

# Statistical Learning Theory

CS4780 – Machine Learning  
Fall 2009

Thorsten Joachims  
Cornell University

Reading: Mitchell Chapter 7 (not 7.4.4 and 7.5)

## Outline

### Questions in Statistical Learning Theory:

- How good is the learned rule after  $n$  examples?
- How many examples do I need before the learned rule is accurate?
- What can be learned and what cannot?
- Is there a universally best learning algorithm?

### In particular, we will address:

#### What is the true error of $h$ if we only know the training error of $h$ ?

- Finite hypothesis spaces and zero training error
- Finite hypothesis spaces and non-zero training error
- Infinite hypothesis spaces and VC dimension

## Can you Convince me of your Psychic Abilities?

### Game

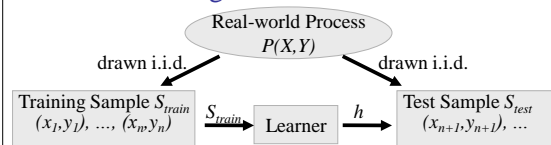
- I think of  $n$  bits
- $|H|$  players try to guess the bit sequence

**0101**

### Question:

- If at least one player guesses the bit sequence correctly, is there any significant evidence that he/she has telepathic abilities?
- How large would  $n$  and  $|H|$  have to be?

## Learning as Prediction Task



- **Goal:** Find  $h$  with small prediction error  $Err_P(h)$  over  $P(X,Y)$ .

- **Strategy:** Find (any?)  $h$  with small error  $Err_{S_{train}}(h)$  on training sample  $S_{train}$ .

- **Training Error:** Error  $Err_{S_{train}}(h)$  on training sample.
- **Test Error:** Error  $Err_{S_{test}}(h)$  on test sample is an estimate of  $Err_P(h)$ .

## Review of Definitions

**Definition:** A particular instance of a learning problem is described by a probability distribution  $P(X,Y)$ .

**Definition:** A sample  $S = ((\vec{x}_1, y_1), \dots, (\vec{x}_n, y_n))$  is independently identically distributed (i.i.d.) according to  $P(X,Y)$ .

**Definition:** The error on sample  $S$   $Err_S(h)$  of a hypothesis  $h$  is  $Err_S(h) = \frac{1}{n} \sum_{i=1}^n \Delta(h(\vec{x}_i), y_i)$ .

**Definition:** The prediction/generalization/true error  $Err_P(h)$  of a hypothesis  $h$  for a learning task  $P(X,Y)$  is

$$Err_P(h) = \sum_{\vec{x} \in X, y \in Y} \Delta(h(\vec{x}), y) P(X = \vec{x}, Y = y).$$

**Definition:** The hypothesis space  $H$  is the set of all possible classification rules available to the learner.

## Useful Formulas

- **Binomial Distribution:** The probability of observing  $x$  heads in a sample of  $n$  independent coin tosses, where in each toss the probability of heads is  $p$ , is

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

- **Union Bound:**

$$P(X_1 = x_1 \vee X_2 = x_2 \vee \dots \vee X_n = x_n) \leq \sum_{i=1}^n P(X_i = x_i)$$

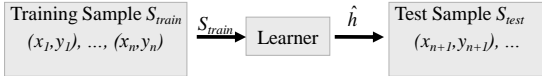
- **Unnamed:**

$$(1 - \epsilon) \leq e^{-\epsilon}$$

## Generalization Error Bound: Finite H, Zero Training Error

- **Setting**
  - Sample of  $n$  labeled instances  $S_{train}$
  - Learning Algorithm  $L$  with a finite hypothesis space  $H$
  - At least one  $h \in H$  has zero training error  $Err_{S_{train}}(h)$
  - Learning Algorithm  $L$  returns zero training error hypothesis  $\hat{h}$
- **What is the probability that the prediction error of  $\hat{h}$  is larger than  $\epsilon$ ?**

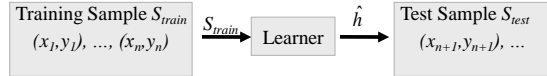
$$P(Err_P(\hat{h}) \geq \epsilon) \leq |H|e^{-\epsilon n}$$



## Sample Complexity: Finite H, Zero Training Error

- **Setting**
  - Sample of  $n$  labeled instances  $S_{train}$
  - Learning Algorithm  $L$  with a finite hypothesis space  $H$
  - At least one  $h \in H$  has zero training error  $Err_{S_{train}}(h)$
  - Learning Algorithm  $L$  returns zero training error hypothesis  $\hat{h}$
- **How many training examples does  $L$  need so that with probability  $(1-\delta)$  it learns an  $\hat{h}$  with prediction error less than  $\epsilon$ ?**

$$n \geq \frac{1}{\epsilon} (\log(|H|) - \log(\delta))$$



## Probably Approximately Correct Learning

**Definition:**  $C$  is **PAC-learnable** by learning algorithm  $\mathcal{L}$  using  $H$  and a sample  $S$  of  $n$  examples drawn i.i.d. from some fixed distribution  $P(X)$  and labeled by a concept  $c \in C$ , if for sufficiently large  $n$

$$P(Err_P(h_{\mathcal{L}(S)}) \leq \epsilon) \geq (1 - \delta)$$

for all  $c \in C, \epsilon > 0, \delta > 0$ , and  $P(X)$ .  $\mathcal{L}$  is required to run in polynomial time dependent on  $1/\epsilon, 1/\delta, n$ , the size of the training examples, and the size of  $C$ .

## Example: Smart Investing

**Task:** Pick stock analyst based on past performance.

**Experiment:**

- Review analyst prediction “next day up/down” for past 10 days.
- Pick analyst that makes the fewest errors.

**Situation 1:**

- 1 stock analyst {A1}, A1 makes 5 errors

**Situation 2:**

- 3 stock analysts {A1,B1,B2}, B2 best with 1 error

**Situation 3:**

- 1003 stock analysts {A1,B1,B2,C1,...,C1000}, C543 best with 0 errors

**Which analysts are you most confident in, A1, B2, or C543?**

## Useful Formula

### • Hoeffding/Chernoff Bound:

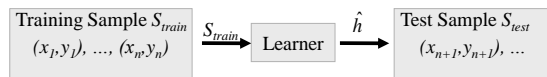
For any distribution  $P(X)$  where  $X$  can take the values 0 and 1, the probability that an average of an i.i.d. sample deviates from its mean  $p$  by more than  $\epsilon$  is bounded as

$$P\left(\left|\left(\frac{1}{n} \sum_{i=1}^n x_i\right) - p\right| > \epsilon\right) \leq 2e^{-2n\epsilon^2}$$

## Generalization Error Bound: Finite H, Non-Zero Training Error

- **Setting**
  - Sample of  $n$  labeled instances  $S$
  - Learning Algorithm  $L$  with a finite hypothesis space  $H$
  - $L$  returns hypothesis  $\hat{h}=L(S)$  with lowest training error
- **What is the probability that the prediction error of  $\hat{h}$  exceeds the fraction of training errors by more than  $\epsilon$ ?**

$$P(|Err_S(h_{\mathcal{L}(S)}) - Err_P(h_{\mathcal{L}(S)})| \geq \epsilon) \leq 2|H|e^{-2\epsilon^2 n}$$



## Example: Smart Investing

**Task: Pick stock analyst based on past performance.**

**Experiment:**

- Have analyst predict "next day up/down" for 10 days.
- Pick analyst that makes the fewest errors.

**Situation 1:**

- 1 stock analyst {A1}, A1 makes 5 errors

**Situation 2:**

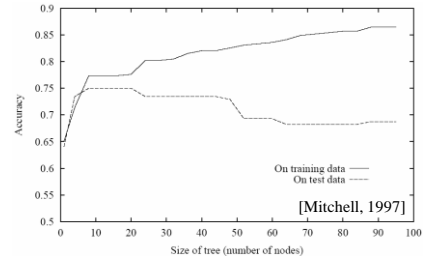
- 3 stock analysts {A1,B1,B2}, B2 best with 1 error

**Situation 3:**

- 1003 stock analysts {A1,B1,B2,C1,...,C1000}, C543 best with 0 errors

**Which analysts are you most confident in, A1, B2, or C543?**

## Overfitting vs. Underfitting



With probability at least  $(1-\delta)$ :

$$Err_P(h_{\mathcal{L}(S_{train})}) \leq Err_{S_{train}}(h_{\mathcal{L}(S_{train})}) + \sqrt{\frac{(\ln(2|H|) - \ln(\delta))}{2n}}$$

## Generalization Error Bound: Infinite H, Non-Zero Training Error

- **Setting**
  - Sample of  $n$  labeled instances  $S$
  - Learning Algorithm  $L$  with a hypothesis space  $H$  with  $VCDim(H)=d$
  - $L$  returns hypothesis  $\hat{h}=L(S)$  with lowest training error
- **Definition:** The **VC-Dimension** of  $H$  is equal to the maximum number  $d$  of examples that can be split into two sets in all  $2^d$  ways using functions from  $H$  (shattering).
- Given hypothesis space  $H$  with  $VCDim(H)$  equal to  $d$  and an i.i.d. sample  $S$  of size  $n$ , with probability  $(1-\delta)$  it holds that

$$Err_P(h_{\mathcal{L}(S)}) \leq Err_S(h_{\mathcal{L}(S)}) + \sqrt{\frac{d \left( \ln \left( \frac{2n}{d} \right) + 1 \right) - \ln \left( \frac{\delta}{4} \right)}{n}}$$