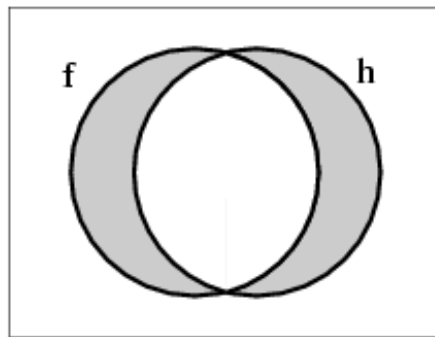


# Statistical Learning Theory

The specific Question we address here is: What are the theoretical bounds on the error rate of  $h$  on new data points?

## General Assumptions (Noise-Free Case)

1. Assumptions on data  $S$ : Examples are iid (independently identically distributed) generated according to a probability distribution  $D(\mathbf{x})$  and labeled according to an unknown function  $y = f(\mathbf{x})$  (classification).
2. Assumptions on learning algorithm: The learning algorithm is given  $m$  examples in  $S$  and outputs a hypothesis  $h \in H$  that is consistent with  $S$  (i.e. correctly classifies them all).
3. Goal Assumption:  $h$  should fit new examples well (low  $\varepsilon$  error rate) that are draw according to the same distribution  $D(\mathbf{x})$



$$error(h, f) = P_{D(\mathbf{x})} [f(\mathbf{x}) \neq h(\mathbf{x})]$$

## PAC (Probably – Approximately Correct) Learning

We allow algorithms to fail with probability  $\delta$

Procedure:

1. Draw  $m$  random samples  $\rightarrow S$
2. Run a learning algorithm and generate  $h$

Since  $S$  is iid we cannot guarantee that the data will be representative  
 -Want to ensure that  $1 - \delta$  of the time, the hypothesis error is less than  $\epsilon$

$$P_D^m [\text{error}(f, h) > \epsilon] < \delta$$

Ex: want to obtain a 90% ( $\epsilon = .9$ ) correct hypothesis 95% ( $\delta = 0.05$ ) of the time.

## Case 1: Finite Hypothesis Spaces

Assume  $H$  is finite

Consider  $h_1 \in H$  such that  $\text{error}(h_1, f) > \epsilon$  - ( $\epsilon$ -bad).

Given one training example  $(\mathbf{x}_1, y_1)$ , the probability that  $h_1$  classifies it correctly is

$$P[h_1(\mathbf{x}_1) = y_1] \leq (1 - \epsilon)$$

Given  $m$  training examples  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)$ , the probability that  $h_1$  classifies it correctly is:

$$P[(h_1(\mathbf{x}_1) = y_1) \wedge \dots \wedge (h_1(\mathbf{x}_m) = y_m)] \leq (1 - \epsilon)^m$$

Now, assume we have a second hypothesis  $h_2$  (also  $\epsilon$ -bad). What is the probability that either  $h_1$  or  $h_2$  will be correct?

$$\begin{aligned} P_D^m [(h_1 \text{ correct}) \vee (h_2 \text{ correct})] &\leq P_D^m [(h_1 \text{ correct})] + P_D^m [(h_2 \text{ correct})] - P_D^m [(h_1 \text{ correct}) \wedge (h_2 \text{ correct})] \\ &\leq P_D^m [(h_1 \text{ correct})] + P_D^m [(h_2 \text{ correct})] \\ &\leq 2(1 - \epsilon)^m \end{aligned}$$

Therefore, for  $k$   $\epsilon$ -bad hypothesis: the probability that any one of them are correct is:

$$\leq k(1 - \epsilon)^m$$

Since  $k \leq |H|$

$$\leq |H|(1 - \epsilon)^m$$

Inequality:  $0 \leq \epsilon \leq 1 \Rightarrow (1 - \epsilon) \leq e^{-\epsilon}$

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

**Lemma:** For a finite hypothesis space  $H$ , given a set of  $m$  training examples drawn independently according to  $D$ , the probability that there exists an hypothesis  $h \in H$  with true error greater than  $\epsilon$  consistent with the training examples is less than or equal to  $|H|e^{-m\epsilon}$ .

Therefore, for probability less than  $\delta$

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

This is true whenever

$$m \geq \frac{1}{\varepsilon} \left[ \ln |H| + \ln \frac{1}{\delta} \right]$$

(Blumer bound – Blumer, Ehrenfeucht, Haussler, and Warmuth 1987).

Therefore, if  $h \in H$  is consistent and all  $m$  samples are independently drawn according to  $D$ , the error rate  $\varepsilon$  on new data points is bounded by

$$\varepsilon \geq \frac{1}{m} \left[ \ln |H| + \ln \frac{1}{\delta} \right]$$

Example applications:

- Boolean Conjunctions over  $n$  features
  - Three possibilities  $x_j, \neg x_j$  or not present. Therefore for  $n$  features

$$|H| = 3^n$$

- $\varepsilon \geq \frac{1}{m} \left[ n \ln 3 + \ln \frac{1}{\delta} \right]$

## Finite Hypothesis Spaces: Inconsistent Hypothesis

If  $h$  does not perfectly fit the data, but has error rate of  $\varepsilon_s$

$$\varepsilon \geq \varepsilon_s + \sqrt{\frac{\ln |H| + \ln \frac{1}{\delta}}{2m}}$$

Therefore larger than the error rate on  $\varepsilon_s$

## Case 2: Infinite Hypothesis Spaces

Even if  $|H| = \infty$ ,  $H$  has limited expressive power, therefore we should still be able to obtain bounds.

**Definition:** Let  $S = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$  be a set of  $m$  examples. A hypothesis space  $H$  can *trivially fit*  $S$ , if for every possible labeling of the examples in  $S$ , there exists an  $h \in H$  that gives this labeling. If so, then  $H$  is said to *shatter*  $S$ .

**Definition:** The Vapnik-Chervonenkis dimension (VC-dimension) of a hypothesis space  $H$  is the size of the largest set of examples that can be trivially fit (shattered) by  $H$ .

Note: if  $H$  is finite,  $VC(H) \leq \log_2 |H|$ .

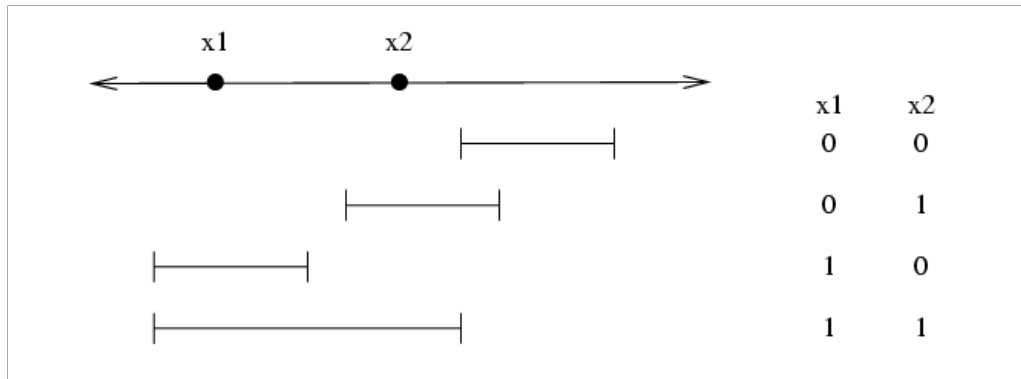
## VC-Dimension Example

Let  $H$  be the set of all intervals on the real line.

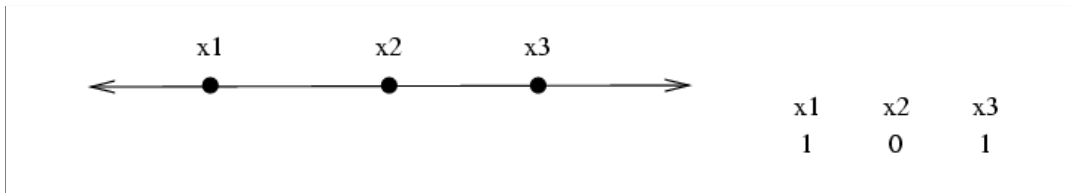
If  $h(x) = 1$  than  $x$  is in the interval.

If  $h(x) = 0$  than  $x$  is NOT in the interval.

$H$  can trivially fit (shatter) any two points.



However, can  $H$  trivially fit (shatter) three points?



No. Therefore the VC-dimension is 2.

## Error Bound for Infinite Hypothesis Spaces

**Theorem:** Suppose that  $VC(H) = d$ . Assume that there are  $m$  training examples in  $S$ , and that a learning algorithm finds an  $h \in H$  with error rate  $\varepsilon_S$  on  $S$ . Then, with probability  $1 - \delta$ , the error rate  $\varepsilon$  on new data is:

$$\varepsilon \leq 2\varepsilon_S + \frac{4}{m} \left[ d \log \frac{2em}{d} + \log \frac{4}{\delta} \right]$$

Called the **Empirical Risk Minimization Principle (Vapnik)**.

However, this does not work well for fixed hypothesis spaces because your learning algorithm will minimize  $\varepsilon_S$ :

- **Underfitting:** Every hypothesis  $H$  has high error  $\varepsilon_S$ . Want to consider  $H'$  that has larger space.
- **Overfitting:** Every hypothesis  $H$  has high error  $\varepsilon_S = 0$ . Want to consider  $H'$  smaller hypothesis spaces that have lower  $d$ .

Suppose we have a nested series of hypothesis spaces:

$$H_1 \subseteq H_2 \subseteq \dots \subseteq H_k \subseteq$$

with corresponding VC dimensions and errors

$$d_1 \leq d_2 \leq \dots \leq d_k \leq \dots$$

$$\varepsilon_S^1 \geq \varepsilon_S^2 \geq \dots \geq \varepsilon_S^k \geq \dots$$

Then, you should use the **Structural Risk Minimization Principle** (Vapnik).

Choose the hypothesis space  $H_k$  that minimizes the combined error bounds:

$$\varepsilon \leq 2\varepsilon_S^k + \frac{4}{m} \left[ d_k \log \frac{2em}{d_k} + \log \frac{4}{\delta} \right]$$