Strong Asymptotic Assertions for Discrete MDL in Regression and Classification

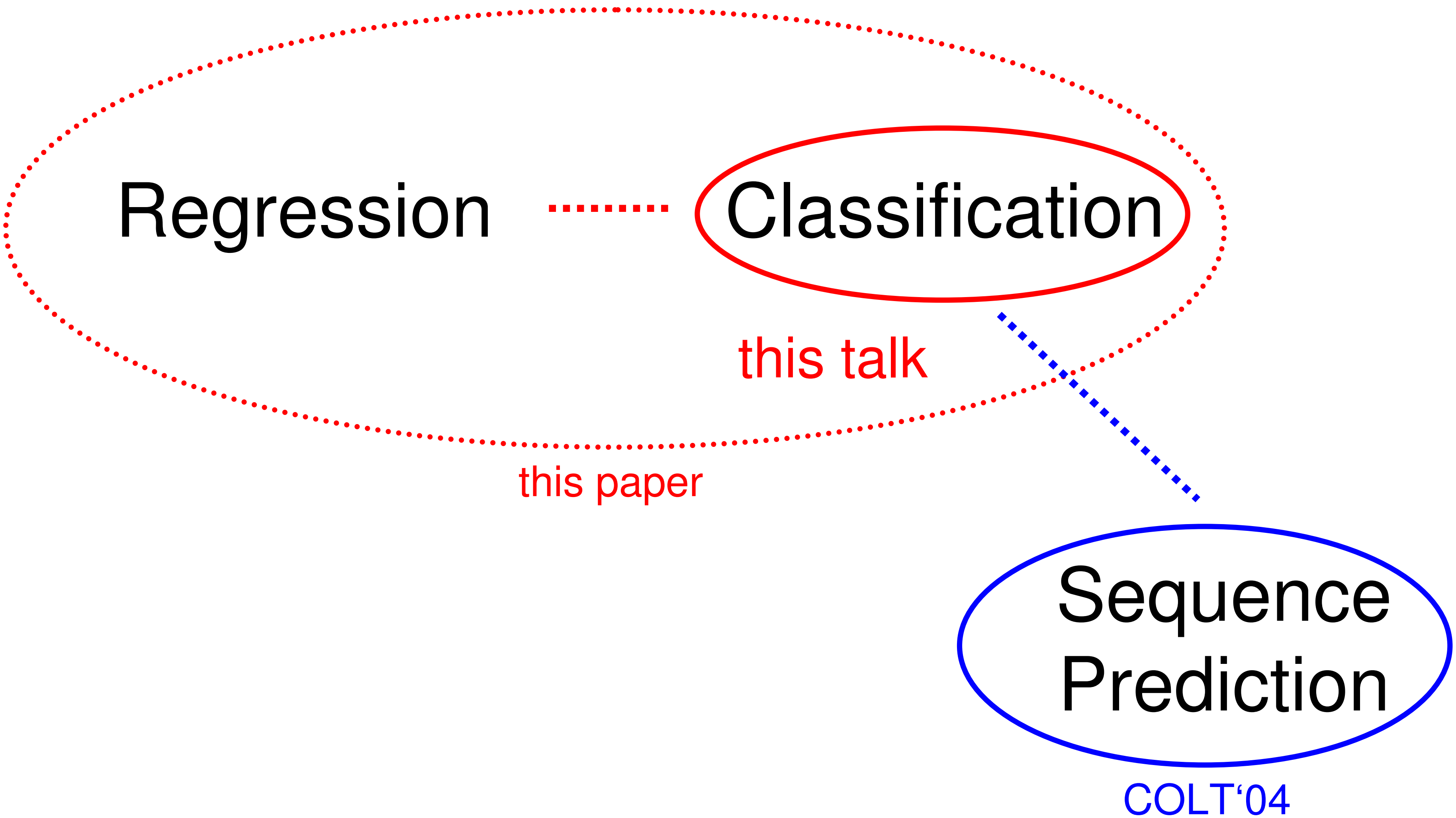
or

A Strange Way of Proving Consistency of MDL Learning

Jan Poland and Marcus Hutter

IDSIA • Lugano • Switzerland

Focus of this Talk



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“?

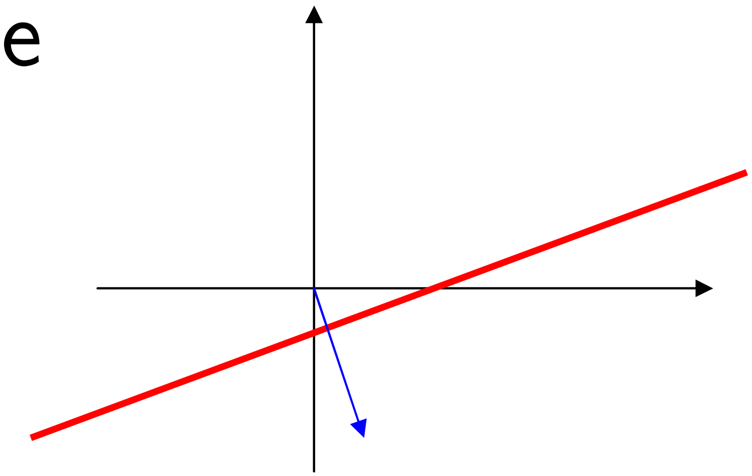
- ~~Identify the exact data generating distribution~~
- Learn the predictive distribution

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 ν 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: $\mathcal{C}^{\mathrm{lin2}} \cong \mathbb{Q}^2$ is the class of rational linear separators on the plane



Proper/Online Learning

Proper Learning assumption:

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

$$\boxed{\mu \in \mathcal{C}}$$

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 *MDL* (or MAP):

$$\varrho^{\text{static}}(y_{n+1}|x_{1:n+1}, y_{1:n}) = \nu^*_{(x_{1:n}, y_{1:n})}(y_{n+1}|x_{n+1})$$

where

$$\boxed{\nu^*_{(x_{1:n}, y_{1:n})} = \arg\max_{\nu \in \mathcal{C}} \{w_\nu \nu(y_{1:n}|x_{1:n})\}}$$

Determine and use the *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:

$$\varrho(y_n|, y_{<n}) = \frac{\varrho(y_{1:n}|x_{1:n})}{\varrho(y_{<n}|x_{<n})}$$

$$\bar{\varrho}(y_n|, y_{<n}) = \frac{\varrho(y_{1:n}|x_{1:n})}{\sum_{y_n} \varrho(y_{1:n}|x_{1:n})}$$

with $\varrho(y_{1:n}|x_{1:n}) = \max_{\nu \in \mathcal{C}} \{w_\nu \nu(y_{1:n}|x_{1:n})\}$.

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:

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

Then the ψ-predictions converge to the μ-predictions *in Hellinger sum* if

$$H^2_{x_{<\infty}}(\mu,\psi) = \sum_{t=1}^{\infty} \mathbf{E}[h_t^2(\mu,\psi)] < \infty.$$

This implies $h_t^2 \to 0$ *almost surely*.

Other Distance Measures

$$s_t(\mu,\psi) \quad = \quad \sum_{y_t\in\{0,1\}} \Big(\mu(y_t|x_{1:t},y_{<t}) - \psi(y_t|x_{1:t},y_{<t})\Big)^2$$

square distance

$$a_t(\mu,\psi) \quad = \quad \sum_{y_t\in\{0,1\}} \Big|\mu(y_t|x_{1:t},y_{<t}) - \psi(y_t|x_{1:t},y_{<t})\Big|$$

absolute distance

$$d_t(\mu,\psi) \quad = \quad \sum_{y_t\in\{0,1\}} \mu(y_t|x_{1:t},y_{<t}) \cdot \ln\frac{\mu(y_t|x_{1:t},y_{<t})}{\psi(y_t|x_{1:t},y_{<t})}$$

Kullback-Leibler divergence

Distance Measures: Properties

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

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

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

- Kullback-Leibler divergence d_t: $\begin{cases} \cancel{\text{triangle inequality}} \\ \cancel{\leq a_t} \end{cases}$

Convergence Theorem

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

$$\varrho^{\text{static}} \longrightarrow \varrho \longrightarrow \bar{\varrho} \longrightarrow \mu$$

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

$$\Rightarrow H^2(\mu, \varrho^{\text{static}}) \leq 21 w_\mu^{-1}$$

Properties of the Proof

- Purely algebraic
- no hidden O-terms
- Inspired by Solomonoffs 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*:

$$L(\varrho) \leq L(\mu) + 42 w_\mu^{-1} + 2\sqrt{42 w_\mu^{-1} L(\mu)}$$

- $\Rightarrow$ expected per-round regret converges to zero almost surely

Discussion

- w_μ^{-1} may be huge
  - Similar bounds hold for the Bayes mixture, e.g.
    $$H^2(\mu,\xi) \le \ln w_\mu^{-1}$$
  - $\Rightarrow$ Bayes mixture converges *much faster* in general
  - The w_μ^{-1} bound for MDL is sharp in general
  - With carefully chosen model class and prior, MDL converges fast, too

Discussion

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

Thank you!