Occam’s Razor Theorem

**IID assumption:** Assume that there is a probability distribution $D$ over pairs $\langle x, y \rangle$ where we would like to predict $y$ given only $x$.

**0-1 Assumption:** $y$ is always either 0 or 1 and a predictive hypothesis $h$ is a rule such that $h(x)$ is either 0 or 1.

An example might be a neural net threshold unit for recognizing the digit 7.

$$\text{err}(h) \equiv \mathbb{P}_{\langle x, y \rangle \sim D} (h(x) \neq y)$$

Let $|h|$ be the number of bits needed to name the rule $h$ in some fixed prefix-free coding language.

Let $S$ be a sample of $m$ pairs $\langle x_1, y_1 \rangle, \ldots, \langle x_m, y_m \rangle$.

Let $\hat{\text{err}}(h)$ and $1[\Phi]$ be defined as follows.

$$\hat{\text{err}}(h) \equiv \frac{1}{m} \sum_{i=1}^{m} 1[h(x_i) \neq y_i]$$

$$1[\Phi] \equiv \begin{cases} 1 & \text{if } \Phi \text{ is true} \\ 0 & \text{otherwise} \end{cases}$$

Let $\forall^\delta S \; \Phi[S, \delta]$ mean that with probability at least $1 - \delta$ over the random choice of the sample $S$ we have that $\Phi[S, \delta]$ holds.

**Theorem:** For samples of size $m$ we have the following.

$$\forall^\delta S \quad \forall h \in H \quad \text{err}(h) \leq \hat{\text{err}}(h) + \sqrt{\frac{(\ln 2)|h| + \ln \frac{1}{\delta}}{2m}}$$

This is a PAC (Probably Approximately Correct) theorem in the sense that probably (with probability $1 - \delta$ over the choice of the sample) the training error rate of simple rules is approximately correct in the sense that it is approximately equal to the generalization error rate.

The Proof uses the Following

- **Chernoff Bound:** $P(\text{err} > \hat{\text{err}} + \epsilon) \leq e^{-2m\epsilon^2}$ We will not prove this here.

- **Union Bound:** $P(\exists x \Phi[x]) \leq \sum_x P(\Phi[x])$ This is a generalization of $P(\Phi \text{ or } \Psi) \leq P(\Phi) + P(\Psi)$. 
• Kraft Inequality: \( \sum_h 2^{-|h|} \leq 1 \)

The Kraft inequality holds for prefix codes — a set of code words where no code word is a proper prefix of any other code word. Null terminated character strings (or byte strings) are prefix codes. To prove the Kraft inequality consider randomly generating one bit at a time and stopping when you have a code for a rule. Then \( 2^{-|h|} = P(h) \).

**Proof:**

We call a rule “bad” (relative to a given sample) if it violates the theorem. More specifically we have the following.

\[
\text{bad}(h) \equiv \left[ \text{err}(h) > \hat{\text{err}}(h) + \sqrt{2^{-|h|} + \frac{\ln 2}{2m}} \right]
\]

\[
P(\text{bad}(h)) \leq e^{-2me^2}
\]
\[
= \delta 2^{-|h|}
\]
\[
P(\exists h \text{ bad}(h)) \leq \sum_h \delta 2^{-|h|}
\]
\[
= \delta \sum_h 2^{-|h|} \leq \delta
\]

1 A Bayesian Interpretation of the Occam Theorem

Let \( P \) range over probability distributions on rules. Define \( |h|_P \) as follows.

\[
|h|_P = \log_2 \frac{1}{P(h)}
\]

We now have the following theorem.

\[
\forall P \forall \delta S \forall h \in H \quad \text{err}(h) \leq \hat{\text{err}}(h) + \sqrt{\frac{(\ln 2)|h|_P + \ln \frac{1}{\delta}}{2m}}
\]

This is now a “Bayesian” theorem in the sense that it is based on an arbitrary “prior”.

2
2 The Realizable Case

\[ \forall \delta \forall h \in H \text{ if } \hat{\text{err}}(h) = 0 \text{ then } \text{err}(h) \leq \frac{(\ln 2)|h| + \ln \frac{1}{\delta}}{m} \]

If \( \hat{\text{err}}(h) = 0 \) this bound is much tighter than the above Occam theorem. Again we say that \( h \) is bad (for a given sample) if it violates the theorem.

\[ \text{bad}(h) \equiv \left[ \hat{\text{err}}(h) = 0 \text{ and } \text{err}(h) > \frac{(\ln 2)|h| + \ln \frac{1}{\delta}}{m} \right] \]

We need only consider rules \( h \) with \( \text{err}(h) > [(\ln 2)|h| + \ln(1/\delta)]/m \). For such rules we have the following.

\[ P(\text{bad}(h)) = P(\hat{\text{err}}(h) = 0) \]

\[ = (1 - \text{err}(h))^m \]

\[ \leq e^{-\text{err}(h)m} \text{ using } 1 - \epsilon \leq e^{-\epsilon} \]

\[ \leq \delta 2^{-|h|} \]

\[ P(\exists h \text{ bad}(h)) \leq \sum h \delta 2^{-|h|} \leq \delta \]

3 Combining the Realizable and the Unrealizable Cases

\[ \forall \delta \forall h \in H \text{ err}(h) \leq \hat{\text{err}}(h) \]

\[ + \sqrt{\frac{2\hat{\text{err}}(h)((\ln 2)|h| + \ln \frac{1}{\delta})}{m}} \]

\[ + \frac{2((\ln 2)|h| + \ln \frac{1}{\delta})}{m} \]

It is interesting to consider the case where \( \hat{\text{err}}(h) = 1/2 \) and \( \hat{\text{err}}(h) = 0 \).
4 The Tightest Version

\[ KL(q||p) = q \ln \frac{q}{p} + (1 - q) \ln \frac{1-q}{1-p} \]

For \( \epsilon \) small (positive or negative) we have the following.

\[ KL(q + \epsilon||q) \approx \frac{\epsilon^2}{2q} \]

Theorem:

\[ \forall \delta \forall \ h \ KL(\hat{err}(h)||err(h)) \leq \frac{(\ln 2)|h| + \ln \frac{2}{\delta}}{m} \]

This theorem can be proved from the following two concentration inequalities.

\[ \text{for } p \leq \text{err}(h) : \quad P(\hat{err}(h) \leq p) \leq e^{-m KL^+(p||err(h))} \]
\[ \text{for } p \geq \text{err}(h) : \quad P(\hat{err}(h) \geq p) \leq e^{-m KL^+(p||err(h))} \]

The preceding theorem can then be proved using the fact that for \( p \leq q \) we have \( KL(p||q) \geq \frac{(p-q)^2}{2q} \).

5 Problem

The following is the “two sided” form of the Chernoff bound.

\[ P(|\hat{err}(h) - \text{err}(h)| \geq \epsilon) \leq 2e^{-2m\epsilon^2} \]

Use this inequality (and the Union bound and Kraft inequality) to prove the following.

\[ \forall \delta \forall \ h \ |\hat{err}(h) - \text{err}(h)| \leq \sqrt{\frac{(\ln 2)|h| + \ln \frac{2}{\delta}}{2m}} \]