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 \( Err_{Strain}(h) \) on training sample \( Strain \).

Review of Definitions

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

Definition: A sample \( S = ((x_1, y_1), \ldots, (x_n, y_n)) \) 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 \( Err_S(h) = \frac{1}{n} \sum_{i=1}^{n} \Delta(h(x_i), y_i) \).

Definition: The prediction/generalization/true error \( Err_P(h) \) of a hypothesis \( h \) for a learning task \( P(X,Y) \) is \( Err_P(h) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \Delta(h(x), y) P(X=x, Y=y) \).

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

• Union Bound:
  \[ P(X_1 = x_1 \lor X_2 = x_2 \lor \ldots \lor X_n = x_n) \leq \sum_{i=1}^{n} P(X_i = x_i) \]

• Unnamed:
  \[ (1 - \epsilon) \leq e^{-\epsilon} \]
**Generalization Error Bound:**

**Finite H, Zero Training Error**

- **Setting**
  - Sample of $n$ labeled instances $S_{\text{train}}$
  - Learning Algorithm $L$ with a finite hypothesis space $H$
  - At least one $h \in H$ has zero training error $Err_{\text{train}}(h)$
  - Learning Algorithm $L$ returns zero training error hypothesis $\hat{h}$

- What is the probability that the prediction error of $\hat{h}$ is larger than $\varepsilon$?

$$P(Err_{\text{test}}(\hat{h}) \geq \varepsilon \leq |H|e^{-\varepsilon^2})$$

<table>
<thead>
<tr>
<th>Training Sample $S_{\text{train}}$ $(x_1,y_1), \ldots, (x_n,y_n)$</th>
<th>Learner</th>
<th>$\hat{h}$</th>
<th>Test Sample $S_{\text{test}}$ $(x_{n+1},y_{n+1}), \ldots$</th>
</tr>
</thead>
</table>

**Sample Complexity:**

**Finite H, Zero Training Error**

- **Setting**
  - Sample of $n$ labeled instances $S_{\text{train}}$
  - Learning Algorithm $L$ with a finite hypothesis space $H$
  - At least one $h \in H$ has zero training error $Err_{\text{train}}(h)$
  - 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 $\varepsilon^2$?

$$n \geq \frac{1}{\varepsilon^2} \log(|H|) - \log(\delta)$$

<table>
<thead>
<tr>
<th>Training Sample $S_{\text{train}}$ $(x_1,y_1), \ldots, (x_n,y_n)$</th>
<th>Learner</th>
<th>$\hat{h}$</th>
<th>Test Sample $S_{\text{test}}$ $(x_{n+1},y_{n+1}), \ldots$</th>
</tr>
</thead>
</table>

**Probably Approximately Correct Learning**

**Definition:** $C$ is **PAC-learnable** by learning algorithm $L$ 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 $n$

$$P(Err_{\text{test}}(\hat{h}_L(S)) \leq \varepsilon \leq (1-\delta))$$

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

**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 $\varepsilon$ is bounded as

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

**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 $\{A_1\}$, $A_1$ makes 5 errors

**Situation 2:**
- 3 stock analysts $\{A_1,B_1,B_2\}$, $B_2$ best with 1 error

**Situation 3:**
- 1003 stock analysts $\{A_1,B_1,B_2,C_1,\ldots,C_{1000}\}$, $C_{543}$ best with 0 errors

Which analysts are you most confident in, $A_1$, $B_2$, or $C_{543}$?

**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 $\varepsilon$?

$$P\left(\left|Err_{\text{test}}(\hat{h}_L(S)) - Err_{\text{train}}(\hat{h}_L(S))\right| \geq \varepsilon\right) \leq 2|H|e^{-\varepsilon^2 n}$$

| Training Sample $S_{\text{train}}$ $(x_1,y_1), \ldots, (x_n,y_n)$ | Learner | $\hat{h}$ | Test Sample $S_{\text{test}}$ $(x_{n+1},y_{n+1}), \ldots$ |
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 \(A_1\), \(A_1\) makes 5 errors

Situation 2:
– 3 stock analysts \(\{A_1,B_1,B_2\}\), \(B_2\) best with 1 error

Situation 3:
– 1003 stock analysts \(\{A_1,B_1,B_2,C_1,\ldots,C_{1000}\}\), C543 best with 0 errors

Which analysts are you most confident in, \(A_1\), \(B_2\), or C543?

Overfitting vs. Underfitting

With probability at least \(1-\delta\):
\[
Err_{\text{train}}(h) \leq Err_{\text{test}}(h) + \sqrt{\frac{(\ln(2d_H)) - \ln(\delta)}{2n}}
\]

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

\[
Err_{\text{test}}(h_{|S}) \leq Err_{\text{train}}(h_{|S}) + \sqrt{\frac{d}{n} \left(\ln\left(\frac{2d}{\delta}\right) + 1\right) - \ln\left(\frac{2}{\delta}\right)}
\]