Statistical Learning Theory and PAC-Learning

CS678 Advanced Topics in Machine Learning
Thorsten Joachims
Spring 2003

Outline:
• What is the true (prediction) error of classification rule \( h \)?
• How to bound the true error given the training error?
• Finite hypothesis space and zero training error
• Finite hypothesis space and non-zero training error
• Infinite hypothesis spaces: VC-Dimension and Growth Function

Learning Classifiers from Examples (Scenario)

Scenario:
• Generator: Generates descriptions \( \tilde{x} \) according to distribution \( P(\tilde{x}) \).
• Teacher: Assigns a value \( y \) to each description \( \tilde{x} \) based on distribution \( P(y|\tilde{x}) \).

Given:
• Training examples \( (\tilde{x}_1, y_1), ..., (\tilde{x}_n, y_n) \sim P(\tilde{x}, y) \quad \tilde{x}_i \in \mathbb{R}^N, y_i \in \{1,-1\} \)
• Set \( H \) of classification rules \( h \) (hypotheses) that map descriptions \( \tilde{x} \) to values \( y \) \( (h(\tilde{x}) \rightarrow y) \).

Goal of Learner:
• Classification rule \( h \) from \( H \) that classifies new examples (again from \( P(\tilde{x}, y) \)) with low error rate!

\[
P(h(\tilde{x}) \neq y) = \int P(h(\tilde{x}) \neq y) dP(\tilde{x}, y) = \text{Err}_P(h)
\]

Principle: Empirical Risk Minimization (ERM)

Learning Principle:
Find the decision rule \( h^* \in H \) for which the training error is minimal:

\[
h^* = \arg\min_{h \in H} \{ \text{Err}_S(h) \}
\]

Training Error:

\[
\text{Err}_S(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(y_i \neq h(\tilde{x}_i))
\]

\[\Rightarrow\] Number of misclassifications on training examples.

Central Problem of Statistical Learning Theory:
When does a low training error lead to a low generalization error?
Sources of Variation

Learning Task:
- Generator: Generates descriptions \( \hat{x} \) according to distribution \( P(\hat{x}) \).
- Teacher: Assigns a value \( y \) to each description \( \hat{x} \) based on \( P(y|\hat{x}) \).

=> Learning Task: \( P(\hat{x}, y) = P(y|\hat{x})P(\hat{x}) \)

Process:
- Select task \( P(\hat{x}, y) \)
- Training sample \( S \) (depends on \( P(\hat{x}, y) \))
- Train learning algorithm \( A \) (e.g., randomized search)
- Test sample \( V \) (depends on \( P(\hat{x}, y) \))
- Apply classification rule \( h \) (e.g., randomized prediction)

What is the true error of classification rule \( h \)?

Includes variation from different test sets.

Problem Setting:
- given rule \( h \)
- given (independent) test sample \( S = (\hat{x}_1, y_1), ..., (\hat{x}_k, y_k) \) of size \( k \)

\( \hat{P}(h(\hat{x}) \neq y) dP(x,y) = Err_P(h) \)

Approach: measure error of \( h \) on test set

\[ Err_V(h) = \frac{1}{k} \sum_{i=1}^{k} \Delta(y_i \neq h(\hat{x}_i)) \]

Binomial Distribution

The probability of observing \( x \) heads in a sample of \( n \) independent coin tosses, when the probability of heads is \( p \) in each toss, is

\[ P(X = x|p, n) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r} \]

Confidence interval:

Given \( x \) observed heads, with at least 95% confidence the true value of \( p \) fulfills

\[ P(X \geq x|p, n) \geq 0.025 \quad \text{and} \quad P(X \leq x|p, n) \geq 0.025 \]

Cross-Validation Estimation

Given:
- training set \( S \) of size \( n \)

Method:
- partition \( S \) into \( m \) subsets of equal size
- for \( i \) from 1 to \( m \)
  - train learner on all subsets except the \( i^{th} \)
  - test learner on \( i^{th} \) subset
  - record error rates on test set

=> Result: average over recorded error rates

Bias of estimate: see leave-one-out

Warning: Test sets are independent, but not the training sets!

=> no strictly valid hypothesis test is known for general learning algorithms (see [Dietterich/97])
Psychic Game

- I guess a 4 bit code
- You all guess a 4 bit code

=> The student who guesses my code clearly has telepathic abilities - right!?

How can You Convince Me of Your Psychic Abilities?

Setting:
- \( n \) bits
- \(|H|\) players

Question: For which \( n \) and \(|H|\) is prediction of zero-error player significantly different from random \((\rho = 0.5)\) with probability \(1 - \delta\)?

=> Hypothesis test for

\[
P(h_1\text{correct} \lor \ldots \lor h_{|H|}\text{correct}, \text{allnonpsychic}) < \delta
\]

or

\[
P(\exists h \in H; \text{Err}_p(h) = 0, \forall h \in H; \text{Err}_p(h) = 0.5) < \delta
\]

PAC Learning

Definition:
- \( C \) = class of concepts \( c: X \rightarrow \{-1, 1\} \) (functions to be learned)
- \( H \) = class of hypotheses \( h: X \rightarrow \{-1, 1\} \) (functions used by learner \( A \))
- \( S \) = training set (of size \( n \))
- \( \epsilon \) = desired error rate of learned hypothesis
- \( \delta \) = probability, with which the learner \( A \) is allowed to fail

\( C \) is PAC-learnable by Algorithm \( A \) using \( H \) and \( n \) examples, if

\[
P(\text{Err}(h_{A(S)}) \leq \epsilon) \geq (1 - \delta)
\]

for all \( c \in C, \epsilon, \delta, \) and \( P(X) \) so that \( A \) runs in polynomial time dependent on \( \epsilon, \delta \), the size of the training examples and the size of the concepts.

=> only polynomially many training examples allowed.

Case: Finite \( H \), Zero Error

- The hypothesis space \( H \) is finite
- There is always some \( h \) with zero training error (\( A \) returns one such \( h \))
- Probability that a (single) \( h \) with \( \text{Err}_p(h) \geq \epsilon \) has training error of zero

\[
(1 - \epsilon)^n
\]

- Probability that there exists \( h \) in \( H \) with \( \text{Err}_p(h) \geq \epsilon \) that has training error of zero

\[
P(\exists h \in H; \text{Err}_p(h) = 0, \text{Err}_p(h) > \epsilon) \leq |H|(1 - \epsilon)^n \leq |H|e^{-\epsilon n}
\]
Case: Finite $H$, Non-Zero Error

Goal:

$$P(\|\text{Err}_S(h_{A(S)}) - \text{Err}_D(h_{A(S)})\| \leq \epsilon) \geq (1 - \delta)$$

<=

$$P(\sup_H \|\text{Err}_S(h) - \text{Err}_D(h)\| \leq \epsilon) \geq (1 - \delta)$$

• Probability that for a fixed $h$, training error and test error differ by more than $\epsilon$ (Hoeffding / Chernoff Bound)

$$P_{\frac{2^\frac{\delta}{\epsilon^2}}{\epsilon^2}} (\sum_{i=1}^n x_i - p) > \frac{\delta}{\epsilon^2} \leq 2e^{-2n\epsilon^2}$$

• Probability over all $h$ in $H$: union bound => multiply by $|H|$

Case: Infinite $H$

• union bound does no longer work.
• maybe not all hypotheses are really different?!

How Many Dichotomies for Fixed Sample?

• Sample $S$ of size $n$
• Hypothesis class $H$

$$\Pi_{\phi}(S) = \{(h(x_1), h(x_2), ..., h(x_n)) : h \in H\}$$

Definition: $H$ shatters $S$, if $|\Pi_{\phi}(S)| = 2^n$ (i.e. hypotheses from $H$ can split $S$ in all possible ways).

Vapnik/Chervonenkis Dimension

Definition: The VC-dimension of $H$ is equal to the maximal number $d$ of examples that can be split into two sets in all $2^d$ ways using functions from $H$ (shattering).

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>...</th>
<th>$x_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_1$</td>
<td>$+$</td>
<td>$+$</td>
<td>...</td>
<td>$+$</td>
</tr>
<tr>
<td>$h_2$</td>
<td>$-$</td>
<td>$+$</td>
<td>...</td>
<td>$+$</td>
</tr>
<tr>
<td>$h_3$</td>
<td>$+$</td>
<td>$-$</td>
<td>...</td>
<td>$+$</td>
</tr>
<tr>
<td>$h_4$</td>
<td>$-$</td>
<td>$-$</td>
<td>...</td>
<td>$+$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$h_n$</td>
<td>$+$</td>
<td>$-$</td>
<td>...</td>
<td>$-$</td>
</tr>
</tbody>
</table>

Growth function $\Phi_{\phi}(S)$: For all $S$

$$|\Pi_{\phi}(S)| \leq \Phi_{\text{VCdim}(H)}(n) \leq \frac{2^\text{VCdim}(H)}{\text{en}^\text{VCdim}(H)}$$
**Linear Classifiers**

Rules of the Form: weight vector $\vec{w}$, threshold $b$

$$h(x) = \text{sign}\left[\sum_{i=1}^{N} \vec{w}_i \vec{x}_i + b\right] = \begin{cases} 1 & \text{if } \sum_{i=1}^{N} \vec{w}_i \vec{x}_i + b > 0 \\ 0 & \text{else} \end{cases}$$

Geometric Interpretation (Hyperplane):

<table>
<thead>
<tr>
<th></th>
<th>+ + + + +</th>
<th>w</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+ + +</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- - - - -</td>
<td>b</td>
</tr>
</tbody>
</table>

**VC-Dimension of Hyperplanes in $\Re^2$**

- Three points in $\Re^2$ can be shattered with hyperplanes.

- Four points cannot be shattered.

$\Rightarrow$ Hyperplanes in $\Re^2 \Rightarrow VCdim=3$

General: Hyperplanes in $\Re^N \Rightarrow VCdim=N+1$

**Error Bound**

**Question:** After $n$ training examples, how close is the training error to the true error?

With probability $\eta$ it holds for all $h \in H$:

$$Err_P(h) - Err_S(h) \leq \Phi(d, n, \eta)$$

$$\Phi(d, n) = \frac{d\sqrt{2n^2} + \frac{5}{3} \ln \frac{n}{4}}{n}$$

- $n$ number of training examples
- $d$ VC-dimension of hypothesis space $H$

$$Err_P(h) \leq Err_S(h) + \Phi(d, n, \eta)$$

**SVM Motivation: Structural Risk Minimization**

$$Err_P(h_1) \leq Err_S(h_1) + \Phi(VCdim(H), n, \eta)$$

**Idea:** Structure on hypothesis space.

**Goal:** Minimize upper bound on true error rate.