Overfitting and Sample Complexity

Columbia COMS4771, 2007

John Langford
Outline

1. The Basic Model

2. The Test Set Bound

3. Occam’s Razor Bound
Model: Definitions

\[ X = \text{input space} \]

\[ Y = \{0, 1\} = \text{output space} \]

\[ c : X \rightarrow Y = \text{classifier} \]

Model: Basic 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:

\[ \hat{c}_S \equiv m \Pr_{x,y \sim S} (c(x) \neq y) = \sum_{i=1}^{m} I [c(x) \neq y] \]

= “train error”, “test error”, or “observed error”, depending on context.

(note: we identify the set $S$ with the uniform distribution on $S$)
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 \( c_D \).
Possible Error distributions

Probability vs. Empirical Error Rate

- red line: true error

[Graph showing distributions with different colors representing possible error distributions]
Model: basic quantities

We use the cumulative:

\[ \text{Bin}(m, k, c_D) = \Pr_{S \sim D^m} (\hat{c}_S \leq k \mid c_D) = \sum_{i=0}^{k} \binom{m}{i} c_D^i (1 - c_D)^{m-i} \]

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

Need confidence intervals ⇒ use the pivot of the cumulative instead

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

= the largest true error such that the probability of observing \( k \) or fewer “heads” (errors) is at least \( \delta \).
Outline

1. The Basic Model

2. The Test Set Bound

3. Occam’s Razor Bound
Test Set Bound: Setting

Standard technique:

1. Cut the data into train set and test set

2. Train on the train set

3. Test on the test set

What do sample complexity say about this method?
Test Set Bound: Theorem

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

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

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, $\overline{\text{Bin}}(m, \hat{c}_S, \delta) \geq c_D$
Observation and Consistent Binomials

Empirical Error Rate
Test Set Bound Notes

Perfectly tight: There exist true error rates achieving the bound

Lower bound of the same form.

Primary use: verification of successful learning
What does Test Set Bound mean?

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

$$\Pr_{S \sim D^m} \left( \text{KL} \left( \frac{\hat{c}_S}{m} \| c_D \right) \leq \frac{\ln \frac{1}{\delta}}{m} \right) \geq 1 - \delta$$

where $\text{KL}(q\|p) = q \ln \frac{q}{p} + (1 - q) \ln \frac{1 - q}{1 - p}$ for $q < p$

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

$$\Pr_{S \sim D^m} \left( c_D \leq \frac{\hat{c}_S}{m} + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}} \right) \geq 1 - \delta$$

**Proof:** Use the Chernoff approximation. Full details in the notes.
Test Set Bound: Example

Suppose $\delta = 0.1$

Suppose $m = 100$

Suppose $\hat{c}_S = 2$

Square root Chernoff bound: $\Rightarrow c_D \in [-0.102, 0.142]$

Exact calculation $\Rightarrow c_D \in [0.0045, 0.0616]$
Outline

1. The Basic Model

2. The Test Set Bound

3. Occam’s Razor Bound
Training Set Bounds in General

• Sometimes a holdout set is critical for learning.

• Sometimes we want bounds to guide learning

⇒ Train set bounds

Occam’s Razor bound is the simplest train set bound.
Occam’s Razor Bound Protocol

Verifier

\[ \delta \]

Draw Training Examples

m examples

"Prior", \( P(c) \)

Evaluate Bound

Learner

Choose \( c \)

classifier, \( c \)
Occam’s Razor Bound

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

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

Compare with test set bound: $\delta \rightarrow \delta P(c)$.

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 \( \Rightarrow \)

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

Test set bound ⇒

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

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

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: \( \Pr(A \text{ or } B) \leq \Pr(A) + \Pr(B) \) repeatedly.
\[ \Pr_{S \sim D^m} \left( \exists c : c_D > \overline{\text{Bin}}(m, \hat{c}_S, \delta P(c)) \right) < \sum_c \delta P(c) = \delta \]
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: \( \Pr(A \text{ or } B) \leq \Pr(A) + \Pr(B) \) repeatedly.

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

Negate again to get proof.

Next: Graphical proof
Each classifier is a Binomial with a different size tail cut. With high probability no error falls in any tail.
The chosen classifier has an unknown true error rate.
Bound = the largest true error rate for which the observation is not in the tail.
Occam’s Razor Bound: Example

Suppose $\delta = 0.1$

Suppose $m = 100$

Suppose $P(c) = 0.1$

Suppose $\hat{c}_S = 2$

Square root Chernoff $\Rightarrow c_D \in [-0.143, 0.183]$

Exact calculation $\Rightarrow c_D \in [0.001, 0.089]$
Conclusion

1. A real confidence interval to compare classifiers is good.

2. Test set bound very simple.

3. Train set bounds tell you something about how to design an algorithm, but are somewhat loose also.

Code for bound calculation at:

http://hunch.net/~jl/projects/prediction_bounds/bound/bound.html
Midterm Thursday!