Confident Bayesian Sequence Prediction

Tor Lattimore
Sequence Prediction

Can you guess the next number?

1, 2, 3, 4, 5, ...
3, 1, 4, 1, 5, ...
1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, ...

<table>
<thead>
<tr>
<th>Thu</th>
<th>Fri</th>
<th>Sat</th>
<th>Sun</th>
<th>Mon</th>
<th>Tue</th>
<th>Wed</th>
<th>Thu</th>
</tr>
</thead>
<tbody>
<tr>
<td>20°</td>
<td>22°</td>
<td>23°</td>
<td>23°</td>
<td>22°</td>
<td>24°</td>
<td>20°</td>
<td>20°</td>
</tr>
<tr>
<td>5°</td>
<td>4°</td>
<td>8°</td>
<td>8°</td>
<td>8°</td>
<td>11°</td>
<td>10°</td>
<td>1°</td>
</tr>
</tbody>
</table>

STOCK MARKET

Google searches:
- can you predict
- can you predict earthquakes
- can you predict your height
- can you predict labour
- can you predict the weather

Press Enter to search.
Sequence Prediction

- $x$ is a sequence over countable alphabet $X$
- $\mu(x)$ is the $\mu$-probability of observing $x$
- $\mu(a|x)$ is the $\mu$-probability of observing $a \in X$ given $x$
- $\mathcal{M}$ is a countable set of measures (sequence generators)

### Hellinger² Distance

$$h_x(\rho, \mu) := \sum_{a \in X} \left( \sqrt{\rho(a|x)} - \sqrt{\mu(a|x)} \right)^2$$

### Total Variation Distance

$$\delta_x(\rho, \mu) := \frac{1}{2} \sum_{a \in X} |\rho(a|x) - \mu(a|x)|$$

$$\sqrt{h_x(\rho, \mu)} \approx \delta_x(\rho, \mu)$$

**Goal:** Construct predictor $\rho$ such that for all $\mu \in \mathcal{M}$

$$\rho(\cdot|x_{<t}) \overset{\text{fast}}{\longrightarrow} \mu(\cdot|x_{<t})$$

when $x$ is sampled from $\mu$
Example

- \( X = \{ \text{heads, tails} \} \)
- \( \mathcal{M} = \{ \mu_\theta \} \) is a countable set of Bernoulli measures (coins)
- \( \rho(\text{heads}|x_{<t}) = \frac{\text{number of heads in } x_{<t}}{\text{number of observations}=t-1} \)

**Theorem (Law of Large Numbers)**

If \( \mu \in \mathcal{M} \), then \( \lim_{t \to \infty} h_{x_{<t}}(\rho, \mu) = 0 \) with \( \mu \)-probability 1
Example (continued)

**Theorem (Corollary of Hoeffding Bound)**

If $\mu \in \mathcal{M}$, then with $\mu$-probability at least $1 - \delta$

$$(\forall t) \quad |\rho(\text{heads}|x_{<t}) - \mu(\text{heads}|x_{<t})| \leq \sqrt{\frac{1}{2t} \log \frac{2t(t+1)}{\delta}}$$

$\delta = \frac{1}{10}$ and $\theta = \frac{1}{2}$
Bayesian Predictors

- No assumption on $\mathcal{M}$ except that it is countable
- $w : \mathcal{M} \to (0, 1]$ is a prior on $\mathcal{M}$
- $\xi(x) := \sum_{\nu \in \mathcal{M}} w(\nu) \nu(x)$ is the Bayes measure

Theorem (Solomonoff & Hutter)

*If* $\mu \in \mathcal{M}$, *then*

- $\lim_{t \to \infty} h_{x < t} (\mu, \xi) = 0$ *with* $\mu$-probability 1
- $\mathbb{E}_\mu \sum_{t=1}^{\infty} h_{x < t} (\mu, \xi) \leq \ln \frac{1}{w(\mu)}$

Done? Not quite, $h_{x < t} (\mu, \xi)$ is unknown. Can you construct a confidence bound on the error like in the Bernoulli case?
Theorem (Bayes Law)

\[ P(H|D) = P(H) \frac{P(D|H)}{P(D)} \]

Posterior belief in hypothesis \( \nu \in \mathcal{M} \) having observed \( x \) is

\[ w(\nu|x) := w(\nu) \frac{\nu(x)}{\xi(x)} \]

Conjecture

*The posterior \( w(\cdot|x) \) concentrates about the truth as data is observed*
Posterior Convergence

Bernoulli measures with true $\theta = \frac{1}{2}$ and $w(\nu) = \frac{1}{21}$
Posterior Convergence

Bernoulli measures with true $\theta = \frac{1}{2}$ and $w(\nu) = \frac{1}{21}$
Posterior Convergence

Bernoulli measures with true $\theta = \frac{1}{2}$ and $w(\nu) = \frac{1}{21}$
Posterior Convergence

Bernoulli measures with true $\theta = \frac{1}{2}$ and $w(\nu) = \frac{1}{21}$
Posterior Convergence

Bernoulli measures with true \( \theta = \frac{1}{2} \) and \( w(\nu) = \frac{1}{21} \)
Bayesian Confidence

**Theorem (Ville)**

*With μ-probability at least 1 − δ*

\[(\forall n), \quad w(\mu|x_n) \geq \delta w(\mu)\]

"With high probability the posterior belief in true hypothesis \(\mu\) never falls below \(\delta w(\mu)\)"

Define set of plausible environments

\[\mathcal{M}(x_n) := \{\nu \in \mathcal{M} : \forall \eta \leq n, \quad w(\nu|x_\eta) \geq \delta w(\nu)\}\]

and confidence bound on error

\[\hat{h}(x_n) := \max \{h_{x_n}(\nu, \xi) : \nu \in \mathcal{M}(x_n)\}\]

**Theorem**

*With μ-probability at least 1 − δ it holds that \(\hat{h}(x_n) \geq h_{x_n}(\mu, \xi)\) for all \(n\)*
Posterior Convergence

bernoulli measures with true $\theta = 1/2$ and $w(\nu) = 1/21$ and $\delta = 1/10$
Posterior Convergence

Bernoulli measures with true $\theta = \frac{1}{2}$ and $w(\nu) = \frac{1}{21}$ and $\delta = \frac{1}{10}$
Posterior Convergence

Bayes predictive distribution

\[ \nu \in \mathcal{M}(x_{<t}) \text{ maximising } h_{x_{<t}}(\nu, \xi) \]

21 Bernoulli measures with true \( \theta = \frac{1}{2} \) and \( w(\nu) = \frac{1}{21} \) and \( \delta = \frac{1}{10} \)
Posterior Convergence

Bayes predictive distribution  \( \nu \in \mathcal{M}(x_{<t}) \) maximising \( h_{x_{<t}}(\nu, \xi) \)  1/2

21 Bernoulli measures with true \( \theta = \frac{1}{2} \) and \( w(\nu) = \frac{1}{21} \) and \( \delta = \frac{1}{10} \)
Posterior Convergence

Posterior at time 20

Bernoulli measures with true $\theta = 1/2$ and $w(\nu) = 1/21$ and $\delta = 1/10$
Bayesian Confidence

\[ \mathcal{M}(x_{<n}) := \{ \nu \in \mathcal{M} : \forall \eta \leq n, w(\nu| x_{<\eta}) \geq \delta w(\nu) \} \]

\[ \hat{h}(x_{<n}) := \max \{ h_{x_{<n}}(\nu, \xi) : \nu \in \mathcal{M}(x_{<n}) \} \]

\[ \delta = \frac{1}{10} \text{ and } \theta = \frac{1}{2}. \]
Bayesian Confidence

$$\mathcal{M}(x_{<n}) := \{ \nu \in \mathcal{M} : \forall \eta \leq n, \ w(\nu | x_{<\eta}) \geq \delta w(\nu) \}$$

$$\hat{h}(x_{<n}) := \max \{ h_{x_{<n}}(\nu, \xi) : \nu \in \mathcal{M}(x_{<n}) \}$$

**Theorem**

*If w is uniform, then with µ-probability at least 1 − δ*

$$\sum_{n=1}^{\infty} \hat{h}(x_{<n}) \lesssim |\mathcal{M}| \left( \ln |\mathcal{M}| + \ln \frac{|\mathcal{M}|}{\delta} \right)$$

Therefore $$\hat{h}(x_{<n})$$ converges *fast* to zero
Knows What It Knows Framework

Environment

agent fails

no

no

$h_{x<t}(\rho, \mu) \leq \epsilon$?

present $x_t$ to agent

Agent

output $\rho(\cdot|x_{<t})$

Am I confident?

yes

yes

no

no

output $\bot$
Knows What It Knows Algorithm

1. **Inputs:** \( \varepsilon, \delta \) and \( M := \{ \nu_1, \nu_2, \cdots, \nu_{|M|} \} \).
2. \( t \leftarrow 1 \) and \( x_{<t} \leftarrow \varepsilon \) and \( w : M \rightarrow [0, 1] \) is uniform
3. **loop**
   4. if \( \hat{h}_t(x_{<t}) \leq \varepsilon \) then
      5. output \( \xi(\cdot | x_{<t}) \)
   6. else
      7. output \( \bot \)
   8. observe \( x_t \) and \( t \leftarrow t + 1 \)

Theorem

The following hold:

1. **The agent fails with probability at most** \( \delta \)
2. **The number of times action** \( \bot \) **is taken is at most**

\[
O \left( \frac{|M|}{\varepsilon} \log \frac{|M|}{\delta} \right)
\]

**with probability at least** \( 1 - \delta \)
constructed frequentist-style confidence intervals for discrete non-i.i.d. Bayes
- Works well in theory and in practise
- Leads to state-of-the-art bounds for KWIK learning
- Generic and applicable elsewhere (Bandits/RL)
- Also have bounds for KL divergence
- Countable case also covered