The Simplest model of Generalization

Model Definition

$X = \text{input space}$

$Y = \{0, 1\} = \text{output space}$

$c : X \rightarrow Y = \text{classifier}$

Model Assumption: All samples are drawn independently from some unknown distribution $D(x, y)$.

$S = (x, y)^m \sim D^m$ is a sample set.
Model: Derived quantities

The thing we want to know:

\[ c_D \equiv \Pr_{x,y \sim D}(c(x) \neq y) = \text{true error} \]
Model: Derived quantities

The thing we want to know:

\[ c_D \equiv \Pr_{x,y \sim D} (c(x) \neq y) = \text{true error} \]

The thing we have:

\[ \hat{c}_S \equiv \sum_{(x,y) \in S} I [c(x) \neq y] \]

= “train error”, “test error”, or “observed error”, depending on context.
Model: Basic Observations

Q: What is the distribution of $\hat{c}_S$?

A: A Binomial.

\[
\Pr_{S \sim D^m}(\hat{c}_S = k | c_D) = \binom{m}{k} c_D^k (1 - c_D)^{m-k}
\]

= probability of $k$ heads (errors) in $m$ flips of a coin with bias $c_D$. 
Model: basic quantities

We use the cumulative:

$$\text{Bin}(m, k, c_D) = \Pr_{S \sim D^m}(\hat{c}_S \leq k \mid c_D)$$

$$= \sum_{i=0}^{k} \binom{m}{i} c_D^i (1 - c_D)^{m-i}$$

= probability of observing $k$ or fewer “heads” (errors) with $m$ coins.
Model: basic quantities

Need confidence intervals ⇒ use the pivot of the cumulative instead

\[
\overline{\text{Bin}}(m, k, \delta) = \max \{p : \text{Bin}(m, k, p) \geq \delta\}
\]

= the largest true error such that the probability of observing \( k \) or fewer “heads” (errors) is at least \( \delta \)
Test Set Bound: Theorem

Theorem: (Test Set Bound) For all classifiers $c$, for all $D$, for all $\delta \in (0, 1]$:

$$\Pr_{S \sim D^m} (c_D \leq \overline{\text{Bin}} (m, \hat{c}_S, \delta)) \geq 1 - \delta$$

World’s easiest proof: (by contradiction).

Assume $\text{Bin} (m, k, c_D) \geq \delta$ (which is true with probability $1 - \delta$).

Then by definition, $\overline{\text{Bin}} (m, \hat{c}_S, \delta) \geq c_D$
Occam’s Razor Bound

Theorem: (Occam’s Razor Bound) For all “priors” $P(c)$ over the classifiers $c$, for all $D$, for all $\delta \in (0, 1]$:

$$\Pr_{S \sim D^m} \left( \forall c : c_D \leq \text{Bin}(m, \hat{c}_S, \delta P(c)) \right) \geq 1 - \delta$$

Corollary: For all $P(c)$, for all $D$, for all $\delta \in (0, 1]$:

$$\Pr_{S \sim D^m} \left( c_D \leq \frac{\hat{c}_S}{m} + \sqrt{\frac{\ln \frac{1}{P(c)} + \ln \frac{1}{\delta}}{2m}} \right) \geq 1 - \delta$$
Occam's Razor Bound: Proof

Test set bound ⇒

∀c \Pr_{S \sim D_m} \left(c_D \leq \text{Bin}(m, \hat{c}_S, \delta P(c)) \right) \geq 1 - \delta P(c)
Occam’s Razor Bound: Proof

Test set bound \( \Rightarrow \)

\[ \forall c \Pr_{S \sim D_m} \left( c_D \leq \text{Bin} \left( m, \hat{c}_S, \delta P(c) \right) \right) \geq 1 - \delta P(c) \]

Negate to get:

\[ \forall c \Pr_{S \sim D_m} \left( c_D > \text{Bin} \left( m, \hat{c}_S, \delta P(c) \right) \right) < \delta P(c) \]
Occam’s Razor Bound: Proof

Test set bound ⇒

\[ \forall c \Pr_{S \sim D^m} \left( c_D \leq \text{Bin} \left( m, \hat{c}_S, \delta P(c) \right) \right) \geq 1 - \delta P(c) \]

Negate to get:

\[ \forall c \Pr_{S \sim D^m} \left( c_D > \text{Bin} \left( m, \hat{c}_S, \delta P(c) \right) \right) < \delta P(c) \]

Apply union bound: \( \Pr(A \text{ or } B) \leq \Pr(A) + \Pr(B) \) repeatedly.

\[ \Pr_{S \sim D^m} \left( \exists c : c_D > \text{Bin} \left( m, \hat{c}_S, \delta P(c) \right) \right) < \sum_c \delta P(c) = \delta \]
Occam’s Razor Bound: Proof

Test set bound ⇒

$$\forall c \quad \Pr_{S \sim D^m} \left( c_D \leq \overline{\text{Bin}}(m, \hat{c}_S, \delta P(c)) \right) \geq 1 - \delta P(c)$$

Negate to get:

$$\forall c \quad \Pr_{S \sim D^m} \left( c_D > \overline{\text{Bin}}(m, \hat{c}_S, \delta P(c)) \right) < \delta P(c)$$

Apply union bound:

$$\Pr_{S \sim D^m} \left( \exists c : c_D > \overline{\text{Bin}}(m, \hat{c}_S, \delta P(c)) \right) < \sum_c \delta P(c) = \delta$$

Negate again to get proof.