Sequential Normalized Maximum Likelihood in Log-loss Prediction

Wojciech Kotłowski[1] Peter Grünwald[2]

[1]Institute of Computing Science, Poznań University of Technology, Poland

[2]Centrum Wiskunde & Informatica, Amsterdam, Netherlands

ITW 2012, Lausanne

Sequential prediction with logarithmic loss

(a.k.a. prequential coding, sequential probability assignment)

Sequential prediction with logarithmic loss

(a.k.a. prequential coding, sequential probability assignment)

- Sequence of outcomes $x_1, x_2, \ldots \in \mathcal{X}$, revealed one by one.

Sequential prediction with logarithmic loss
(a.k.a. prequential coding, sequential probability assignment)

- Sequence of outcomes $x_1, x_2, \ldots \in \mathcal{X}$, revealed one by one.
- In each iteration, after observing $x^n = x_1, x_2, \ldots, x_n$, a forecaster predicts x_{n+1} by assigning a distribution $P(\cdot|x^n)$.

Sequential prediction with logarithmic loss

(a.k.a. prequential coding, sequential probability assignment)

- Sequence of outcomes $x_1, x_2, \ldots \in \mathcal{X}$, revealed one by one.
- In each iteration, after observing $x^n = x_1, x_2, \ldots, x_n$, a forecaster predicts x_{n+1} by assigning a distribution $P(\cdot|x^n)$.
- After x_{n+1} is revealed, the forecaster incurs logarithmic loss $-\log P(x_{n+1}|x^n)$.

Sequential prediction with logarithmic loss

(a.k.a. prequential coding, sequential probability assignment)

- Sequence of outcomes $x_1, x_2, \ldots \in \mathcal{X}$, revealed one by one.
- In each iteration, after observing $x^n = x_1, x_2, \ldots, x_n$, a forecaster predicts x_{n+1} by assigning a distribution $P(\cdot|x^n)$.
- After x_{n+1} is revealed, the forecaster incurs logarithmic loss $-\log P(x_{n+1}|x^n)$.
- Regret of the forecaster relative to a set of distributions $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$:

$$\mathcal{R}(P, x^n) = \sum_{i=1}^{n} -\log P(x_i|x^{i-1}) - \inf_{\theta \in \Theta} \sum_{i=1}^{n} -\log P_\theta(x_i|x^{i-1}).$$

Sequential prediction with logarithmic loss

(a.k.a. prequential coding, sequential probability assignment)

- Sequence of outcomes $x_1, x_2, \ldots \in \mathcal{X}$, revealed one by one.
- In each iteration, after observing $x^n = x_1, x_2, \ldots, x_n$, a forecaster predicts x_{n+1} by assigning a distribution $P(\cdot|x^n)$.
- After x_{n+1} is revealed, the forecaster incurs logarithmic loss $-\log P(x_{n+1}|x^n)$.
- Regret of the forecaster relative to a set of distributions $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$:

$$\mathcal{R}(P, x^n) = \sum_{i=1}^{n} -\log P(x_i|x^{i-1}) - \inf_{\theta \in \Theta} \sum_{i=1}^{n} -\log P_\theta(x_i|x^{i-1}).$$

Goal: minimize the worst-case regret.

Sequential prediction with logarithmic loss

(a.k.a. prequential coding, sequential probability assignment)

- Sequence of outcomes $x_1, x_2, \ldots \in \mathcal{X}$, revealed one by one.
- In each iteration, after observing $x^n = x_1, x_2, \ldots, x_n$, a forecaster predicts x_{n+1} by assigning a distribution $P(\cdot|x^n)$.
- After x_{n+1} is revealed, the forecaster incurs logarithmic loss $-\log P(x_{n+1}|x^n)$.
- Regret of the forecaster relative to a set of distributions $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$:

$$\mathcal{R}(P, x^n) = \sum_{i=1}^{n} -\log P(x_i|x^{i-1}) - \inf_{\theta \in \Theta} \sum_{i=1}^{n} -\log P_\theta(x_i|x^{i-1}).$$

  Goal: minimize the worst-case regret.
- We choose $\mathcal{P}$ to be an exponential family of distributions (Gaussian, Bernoulli, Poisson, binomial, Gamma, etc.)

Sequential prediction with logarithmic loss

(a.k.a. prequential coding, sequential probability assignment)

- Sequence of outcomes $x_1, x_2, \ldots \in \mathcal{X}$, revealed one by one.
- In each iteration, after observing $x^n = x_1, x_2, \ldots, x_n$, a forecaster predicts x_{n+1} by assigning a distribution $P(\cdot|x^n)$.
- After x_{n+1} is revealed, the forecaster incurs logarithmic loss $-\log P(x_{n+1}|x^n)$.
- Regret of the forecaster relative to a set of distributions $\mathcal{P} = \{P_\theta | \theta \in \Theta\}$:

$$\mathcal{R}(P, x^n) = \sum_{i=1}^{n} -\log P(x_i|x^{i-1}) - \inf_{\theta \in \Theta} \sum_{i=1}^{n} -\log P_\theta(x_i|x^{i-1}).$$

  Goal: minimize the worst-case regret.
- We choose $\mathcal{P}$ to be an exponential family of distributions (Gaussian, Bernoulli, Poisson, binomial, Gamma, etc.)
- No assumptions on the process generating the outcomes!

Algorithms

The minimax algorithm

Normalized maximum likelihood (NML) achieves the minimal worst-case regret:

$$P_{\mathrm{NML}} = \arg\min_{P} \max_{x^n} \mathcal{R}(P, x^n) = \frac{k}{2} \log n + O(1)$$

- ☺ Optimal
- ☹ Hard to calculate, often impractical
- ☹ Requires knowledge of time horizon

Algorithms

Maximum likelihood (ML) strategy

Predicts with the best distribution on past outcomes:

$$P(x_{n+1}|x^n) = P_{\hat{\theta}_n}(x_{n+1}),$$

where $\hat{\theta}_n = \arg\min_\theta \sum_{i=1}^n - \log P_\theta(x_i)$.

- ☺ Simple to calculate, often used in practice
- ☹ Suboptimal: the constant in $O(\log n)$ much larger than $\frac{k}{2}$
- ☹ Requires bounding the data to achieve logarithmic regret

Our contribution

Include the current (to be predicted) outcome into calculation of the maximum likelihood

Our contribution

Include the current (to be predicted) outcome into calculation of the maximum likelihood

Sequential normalized maximum likelihood (SNML):

$$P(x_{n+1}|x^n) \propto P_{\hat{\theta}_{n+1}}(x^{n+1}),$$

where $\hat{\theta}_{n+1} = \arg\min_\theta \sum_{i=1}^{n+1} - \log P_\theta(x_i)$

Our contribution

Include the current (to be predicted) outcome into calculation of the maximum likelihood

Sequential normalized maximum likelihood (SNML):

$$P(x_{n+1}|x^n) \propto P_{\hat{\theta}_{n+1}}(x^{n+1}),$$

where $\hat{\theta}_{n+1} = \arg\min_\theta \sum_{i=1}^{n+1} -\log P_\theta(x_i)$

- Achieves asymptotically optimal regret $\frac{k}{2}\log n + O(1)$.
- SNML coincides with NML given that the current iteration is the last iteration.
- Relationship to Bayesian strategy with Jeffreys' prior.