Online Methods in Learning Theory
or
Learning in Countable Classes of Stochastic Models

Jan Poland

IDSIA • Lugano • Switzerland

Overview

- Bayesian framework for sequence prediction (no i.i.d. assumption!)
- Connect this to classification
- Also keep in mind: Universal prediction
- Bayesian predictors: Mixture and MDL
- Online convergence results (asymptotics, loss bounds)
- Proceeding further: offline theorems, active learning, ...?

Rough Problem Setup

- *Prediction*: Given an initial part $x_{1:t} = x_1 x_2 \ldots x_t$ of a sequence, predict the next symbol x_{t+1}. For example
  - $x_{1:t} = 01010101010101$
  - $x_{1:t} = 1100100100001111110110101010001000100001$
  - $x_{1:t} = 0001111001010010001111110110101001001111$

- *Classification*: Given some training data

$$(u, x)_{1:t} = [(u_1, x_1), \ldots, (u_t, x_t)]$$

and an input u_{t+1}, predict output x_{t+1}.

Prediction: (Semi)Measures

- Restrict to binary (output) alphabet $\mathbb{B} = \{0,1\}$
- $\mathbb{B}^\infty = \{$binary sequences$\}$, $\mathbb{B}^* = \{$binary strings$\}$, ϵ is the empty string
- A *measure* μ is a function $\mu : \mathbb{B}^* \to [0,1]$ s.t.

$$\mu(\epsilon) = 1 \text{ and } \mu(x) = \mu(x0) + \mu(x1) \text{ for all } x$$

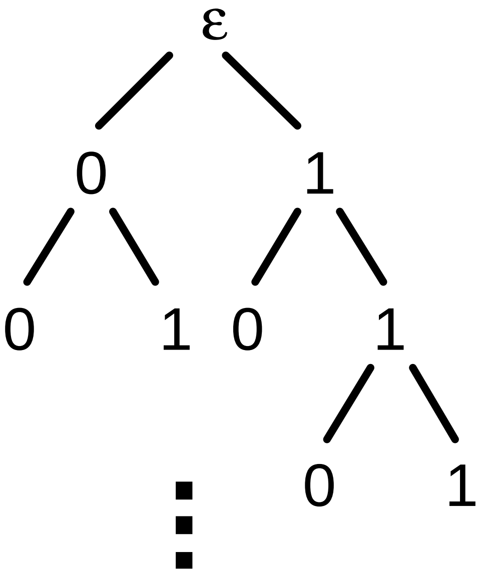
- A *semimeasure* ν has

$$\nu(\epsilon) \leq 1 \text{ and } \nu(x) \geq \nu(x0) + \nu(x1) \text{ for all } x$$

Examples: (Semi)Measures

- $\lambda(x) = 2^{-length(x)}$ is the uniform measure
- $\mu_1(111...1) = 1$ and $\mu_1(x) = 0$ if x contains at least one 0, is a deterministic measure

- $M_U(x) =$ the probability that some universal Turing machine (UTM) U outputs a string starting with x when the input is random coin flips
- The latter is a semimeasure, not a measure, since U does not halt on each input!



Binary Classification

- Need to add input
- $\mu(1|u)$ is i.i.d. given some input $u \in U$
- Conditionalized measure, depends only on input, no history
- Input space U arbitrary, thus may contain history
- Can recover full (non-i.i.d.) sequence prediction setup by letting $U = \mathbb{B}^*$ and $u_{t+1} = x_{1:t}$
- Conversely: All online results also hold with input

Classes of (Semi)Measures

- Let $\mathcal{C}$ be a *countable* class of (semi)measures
- Each $\nu \in \mathcal{C}$ is assigned a *prior weight* $w_\nu > 0$
- Kraft inequality: $\sum_{\nu \in \mathcal{C}} w_\nu \leq 1$
- Universal setup: $\mathcal{C} = \mathcal{M} \cong$ all programs on a UTM U
- $w_\nu = 2^{-K(\nu)}$ where $K(\nu)$ is the *prefix Kolmogorov Complexity* of ν, i.e. the length of the shortest self-delimiting program defining ν

Assumptions

- We make *no probabilistic* assumption on $\mathcal{C}$
- We show bounds for given *true distribution μ*
- which is a *measure* (not a semimeasure)
- *and assumed to be in $\mathcal{C}$*
- Thus, bounds depend on the complexity (or prior weight w_μ) of the true distribution
- Occam's razor
- priors correspond to regularization

Bayes Mixtures

- Bayes mixture $\xi(x) = \sum_{\nu \in \mathcal{C}} w_\nu \nu(x)$
- Bayes mixture prediction:

$$\xi(a|x) = \frac{\sum_\nu w_\nu \nu(xa)}{\sum_\nu w_\nu \nu(x)}$$

  for $a \in \{0,1\}$.
- ξ is (semi)measure
- "Committee of all models"

Minimum Description Length

- Minimum Description Length (MDL) estimator

$$\begin{aligned} \nu^x &= \arg\max\{w_\nu \nu(x)\} \\ \varrho(x) &= \max\{w_\nu \nu(x)\} \end{aligned}$$

- ν^x is *maximizing element*
- $-\log \varrho(x) = \min\{-\log w_\nu - \log \nu(x)\}$
- $-\log w_\nu$ ↔ code of the model
- $-\log \nu(x)$ ↔ code of data given

Prediction using MDL

- Dynamic MDL predictor: $\varrho(a|x)=\frac{\varrho(xa)}{\varrho(x)}$
  not a semimeasure!
- Normalized dynamic MDL: $\varrho(a|x)=\frac{\varrho(xa)}{\varrho(x0)+\varrho(x1)}$
  measure
  search new model for each next symbol
- Static MDL predictor: $\varrho^x(a|x)=\frac{\nu^x(xa)}{\nu^x(x)}$
  (semi)measure
  find best model and use this for prediction
- $\Rightarrow$ Static MDL is computationally more efficient

Bayes Mixture Predictions

Theorem (Solomonoff): Let $\mu \in \mathcal{C}$ be a measure, then

$$\sum_{t=0}^{\infty} \mathbf{E} \sum_{a\in\{0,1\}} \big(\mu(a|x_{1:t}) - \xi(a|x_{1:t})\big)^2 \leq \ln(w_\mu^{-1})$$

$\Rightarrow$ The posteriors *almost surely* converge to the true probabilities *fast*

Proof of Solomonoff's Theorem

$$\begin{aligned}
&\sum_{t=0}^{T} \mathbf{E} \sum_{a\in\{0,1\}} \big(\mu(a|x_{1:t}) - \xi(a|x_{1:t})\big)^2 \\
\leq\;& \sum_{t=0}^{T} \mathbf{E} \sum_{a\in\{0,1\}} \mu(a|x_{1:t}) \ln \frac{\mu(a|x_{1:t})}{\xi(a|x_{1:t})} \;=\; \sum_{t=0}^{T} \mathbf{E} \ln \frac{\mu(x_t|x_{1:t})}{\xi(x_t|x_{1:t})} \\
=\;& \mathbf{E} \ln \left(\prod_{t=0}^{T} \frac{\mu(x_t|x_{1:t})}{\xi(x_t|x_{1:t})} \right) \;=\; \mathbf{E} \ln \frac{\mu(x_{1:T+1})}{\xi(x_{1:T+1})} \;\leq\; \ln w_\mu^{-1}
\end{aligned}$$

Lemma:
The quadratic distance is bounded by the relative entropy.

Observation:
x dominates µ, i.e.
x(*x*) ≥ *w*µ µ(*x*) for all *x*

MDL: Main Theorem

Theorem: $\mu \in \mathcal{C}$ measure, then

$$(i) \qquad \sum_{t=0}^{\infty} \mathbf{E} \sum_{a\in\{0,1\}} \big(\mu(a|x_{1:t}) - \varrho_{\text{norm}}(a|x_{1:t})\big)^2 \leq \ln w_\mu^{-1} + w_\mu^{-1},$$

normalized dynamic

$$(ii) \qquad \sum_{t=0}^{\infty} \mathbf{E} \sum_{a\in\{0,1\}} \big(\mu(a|x_{1:t}) - \varrho(a|x_{1:t})\big)^2 \leq 8 \cdot w_\mu^{-1},$$

dynamic

$$(iii) \qquad \sum_{t=0}^{\infty} \mathbf{E} \sum_{a\in\{0,1\}} \big(\mu(a|x_{1:t}) - \varrho^{x_{1:t}}(a|x_{1:t})\big)^2 \leq 21 \cdot w_\mu^{-1}$$

static

$\Rightarrow$ The posteriors *almost surely* converge to the true probabilities, but convergence is *slow* in general

Proof Idea

- For ϱ_{norm}:
  - use relative entropy bound
  - decompose ϱ_{norm} in ϱ and normalizer
  - ϱ-contribution bounded by $\ln w_\mu^{-1}$
  - normalizer contribution bounded by w_μ^{-1}
- Then bound the cumulative absolute difference $|\varrho - \varrho_{\mathrm{norm}}|$ by $2w_\mu^{-1}$
- Finally bound the cumulative absolute difference $|\varrho^x - \varrho|$ by $3w_\mu^{-1}$
- square distances may be chained

Loss Bounds

- **Theorem** (Hutter): $\mu \in \mathcal{C}$ measure $\Rightarrow$

$$L^{\xi}(T) \leq L^{\mu}(T) + 2\sqrt{L^{\mu}(T)\ln w_{\mu}^{-1}} + 2\ln w_{\mu}^{-1}$$

for $0/1$ los and arbitrary loss

- **Corollary**: For arbitrary loss,

$$L^{\varrho_{\mathrm{norm}}}(T) \leq L^{\mu}(T) + O(\sqrt{L^{\mu}(T)w_{\mu}^{-1}}) + O(w_{\mu}^{-1})$$

Loss Bounds

- **Corollary**: For 0/1 loss,

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

$$L^{\varrho^{x}}(T) \leq L^{\mu}(T) + O(\sqrt{L^{\mu}(T) w_{\mu}^{-1}}) + O(w_{\mu}^{-1})$$

- Arbitrary loss open!
- Compare to prediction with expert advice: *worst-case* loss for *individual* sequences

$$L^{PEA}(T) \leq L^{\mu}(T) + 2\sqrt{2L^{\mu}(T) \ln w_{\mu}^{-1}} + O(\ln w_{\mu}^{-1})$$

Exponential Bounds are Sharp

- MDL bound exponentially worse than Bayes mixture
- This bound is sharp!
- Classification example
  - input space $U = \{1, 2, 3, 4, 5, 6, 7\}$
  - $\nu_1, \ldots, \nu_8$ are *deterministic*
  - true distribution is $\mu = \nu_8$
- A prediction example where $\mathcal{C}$ contains only Bernoulli distributions is possible
- (But: Bernoulli $\Rightarrow$ good bounds hold under mild assumptions)

Exponential Bounds are Sharp

$$
\begin{array}{r ccccccc l}
\text{Input } u & 1 & 2 & 3 & 4 & 5 & 6 & 7 & \\
\nu_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & w_1 = \frac{1}{8} \\
\nu_2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & w_2 = \frac{1}{8} \\
\nu_3 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & w_3 = \frac{1}{8} \\
\nu_4 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & w_4 = \frac{1}{8} \\
\nu_5 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & w_5 = \frac{1}{8} \\
\nu_6 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & w_6 = \frac{1}{8} \\
\nu_7 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & w_7 = \frac{1}{8} \\
\mu = \nu_8 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & w_8 = \frac{1}{8}
\end{array}
$$

Hybrid MDL predictions

- Hybrid MDL predictor: $\varrho^{hybrid}(a|x) = \frac{\nu^{xa}(xa)}{\nu^{x}(x)}$
- “Dynamic MDL but drop weights”
- Predictive properties? Poorer!
- Only converges if the maximizing element *stabilizes*
- This happens almost surely if
  - all (semi)measures in $\mathcal{C}$ are independent of the past (factorizable)
  - μ is uniformly stochastic, i.e. in each time step either deterministic or noisy with at least a certain amplitude

Complexity and Randomness

Universal case: $\mathcal{C} = \mathcal{M}$, and $\tilde{\mathcal{C}}$ is $\mathcal{C}$ restricted to computable measures

$$\Rightarrow 2^{Km(x)} \overset{\times}{=} \tilde{\varrho}(x) \overset{\times}{\leq} \tilde{\xi}(x) \overset{\times}{\leq} \varrho(x) \overset{\times}{=} \xi(x) \overset{\times}{=} M(x)$$

Gács: $\xcancel{\overset{\times}{=}}$ $\Rightarrow$ which inequality is proper?

$\Rightarrow$ all quantities define Martin-Löf randomness by $f(x_{1:n}) \leq C\mu(x_{1:n})$ for all n and some C

Offline bounds?

- We want something like

$$\left|\xi(u_t|u_{<t}, x_{<t}) - \mu(u_t)\right| \leq \frac{\ln w_\mu^{-1} + \ln\frac{1}{\delta}}{t}$$

with probability $1 - \delta$

- Abuse notation: $\mu(u_t) = \mu(1|u_t)$
- Generally, $\left|\xi(u_t|u_{<t}, x_{<t}) - \mu(u_t)\right|$ is *not decreasing* in t
- $\Rightarrow$ no direct conclusion from cumulative online bound possible

Decrease of error?

- Assume $u \overset{i.i.d.}{\sim} D$
- Assume deterministic case, w.l.o.g. $\mu \equiv 1$

- $\xi(u_t|u_{<t}, x_{<t}) \nearrow$
- $\mathbf{E}_t \xi(u_t|u_{<t}, x_{<t}) \nearrow$
- $\mathbf{E}_{1:t} \xi(u_t|u_{<t}, x_{<t}) \nearrow$
- $\mathbf{E}_{1:t} \big(\xi(u_t|u_{<t}, x_{<t}) - 1\big)^2 \searrow$
- Error rate $\searrow$

?

No decrease of error!

Input u	1	2	
$D(u)$	$\frac{1}{3}$	$\frac{2}{3}$	
ν_1	0	1	$w_1 = 0.89$
ν_2	1	0	$w_2 = 0.1$
$\mu = \nu_3$	1	1	$w_3 = 0.01$

$$\mathbf{E}_1 \xi(u_1|\emptyset) = 0.66 \quad \mathbf{E}_{1:2} \xi(u_2|u_1) = 0.58$$

$$\mathbf{E}_1 \Big(1 - \xi(u_1|\emptyset)\Big)^2 = 0.33 \quad \mathbf{E}_{1:2} \Big(1 - \xi(u_2|u_1)\Big)^2 = 0.41$$

Active Learning

$$\begin{array}{r|ccccccc|l}
\text{Input } u & 1 & 2 & 3 & 4 & 5 & 6 & 7 & \\
\nu_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & w_1 = \frac{1}{8} \\
\nu_2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & w_2 = \frac{1}{8} \\
\nu_3 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & w_3 = \frac{1}{8} \\
\nu_4 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & w_4 = \frac{1}{8} \\
\nu_5 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & w_5 = \frac{1}{8} \\
\nu_6 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & w_6 = \frac{1}{8} \\
\nu_7 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & w_7 = \frac{1}{8} \\
\mu = \nu_8 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & w_8 = \frac{1}{8}
\end{array}$$

The End

Thank you!