$
- $k = 10000$
- $\epsilon = .005$

Example

How many instances of training data do you need to get with .01 accuracy of the best hypothesis in H where $|H| = 10000$ using the ERM algorithm. Set the confidence interval to 99%.

Solution

- $\delta = .01$
- $k = 10000$
- $\epsilon = .005$
- $$m \geq \frac{\log(20000/.01)}{2(.005)^2} \geq 290174$$

Observations

- Chernoff bound does not take into account variance of the the random variables. For example, if the variance of each variable is zero then the sum will have perfect convergence. The variance of a Bernoulli variable is $p(1-p)$. Therefore the sum of random variables should have better convergence when the error of the hypothesis is small.
- The PAC setting has no noise and is realizable ($F \subset H$). This means that h^* has zero error. This gives an improved bound of $m \geq \frac{\log(|H|/\delta)}{2\epsilon}$.

Observations Cont.

- The finite H bound has a logarithmic dependence on the number of hypotheses. This is not as good as it seems. If a learning problem has n binary attributes then max $|H| = 2^{2^n}$. This gives a bound much greater than 2^n. In other words, we need to see every instance many times.
- For these results to be useful, we need to select a set H that performs well with real data, but is much smaller than $2^{|X|}$.
- Many algorithms have parameters that greatly effect performance. ERM can be used to justify parameter selection with a validation set.

Computers

Every computer algorithm uses a finite set of hypotheses.

Example

Your algorithm represents a hypothesis with 100 floats. Derive a bound on the error of the hypothesis returned by the algorithm. Notice that $|H| = 2^{32(100)}$ since a float is represented with 32 bits.

Semi-parametric Learning

You can allow the size of the hypotheses considered by the algorithm to grow as a function of m and still get uniform convergence bounds. (Occam Algorithms)

Infinite H

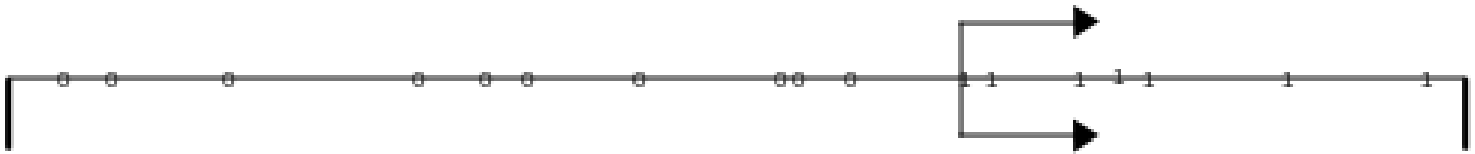
Example

- Let $T = \{[x \geq a] | a \in [0, 1]\}$.
- PAC model where $F = T$.
- With probability $1 - \delta$ find a $h \in T$ such that $error(h) \leq \epsilon$.

Infinite H

Example

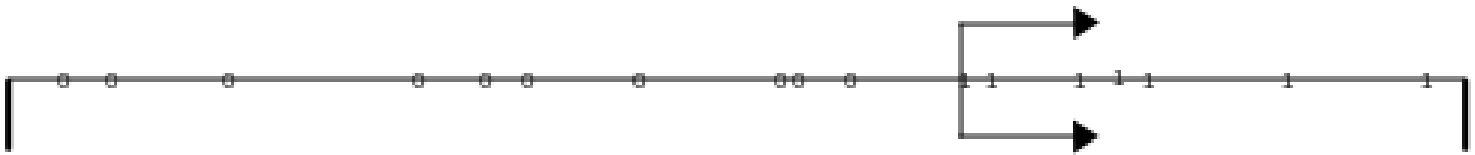
- Let $T = \{[x \geq a] | a \in [0, 1]\}$.
- PAC model where $F = T$.
- With probability $1 - \delta$ find a $h \in T$ such that $error(h) \leq \epsilon$.



Infinite H

Example

- Let $T = \{[x \geq a] | a \in [0, 1]\}$.
- PAC model where $F = T$.
- With probability $1 - \delta$ find a $h \in T$ such that $error(h) \leq \epsilon$.



- Algorithm predicts 1 for examples greater or equal to smallest y=1 training instance. This is shown with arrows in picture.
- True hypothesis in gap between biggest y=0 and smallest y=1.
- The probability of an error is at most the probability that an example lands in this gap. (Real error gap is the gap between the true hypothesis and the algorithm hypothesis.)
- If probability of an error is ϵ or greater then the probability that all samples miss the real error gap is $(1 - \epsilon)^m$. Set $(1 - \epsilon)^m \leq \delta$ to derive bounds for a given δ.
- Use the fact that for $\epsilon \in [0, 1/2]$ that $-2\epsilon \leq \lg(1 - \epsilon)$ to show $m \geq \frac{\lg(1/\delta)}{2\epsilon}$ for $\epsilon \in [0, 1/2]$.

More Infinite H

Example

- Let $U = \{u | u : N \rightarrow \{0, 1\}\}$.
- We can represent U using the set $[0, 2]$.
  - Let $a \in [0, 2]$ represent a hypothesis of U.
  - Let $u_a(n)$ be equal to bit n of a when using a binary representation for a. (We can use first bit to solve technical problem caused by certain real numbers having two binary representations.)
- ERM does not work for U.

Example

ERM does not work for $S = \{[\sin(ax) \geq 0] | a \in R\}$ when one uses a carefully selected set of examples.

VC Dimension

Growth Function

- Given a set of examples $X_1, X_2, \ldots, X_m$ let $\Phi(X_1, \ldots, X_m)$ be the number of unique labelings.
- Let $S_H(m)$ be the maximum number of labelings that can be created on m instances. This is called the growth function.

Example

- $S_T(1) = 2$, $S_T(2) = 3$, $S_T(3) = 4$, $S_T(m) = m + 1$.
- $S_U(1) = 2$, $S_U(2) = 4$, $S_U(3) = 8$, $S_U(m) = 2^m$.

We say a set of examples $X_1, X_2, \ldots, X_m$ is shattered if $\Phi(X_1, \ldots, X_m) = 2^m$.

VC Dimension Cont

VC Dimension

Define $\text{VC}(H)$ as the largest m such that $S_H(m) = 2^m$. In other words, the largest set of examples that can be shattered.

Example

- $\text{VC}(T) = 1$.
- $\text{VC}(U) = \infty$.

Example

- Find the VC dimension of 2 dimensional half-spaces.

VC Dimension Cont

VC Dimension

Define VC(H) as the largest m such that $S_H(m) = 2^m$. In other words, the largest set of examples that can be shattered.

Example

- VC(T) $= 1$.
- VC(U) $= \infty$.

Example

- Find the VC dimension of 2 dimensional half-spaces. 3.
- Find the VC dimension of n dimensional half-spaces.

VC Dimension Cont

VC Dimension

Define VC(H) as the largest m such that $S_H(m) = 2^m$. In other words, the largest set of examples that can be shattered.

Example

- VC(T) $= 1$.
- VC(U) $= \infty$.

Example

- Find the VC dimension of 2 dimensional half-spaces. 3.
- Find the VC dimension of n dimensional half-spaces. $n + 1$ (Radon's Lemma)

VC Dimension Cont

A common notation is to use d as shorthand for the VC dimension.

Sauer's Lemma

$$S_H(m) \leq \sum_{i=0}^{d} \binom{m}{i}.$$

Proves how VC dimension limits the possible example labelings.

Polynomial Upper Bound

$$S_H(m) \leq \sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{em}{d}\right)^d$$

As m increase the number of possible labelings grows only polynomially in m.

VC Dimension Cont

Example

- Let $H_0 = \{h | h(x) = 0\}$.
- Let $H_1 = \{h | h(x) = 1$ for one x value $\}$.
- Let $H_2 = \{h | h(x) = 1$ for two x values $\}$.
- $\vdots$
- Let $H_d = \{h | h(x) = 1$ for d x values $\}$.

Let $H = H_0 \cup \cdots \cup H_d$.

- $\text{VC}(H) = d$.
- $S_H(m) = \sum_{i=0}^{d} \binom{m}{i}$.

If I add any function to H then the VC dimension must increase.

Error Bound

Main Result

$$\Pr(\exists h \in H \; |E[\bar{L}(h)] - \bar{L}(h)| > \epsilon) \leq 2\exp(\log(S_H(2m)) - m\epsilon^2/8).$$

We can use δ notation and simplify.

$$m \geq 8\frac{d\log(2em/d) + \log(2/\delta)}{\epsilon^2}.$$

Notice that m is on both sides of the inequality.

$$\epsilon \geq \sqrt{8\frac{d\log(2em/d) + \log(2/\delta)}{m}}.$$

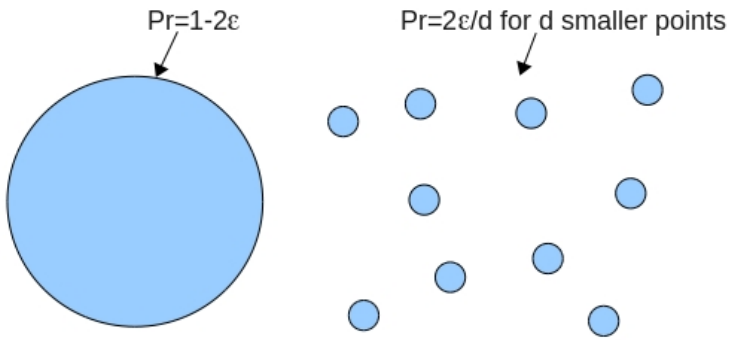
Best current bound using Rademacher averages. There exists a K

$$\epsilon \geq K\left(\sqrt{\frac{d + \log(1/\delta)}{m}}\right).$$

Lower Bounds

- PAC learning must take at least $\frac{d}{4\epsilon}$ samples (with δ close to $1/2$) in the worst case.

  Pick d smaller points so that they are shattered by H. Must see at least $d/2$ smaller points to get within ϵ of perfect. Roughly, every $\frac{1}{2\epsilon}$ samples will be from the smaller points. The remaining samples will be from the high probability point. The label of the high probability point doesn't matter (we learn it pretty quickly.)



- General statistical learning must take $\Omega\left(Kd\eta/\epsilon^2\right)$ samples in the worst case where η is the error rate of the best hypothesis in H.

Structural Risk Minimization

ERM gives a single confidence interval for all $h \in H$. It can be useful to give different confidence intervals to different hypotheses.

Example: Relevant Attribute Half-Spaces

- Order the attributes so that initial attributes are more relevant for learning your concept.
- Let $H_k = \{[\sum_{i=1}^{k} w_i x_i \geq b] | w_1, \ldots, w_k, b \in R\}$.
- Give δ/n confidence to $H_1, \ldots, H_n$ and apply union bound.
- Return hypothesis that has best upper bound on error.

Regularization

- Can extend this in a continuous way using regularization.
- No-noise SVM proof is based on regularization.

Computational Issues

Time

Just because a problem needs only a polynomial number of samples does not mean one can solve ERM in polynomial amount of time.

Example: Half-Spaces

It is NP hard to find a half-space that has error much less than 50% with arbitrary instances even if those instances can almost be perfectly separated by a half-space.

Example: Conjunctions

PAC agnostic model ($H \subset F$) cannot efficiently learn conjunctions unless RP=NP.

Example: Discrete Cube Root Problem

PAC based discrete cube root learning problem can only be solved in polynomial time if we can efficiently compute discrete cube roots.

Other Popular Learning Theory Models

- Bayesian Learning
  - Each $h \in H$ has a prior probability.
  - Has complex modeling requirements which can be good and bad.
- On-line Learning.
  - No statistical assumptions. An adversary can generate instances including noisy instances.
  - Requires label information for most examples. Ideal for problems that predict the future.