Strong Asymptotic Assertions for Discrete MDL in Regression and Classification

or

A Strange Way of Proving Consistency of MDL Learning

Jan Poland and Marcus Hutter

IDSIA • Lugano • Switzerland
Focus of this Talk

Regression Classification

Sequence Prediction

this talk
this paper
COLT’04
Why Consistency?

- Consistent learners will learn the right thing (at least) in the limit
- Not all learners are consistent
- The learner should have at least the chance to be consistent (proper learning)
- Consistency is a desirable property

What is “learning the right thing“?

- Identify the exact data generating distribution
- Learn the predictive distribution
**Setup**

- Given some *training data* \((x_{1:n}, y_{1:n})\)
- where \(x_i \in \mathcal{X}\) and \(y_i \in \{0, 1\}\) for \(1 \leq i \leq n\)
- Given a new input \(x \in \mathcal{X}\), what is the corresponding output \(y\)?
- More advanced question: What is the probability that \(y(x) = 1\)?
- Solution: Train a SVM, a Neural Net, ...
Bayesian Framework

- A *model* is a function $\nu$ from $\mathcal{X}$ to the probability measures on $\{0, 1\}$
- Let $\mathcal{C}$ be a *countable* model class
- Each $\nu \in \mathcal{C}$ is assigned a *prior weight* $w_\nu > 0$
- Kraft inequality: $\sum_{\nu \in \mathcal{C}} w_\nu \leq 1$
- Example: $\mathcal{C}^{\text{lin}^2} \cong \mathbb{Q}^2$ is the class of rational linear separators on the plane
Proper Learning assumption:

- The inputs $x \in \mathcal{X}$ are generated by some arbitrary mechanism
- The outputs $y$ are generated by a distribution $\mu \in \mathcal{C}$

Online learning: Learn predictive distribution $\mu(\cdot | x_{1:t}, y_{<t})$ for increasing data $(x_{<t}, y_{<t})$
Bayes Mixture

- Then, given \((x_{1:n}, y_{1:n})\), predict according to the *Bayes mixture*

\[
\xi(y_{n+1} | x_{1:n+1}, y_{1:n}) = \frac{\sum_\nu w_\nu \prod_{t=1}^{n+1} \nu(y_t | x_t)}{\sum_\nu w_\nu \prod_{t=1}^{n} \nu(y_t | x_t)}
\]

- The Bayes mixture is the *best* we can do under the Bayesian assumptions, *but*:
  - it is costly to evaluate and to approximate
  - it may output a distribution not within \(\mathcal{C}\) (in particular for regression)
Static MDL

Therefore, we might prefer \textit{MDL} (or MAP):

\[
Q_{\text{static}}(y_{n+1}|x_{1:n+1}, y_{1:n}) = \nu^*(x_{1:n}, y_{1:n})(y_{n+1}|x_{n+1})
\]

where

\[
\nu^*(x_{1:n}, y_{1:n}) = \arg \max_{\nu \in \mathcal{C}} \left\{ w_{\nu} \nu(y_{1:n}|x_{1:n}) \right\}
\]

Determine and use the \textit{most plausible model} from \( \mathcal{C} \).
Dynamic MDL

The term static MDL is opposed to non-normalized and normalized dynamic MDL, which we need for the proofs:

\[
\mathcal{Q}(y_n | y < n) = \frac{\mathcal{Q}(y_{1:n} | x_{1:n})}{\mathcal{Q}(y < n | x < n)}
\]

\[
\bar{\mathcal{Q}}(y_n | y < n) = \frac{\mathcal{Q}(y_{1:n} | x_{1:n})}{\sum_{y_n} \mathcal{Q}(y_{1:n} | x_{1:n})}
\]

with \( \mathcal{Q}(y_{1:n} | x_{1:n}) = \max_{\nu \in \mathcal{C}} \{ \omega_{\nu} \nu(y_{1:n} | x_{1:n}) \} \).

This means: compute a new estimate for each possible \( y_n \). Note that the dynamic MDL predictor may be not a probability density (mass more than 1).
**Distance and Convergence**

**Hellinger distance** of two predictive distributions:

\[
h^2_t(\mu, \psi) = \sum_{y_t \in \{0,1\}} \left( \sqrt{\mu(y_t|x_{1:t}, y_{<t})} - \sqrt{\psi(y_t|x_{1:t}, y_{<t})} \right)^2.
\]

Then the \(\psi\)-predictions converge to the \(\mu\)-predictions in **Hellinger sum** if

\[
H^2_{x<\infty}(\mu, \psi) = \sum_{t=1}^{\infty} \mathbb{E}[h^2_t(\mu, \psi)] < \infty.
\]

This implies \(h^2_t \to 0\) **almost surely**.
Other Distance Measures

\[ s_t(\mu, \psi) = \sum_{y_t \in \{0,1\}} \left( \mu(y_t|x_{1:t}, y_t) - \psi(y_t|x_{1:t}, y_t) \right)^2 \]

square distance

\[ a_t(\mu, \psi) = \sum_{y_t \in \{0,1\}} \left| \mu(y_t|x_{1:t}, y_t) - \psi(y_t|x_{1:t}, y_t) \right| \]

absolute distance

\[ d_t(\mu, \psi) = \sum_{y_t \in \{0,1\}} \mu(y_t|x_{1:t}, y_t) \cdot \ln \frac{\mu(y_t|x_{1:t}, y_t)}{\psi(y_t|x_{1:t}, y_t)} \]

Kullback-Leibler divergence
Distance Measures: Properties

- Hellinger distance $h_t$:
  \[
  \begin{cases}
  \text{triangle inequality} \\
  \leq a_t \\
  \leq d_t
  \end{cases}
  \text{implies arbitrary loss bounds}

- Quadratic distance $s_t$:
  \[
  \begin{cases}
  \text{triangle inequality} \\
  \leq a_t \\
  \leq d_t
  \end{cases}
  \text{implies arbitrary loss bounds}

- Absolute distance $a_t$:
  \[
  \begin{cases}
  \text{triangle inequality} \\
  \leq d_t
  \end{cases}
  \]

- Kullback-Leibler divergence $d_t$:
  \[
  \begin{cases}
  \text{triangle inequality} \\
  \leq a_t
  \end{cases}
  \]
Convergence Theorem

Recall $\mu \in \mathcal{C}$ (proper learning), and $w_\mu$ is the prior weight of $\mu$, then

$$\mathcal{Q}^{\text{static}} \xrightarrow{} \mathcal{Q} \xrightarrow{} \overline{\mathcal{Q}} \xrightarrow{} \mu$$

$$\sum_t a_t \leq 3w_\mu^{-1} \quad \sum_t a_t \leq 2w_\mu^{-1} \quad \sum_t d_t \leq 2w_\mu^{-1}$$

$$\Rightarrow H^2(\mu, \mathcal{Q}^{\text{static}}) \leq 21w_\mu^{-1}$$
Properties of the Proof

- Purely algebraic
- no hidden O-terms
- Inspired by Solomonoff’s proof for universal induction
Loss bounds

- Assume that predictions entail a loss $\ell(y, \tilde{y}|x)$
- Loss depends on input $x$, true output is $y$, and prediction $\tilde{y}$
- Then we should predict in order to minimize expected loss wrt. our current believe (Bayes-optimal)
- $L$ denotes cumulative expected loss
- **Loss bound:**

  \[
  L(\rho) \leq L(\mu) + 42w_{\mu}^{-1} + 2\sqrt{42w_{\mu}^{-1} L(\mu)}
  \]

- $\Rightarrow$ expected per-round regret converges to zero almost surely
• $w_{\mu}^{-1}$ may be huge
  - Similar bounds hold for the Bayes mixture, e.g.
    \[ H^2(\mu, \xi) \leq \ln w_{\mu}^{-1} \]
  - $\Rightarrow$ Bayes mixture converges much faster in general
  - The $w_{\mu}^{-1}$ bound for MDL is sharp in general
  - With carefully chosen model class and prior, MDL converges fast, too
Discussion

- \( \mu \in C \)
- This condition may be important!
- Weak condition for *universal model class* \( \cong \) all programs on some universal Turing machine

Thank you!